RÉSUME

Aujourd'hui, le World Wide Web se veut l'une des plates-formes les plus sollicitées pour la diffusion et la recherche d'information. De ce fait, de nombreux chercheurs le considèrent comme la meilleure source d'information, sur laquelle ils s'appuient pour leurs fouilles des données. Il convient toutefois de souligner que le choix et la sélection des données sources sont cruciaux, dans la mesure où ils impactent directement le résultat final. En nous basant sur des études théoriques sur la fouille de texte, nous avons essayé d'élaborer un modèle sur lequel le chercheur pourrait se baser pour concevoir ses projets de fouille de contenu du web, et pour faciliter le choix et sélection des données sources selon ses objectifs de recherche. Par ailleurs, compte tenu du fait que la fouille de contenu web se distingue de la fouille texte uniquement par ses méthodes de collecte des données, nous avons développé des méthodes dédiées à la collecte automatique et à l'extraction intelligente des données, notamment le projet RCrawler développé avec le langage R. En dépit des difficultés et des limites relatives à ce processus, les expérimentations réalisées ont témoigné de la performance et l'efficacité des approches proposées.

Mots-clés : fouille de contenu du Web, fouille de texte, analyse de texte, collection des données, robot d'indexation, extraction des données

N°d'ordre : 292/20

Contribution au Développement d'Approches pour la Collecte et l'Extraction de contenue Web.

KHALIL SALIM

N° d'ordre : 292/20

UNIVERSITE SULTAN MOULAY SLIMANE

Faculté des Sciences et Techniques

Béni-Mellal

Centre d'Études Doctorales : Sciences et Techniques

Formation Doctorale : **Mathématiques et Physiques Appliquées**

THÈSE

Présentée par

JABIR BRAHIM

Pour l'obtention du grade de

DOCTEUR

Discipline : Informatique

Spécialité : Informatique

Contribution au Développement d'Approches pour la Collecte et l'Extraction de contenue Web.

Soutenue le Samedi 28 Novembre 2020 à 10h devant la commission d'examen:

Pr.Said SAFI	Professeur, Université Sultan Moulay Slimane, F.P Béni-Mellal, Maroc	Président
Pr.Youssef ES-SAADY	Professeur, Université Ibn Zohr, Taroudant , Maroc	Rapporteur
Pr.Rachid EL AYACHI	Professeur, Université Sultan Moulay Slimane, F.P. Béni-Mellal, Maroc.	Rapporteur
Pr.Mohamed BASLAM	Professeur, Université Sultan Moulay Slimane, F.S.T Béni-Mellal, Maroc	Rapporteur
Pr.Mohamed FAKIR	Professeur, Université Sultan Moulay Slimane, F.S.T. Béni-Mellal, Maroc.	Directeur de Thèse

Dédicace

A à ma source de joie et de fierté.

A mon fils Sami KHALIL

Si une chose est susceptible à me rendre heureux, c'est le bonheur de t'avoir dans ma vie.

Tu m'as toujours soutenu avec tes sourires et ta présence.

Puisse DIEU te garder, te préserver, et te procurer santé et bonheur.

Que cette dédicace soit ta source d'inspiration, d'aspiration et de motivation.

Remerciements

Mes tous premiers mots de remerciements vont à l'endroit de mon directeur de thèse Mr Mohamed FAKIR pour ses précieux conseils, sa compréhension et sa bonne patience tout au long de mon parcours doctoral. Grâce à son soutien, son expertise, ses remarques et corrections, mes travaux de thèse ont pu être terminés

Je remercie ma maman Fettouma RAHOULI pour son encouragement continue et pour ses prières fortifiantes, j'éprouve une immense gratitude pour sa présence à mes côtés. Je remercie également ma femme Fatima Zahra DAZAHRA pour son dévouement envers notre foyer et pour son amour qui ont joué un rôle important dans ma carrière.

Je ne pourrai pas oublier ma grand-mère FATIMA ainsi que ma tante Touria BENTOUILA et son mari Stefano Silvestri qui m'ont toujours encouragé et soutenu moralement. Je leur exprime ma profonde admiration.

Je tiens à exprimer ma profonde gratitude les membres du Centre d'Etudes Doctorales de la FST de Beni Mellal et plus spécialement les membres du laboratoire TIAD, les doctorants, les enseignants chercheurs qui, par leur curiosité intellectuelle, leur savoir-faire et leurs conseils et remarques, m'ont incité à arriver jusqu'au bout de mes travaux de thèse.

Un grand merci à tous les professeurs qui ont accepté de participer à ce jury de thèse. Leurs remarques pertinentes ont contribué à valoriser mon travail de recherche ainsi surmonter quelques points de faiblesse. Soyez assurés de ma plus profonde reconnaissance pour l'attention que vous avez porté à ce manuscrit et pour le temps que vous avez consacré à son évaluation.

Tous mes remerciements à l'ensemble des professeurs, qui m'ont enseigné dès mes études primaires, secondaires jusqu'aux études supérieurs, grâce à eux j'ai pu arriver à ce niveau d'étude. Je leur exprime ma gratitude et ma reconnaissance.

Résumé

Aujourd'hui, le World Wide Web se veut l'une des plates-formes les plus sollicitées pour la diffusion et la recherche d'information. De ce fait, de nombreux chercheurs le considèrent comme la meilleure source d'information, sur laquelle ils s'appuient pour leurs fouilles des données. Il convient toutefois de souligner que le choix et la sélection des données sources sont cruciaux, dans la mesure où ils impactent directement le résultat final. En nous basant sur des études théoriques sur la fouille de texte, nous avons essayé d'élaborer un modèle sur lequel le chercheur pourrait se baser pour concevoir ses projets de fouille de contenu du web, et pour faciliter le choix et sélection des données sources selon ses objectifs de recherche. Par ailleurs, compte tenu du fait que la fouille de contenu web se distingue de la fouille texte uniquement par ses méthodes de collecte des données, nous avons développé des méthodes dédiées à la collecte automatique et à l'extraction intelligente des données, notamment le projet RCrawler développé avec le langage R. En dépit des difficultés et des limites relatives à ce processus, les expérimentations réalisées ont témoigné de la performance et l'efficacité des approches proposées.

Mots-clés : fouille de contenu du Web, fouille de texte, analyse de texte, collection des données, robot d'indexation, extraction des données.

Abstract

Today, the World Wide Web is one of the most popular platforms for sharing and publishing of information. As a result, many researchers consider it the best source of information for their data mining projects. However, the choice and the selection of data source data is as crucial step as it directly impacts the result quality. Based on our theoretical studies on text mining, we have tried to develop a model on which the researcher could rely on to design his own web content mining projects according to his own research objectives, by making the data source selection very simple. In addition, given the fact that web content mining differs from text mining only by its data collection methods, we have developed methods dedicated to automatic data collection and intelligent data extraction, in particular the project RCrawler, which is a multithreaded web crawler and web scraper based on R platform. Despite the difficulties and limits relating to this process, the experiments we have carried out have demonstrated the performance and effectiveness of our proposed approaches.

Keywords: web content mining, text mining, text analysis, data collection, web crawling, web scraping

Table des matières

Liste des tableaux

Liste des figures

Liste des tableaux

Liste des figures

Introduction

Le World Wide Web (ou le Web tout court) a eu un immense impact sur l'ensemble des aspects de nos vies. Il s'agit de la source d'informations la plus importante et la plus connue, facilement accessible et consultable. Il se compose de milliards de documents interconnectés (appelés pages Web) rédigés par des millions de personnes. Depuis son introduction, le Web a radicalement changé notre comportement vis-à-vis de la recherche d'informations.

La fouille de la structure Web extrait des modèles des structures de liaison entre les pages et présente le Web comme étant un graphique dirigé dans lequel les nœuds représentent des pages et les bords dirigés représentent des liens. La fouille d'utilisation Web exploite les modèles d'activité des utilisateurs recueillis à partir de l'analyse des enregistrements de journaux Web, afin de comprendre le comportement des utilisateurs lors des visites de sites Web. La fouille de contenu Web extrait des informations précieuses du contenu Web. Ce dernier est réalisé avec deux objectifs : la fouille des résultats de recherche, qui est exploitée pour améliorer les moteurs de recherche et les champs de recherche d'informations ; et la fouille de contenu de page Web, qui extrait le contenu de page Web à des fins d'analyse et d'exploration.

Dans la fouille de contenu Web, la collecte de données constitue une tâche importante. En effet, plusieurs applications de fouille de données Web utilisent des robots d'indexation Web pour le processus de récupération et de collecte de données à partir du Web. Une fois les données collectées, nous suivons le même processus en trois étapes : le prétraitement des données, l'exploration des données Web et le post-traitement. Cependant, les techniques utilisées pour chaque étape peuvent être très différentes de celles utilisées dans la fouille de données traditionnelle (voir figure ci-dessous).

Les robots d'indexation, appelés également araignées, sont des programmes qui parcourent et téléchargent automatiquement les pages Web en suivant des hyperliens de manière méthodique et automatisée. Il existe différents types de robots d'indexation Web. Les robots universels sont destinés à explorer et à indexer toutes les pages Web, quel que soit leur contenu. D'autres, appelés robots d'exploration préférentiels, sont

davantage ciblés sur un sujet ou un thème spécifique. Les robots d'indexation Web sont connus principalement pour soutenir les actions des moteurs de recherche, en particulier dans l'indexation.

Afin d'améliorer la précision des résultats de la fouille de contenu web, seules les données précieuses et significatives doivent être extraites. Certaines données non pertinentes, telles que les bannières de barre de navigation et les publicités, doivent être exclues et marginalisées, ce qui implique un processus la fouille des données efficace. L'extraction de données, ou “scraping de données“, est le problème de l'extraction d'informations cibles à partir de pages Web afin de produire des données structurées qui sont prêtes pour le post-traitement.

Pour répondre aux questions de recherche, nous consacrerons le **Chapitre 1** aux différentes méthodologies de fouille et d’analyse de texte, et présentera une esquisse générale concernant les principales approches de l'analyse de texte, ainsi que les risques relatifs à l'analyse des données issues des sources en ligne.

Le **chapitre 2** aura pour but d’aider le chercheur à concevoir et à modéliser un problème de fouille de contenu web, à savoir où les décisions les plus critiques sont prises dans ce type de projet de recherche de ce type**.** Cela permettra au chercheur à identifier les décisions les plus délicates lors de la recherche de la fouille de texte, et à s’accoutumer aux divers modèles de recherche (recherche qualitative, quantitative et mixte. Dans le **chapitre 3**, nous allons mettre la lumière sur les méthodes et les techniques de fouille de texte, à savoir la collecte et le traitement et l'analyse des données. Nous mettrons également en exergue des études de projets de fouille d'opinion, d’extraction de l’information et l’analyse de sujets.

Le chapitre 4 mettra le phare sur les travaux de recherches que nous avons réalisés sur le robot d’indexation RCrawler, détenteur d’un ensemble de fonctions consubstantielles à l’exploration Web, le grattage web et l’analyse des liens potentiels. Et dans le **chapitre 5**, nous allons effectuer quelques tests préliminaires, procéder à une évaluation multithreading, une évaluation de la performance en termes de vitesse, et une évaluation du grattage web. En dernier lieu, nous discuterons les résultats de l'évaluation de ces techniques.

Chapitre 1 : Les méthodologies de fouille et d'analyse de texte

1. Introduction

La fouille de texte est un domaine intéressant qui met en place de nouvelles méthodes de recherche, et des outils logiciels émergents pour une utilisation accrue dans le monde universitaire, professionnel ainsi que les organismes gouvernementaux. Aujourd'hui, les chercheurs et scientifiques utilisent les outils de fouille de texte dans de grands projets pour des fins de prédiction tel que la direction des marchés boursiers (Bollen, Mao et Zeng, 2011) ou la survenue des protections politiques (Kallus, 2014).

La fouille de texte est également utilisée dans différents domaines, tel que le marketing, le commerce et les travaux du gouvernement et de la défense. Au cours des dernières années, la fouille de texte a inauguré sa trajectoire dans les sciences sociales ainsi que dans divers disciplines universitaires tel que l'anthropologie (Acerbi, Lampos, Garnett et Bentley, 2013; Marwick, 2013), les communications (Lazard, Scheinfeld, Bernhardt, Wilcox et Suran, 2015), l'économie (Levenberg, Pulman, Moilanen, Simpson et Roberts, 2014), l'éducation (Evison, 2013), la science politique (Eshbaugh-Soha, 2010; Grimmer et Stewart, 2013), la psychologie (Colley & Neal, 2012; Schmitt, 2005) et finalement la sociologie (Bail, 2012; Heritage & Raymond, 2005; Mische, 2014).

L'analyse de texte est une analyse systématique des mots dans un texte donné. Elle combine généralement des méthodes statistiques formelles et d'autres techniques d'interprétation moins formelles et plus humanistes. L'analyse de texte a commencé dès les années 1200 avec le dominicain Hugh de Saint-Cher et son équipe constituée d'une centaine de frères qui ont créé la première concordance biblique. Il existe également des preuves d'études européennes sur les journaux à la fin des années 1600, et dans ce contexte, le premier bien documenté a été une analyse quantitative du texte réalisée en Suède dans les années 1700 lorsque l'église d'État suédoise a analysé le contenu idéologique des hymnes populaires qui semblaient remettre en question l'orthodoxie de l'église (Krippendorff, 2013, p. 10-11).

Le domaine de l'analyse de texte a connu un développement exponentiel au XXe siècle de telle sorte que les chercheurs en sciences sociales et humaines ont développé un large éventail de techniques d'analyse de textes, notamment des méthodes reposant

fortement sur l'interprétation humaine des textes ainsi que sur des méthodes statistiques formelles. D'un autre côté, une analyse de nature quantitative systématique des journaux a été réalisée à la fin des années 1800 et au début des années 1900 par des chercheurs tel que le chercheur Speed (1893). Ces chercheurs ont démontré à fin des années 1800 que les journaux de New York avaient diminué leur couverture des questions littéraires, scientifiques et religieuses en faveur du sport, des potins et scandales. Cependant, des études d'analyse de texte similaires ont été réalisées par Wilcox (1900), Fenton (1911) et White (1924) qui ont quantifié l'espace journalistique consacré à différentes catégories de nouvelles.

À partir des années 1920 et jusqu'à 1940, Lasswell et ses collègues ont mené des études d'analyse de contenu révolutionnaires sur les messages politiques et la propagande (par exemple, Lasswell, 1927). Les travaux de Lasswell ont inspiré des projets d'analyse de contenu à un grand échelle, notamment le projet General Inquirer à Harvard, qui est un lexique visant à attacher des informations syntaxiques, sémantiques et pragmatiques aux mots tagués d'une partie du discours (Stone, Dunphry, Smith et Ogilvie, 1966).

2. Les méthodes d'analyse de texte relatives aux sciences humaines et sociales

2.1. L'analyse narrative

A. Introduction

L'analyse narrative se concentre sur la façon dont les gens créent et utilisent des histoires pour interpréter le monde. Il convient de rappeler que l'analyse narrative ne traite pas les récits comme des histoires qui transmettent un ensemble de faits sur le monde et n'est donc généralement pas intéressée à savoir si les histoires sont objectivement vraies ou non. Les chercheurs narratifs considèrent les récits comme des produits sociaux mis en place par des personnes dans différents contextes : sociaux, historiques et culturels. Les récits sont considérés comme des dispositifs d'interprétation que les gens utilisent pour se représenter eux-mêmes et leur monde auprès des autres. Les spécialistes en narratologie soutiennent le fait que les représentations que les gens ont d'eux-mêmes prennent souvent la forme d'histoires et que les histoires publiques circulant dans la culture populaire fournissent des ressources que les gens utilisent à la fois pour construire leurs récits et identités personnels d'une part (Ricoeur, 1991) et pour

lier le présent au passé d'une autre part. De telles histoires se retrouvent souvent dans les comptes rendus d'entretiens (Gee, 1991).

Les théories du récit suggèrent que les principaux éléments qui donnent une forme narrative aux textes sont les séquences et les conséquences des événements par lesquels les récits organisent, connectent et évaluent les événements comme significatifs pour des publics particuliers. Avec ces éléments, les conteurs interprètent le monde social pour leur public. Les spécialistes en narratologie caractérisent les histoires en termes de transformation (changement dans le temps) des personnages et des actions qui sont rassemblées dans une intrigue. Les histoires rassemblent de nombreux éléments de l'intrigue, y compris les digressions et les sous-intrigues, c'est ce qu'on appelle un processus "emplotmnet" (White, 1978). Les récits doivent avoir un point, et leur point prend souvent la forme d'un message moral.

Concepts basiques

- Les récits sont des histoires que les gens se racontent entre eux pour interpréter et motiver le comportement social.
- Les histoires publiques sont des récits qui circulent dans la culture populaire.
- "Empltment" fait référence au processus de rassemblement des personnages et des actions dans une intrigue qui implique un changement au fil du temps.

Une grammaire narrative est une structure narrative de base qui se répète dans de nombreux genres narratifs divers.

B. Les approches de l'analyse narrative

Il existe plusieurs approches principales de l'analyse narrative, trois des approches les plus influentes étant les approches structurelles, fonctionnelles et sociologiques. Les approches structurelles de l'analyse narrative opèrent principalement au niveau de l'analyse textuelle, c'est-à-dire qu'elles se concentrent sur les textes eux-mêmes plutôt que sur les contextes sociaux et historiques dans lesquels les histoires émergent, circulent et changent. L'analyse narrative structurelle se concentre sur ce que l'on appelle une grammaire narrative. Un des premiers théoriciens de la grammaire des histoires, Propp (1968) a soutenu que le conte de fées a une forme narrative qui est au centre de toute narration.

Le conte de fées n'est pas structuré par la nature des personnages mais par la fonction qu'ils jouent dans l'intrigue, et le nombre de fonctions possibles est assez faible. Dans son approche structurelle influente du récit, Labov (1972) a défini le récit comme

« une méthode pour récapituler l'expérience passée en faisant correspondre une séquence verbale de clauses à la séquence d'événements qui s'est effectivement produit» (Labov 1972, pp. 359 –360; voir aussi Labov et Waletzky, 1967, p. 20). Pour Labov (1972), un récit minimal est « une séquence de deux clauses qui sont ordonnées dans le temps ». Le squelette d'un récit consiste donc en une série de clauses ordonnées temporellement appelées clauses narratives (Labov, 1972, pp. 360–361). Alors que les récits nécessitent des clauses narratives, toutes les clauses trouvées dans le récit ne sont pas des clauses narratives. Labov a fourni l'exemple suivant :

1. Je connais un garçon nommé Harry.
2. Un autre garçon lui a lancé une bouteille dans la tête.
3. Il a obtenu sept points.

Dans ce passage narratif, seules les clauses B et C sont des clauses narratives. La clause A est une clause libre dans la terminologie de Labov car elle n'a pas de composante temporelle. Elle peut être déplacé librement dans le texte sans altérer la signification du texte. Ce n'est pas le cas avec les clauses narratives, où une restitution des clauses entraîne généralement un changement de sens (Labov, 1972, p. 360).

Labov a également proposé qu'il y ait six parties fonctionnelles distinctes dans un récit entièrement formé : (1) l'abstrait, (2) l'orientation, (3) l'action compliquée , (4) l'évaluation, (5) le résultat ou la résolution, et (6) la coda . De ces six parties, seule l'action compliquée, qui constitue le corps principal des clauses et « comprend généralement une série d'événements» (Labov et Waletzky, 1967, p. 32), est «essentielle si nous voulons reconnaître un récit» (Labov, 1972, p. 370). Voir le tableau 3.1

L'approche fonctionnelle de l'analyse narrative a été lancée par le psychologue Bruner (1990). Bruner a fait valoir que l'ordre de l'expérience des humains se produit dans deux modes de base. Le premier est le mode paradigmatique, ou mode logico-scientifique. Ce mode tente de réaliser l'idéal d'un système mathématique formel de description et d'explication. Ce mode est typique pour l'argumentation en philosophie et en sciences naturelles. En revanche, dans le mode narratif d'organisation des expériences dans lequel la particularité et la spécificité des événements ainsi que l'implication, la responsabilisation et la responsabilité des personnes dans la réalisation d'événements spécifiques sont plus importantes que les considérations logiques.

Tableau 1.1 Exemples d'études d'analyse narrative

	Étude qualitative	**Étude sur méthodes mixtes**
Analyse narrative structurelle	Laird, McCance, McCormack, et Gribben, 2015	Frazosi, De Fazio, et Vicari, 2012
Analyse narrative fonctionnelle	Stroet, Opdenakker et Minnaer, 2015	
Analyse narrative sociologique	Andersen, 2015	Mische, 2014

L'analyse fonctionnelle du récit diffère de l'analyse structurelle car elle se concentre sur les éléments structurels plus que les textes eux-mêmes, elle se concentre sur ce que font des histoires particulières dans le contexte de la vie quotidienne des gens. Pour Bruner, certaines des fonctions du récit comprennent la résolution de problèmes, la réduction de la tension et la résolution de dilemmes. Les récits permettent aux gens d'expliquer les décalages entre l'exceptionnel et l'ordinaire.

Les récits ne sont pas nécessaires lorsque des événements se produisent qui sont perçus comme ordinaires, mais sont nécessaires pour permettre aux gens de refondre des expériences inconnues ou chaotiques en histoires causales afin de donner un sens à de telles expériences et de les rendre familières et sûres. La théorie du développement social fonctionnelle de Vygotsky (Wertsch 1985) et la grammaire fonctionnelle systématique de Halliday (1985), ainsi que la recherche du psychologue Michael Bamberg qui utilise des méthodes narratives fonctionnelles, sont étroitement liées à l'approche de Bruner. Bamberg étudie la formation de l'identité des adolescents ainsi que l'émergence des identités professionnelles (Bamberg, 2004).

Les approches fonctionnelles du récit ont influencé ce que l'on appelle la «tradition de l'histoire de la vie» dans les domaines de la narratologie, de la psychologie et de la recherche en gestion. Les psychologues ont utilisé des récits autobiographiques à la fois pour la recherche et dans leur pratique thérapeutique. Leur intérêt n'est pas principalement dans le contenu des histoires de vie en soi mais dans la façon dont les individus racontent leurs histoires: ce qu'ils soulignent et omettent, leur position en tant que protagonistes ou victimes, et finalement les relations que leurs histoires établissent entre le conteur et le public (voir Rosenwald & Ochberg, 1992). Pour les chercheurs

d’histoires de vie, les histoires personnelles ne sont pas seulement des moyens de partager avec une autre personne, ou avec soi-même, des informations sur sa vie, mais ils sont aussi des moyens de façonner des identités personnelles.

La troisième et dernière approche de l'analyse narrative que nous allons aborder est une approche sociologique qui se concentre sur les contextes culturels, historiques et politiques dans lesquels des histoires particulières sont racontées par des narrateurs précis à des publics particuliers. Le sociologue britannique Plummer’s Telling Sexual Stories (1995) est une analyse narrative sociologique des «coming out stories». Plummer (1995) a soutenu le fait que de telles histoires sont des «rites d'une culture de narration sexuelle » (p. I). Cette culture, qui a émergé à la fin du 20e siècle, a promu la transformation d'expériences qui étaient autrefois considérées comme personnelles, privées et pathologiques en histoires publiques et politiques. L'étude de Plummer sur la prolifération de ces types d'histoires est basée sur des entrevues transcrites avec des personnes racontant leurs propres histoires biographiques impliquant de se remettre d'un viol ou de sortir gaies ou lesbiennes et de façonner des identités personnelles basées sur la participation à des communautés basées sur une identité sexuelle ou des objectifs politiques liés à l'expérience intime.

C. Planification d’un projet de recherche sur l’analyse narrative

Il n'y a pas de procédure convenue pour planifier un projet de recherche sur l'analyse narrative. Comme c'est le cas pour de nombreuses méthodes de recherche qualitative (Creswell, 2014, p. 46), la conception de la recherche en analyse narrative est souvent émergente dans ce sens. Le plan initial d'un projet d'analyse narrative ne peut pas être strictement prescrit, mais il est probable que le plan change pendant la collecte des données et l'analyse préliminaire.

Les questions de recherche peuvent également changer à fur et à mesure que des modèles émergent dans les données et, par conséquent, nous pouvons choisir de modifier la stratégie et tactiques de collecte de données. Néanmoins, l'analyse narrative suivra idéalement le concept de congruence méthodologique avancé par Richards et Morse (2013 ; voir Creswell, 2014, p. 50). La congruence méthodologique se réfère aux objectifs, questions et méthodes d'un projet de recherche interconnectés de sorte que l'étude apparaisse comme un ensemble cohérent plutôt qu’un ensemble de parties indépendantes.

Malgré qu'il n'y ait pas de recette convenable pour effectuer une analyse narrative, les chercheurs narratifs sont censés concevoir leur étude avec une

compréhension générale de l'intention et de la justification de la recherche qualitative. Les modèles de recherche en analyse narrative suivent généralement l'approche traditionnelle consistant à présenter un problème, poser une question, collecter des données pour répondre à la question et enfin analyser les données pour répondre à cette dernière.

Les critères de Creswell (2014, pp. 53-55) pour une bonne recherche qualitative s'appliquent à l'analyse narrative, y compris la nécessité d'employer des procédures de collecte de données rigoureuses, en commençant par un objectif ou un concept unique à explorer, puis en ayant une section de méthodes détaillées et en respectant finalement les normes éthiques les plus élevées possibles (voir tableau 3.2).

Tableau 1.2: Les principes du design de la recherche d'analyse narrative

S’assurer que les procédures de collecte de données sont rigoureuses au maximum
Utiliser une approche de reconnaissance pour les analyses narratives
Analyser les données sur sur différents niveaux d’abstraction et divers angles
Adhérer le maximum possible aux standards de l’éthique

D. L’analyse narrative qualitative

Les approches de l'analyse narrative de nature qualitative dépendent de l'interprétation humaine des textes et sont largement utilisées dans les sciences sociales et humaines. En sciences sociales, les approches qualitatives sont particulièrement populaires dans les domaines de la recherche appliquée tels que l'éducation, la santé, la gestion et la recherche touristique.

Un exemple d'étude qualitative qui adopte une approche fonctionnelle de l'analyse narrative est un article publié en 2015 par Stroet, Opdenakker et Minnaert , des chercheurs en éducation, (2015), qui ont mené une analyse narrative des interactions enseignant-élève dans deux classes organisées différemment. Leur objectif était de mieux comprendre les manifestations des dimensions positives et négatives du besoin de l'enseignement de soutien en reliant ces dimensions aux approches pédagogiques des écoles. Pour les deux écoles, Stroet et ses collègues ont analysé les cours de mathématiques et de langue de septième année et ont trouvé à la fois des différences et

des similitudes frappantes entre les classes dans les manifestations de l'enseignement de soutien.

Les chercheurs en santé et en performance humaine Busanich, McGannon et Schinke (2014) ont adopté une approche différente de celle de Stroet et ses collègues; Busanich et ses collègues ont utilisé une analyse narrative structurelle et un constructivisme social pour comparer les expériences alimentaires désordonnées d'un coureur d'élite masculin et féminin. Leurs données étaient des transcriptions de quatre entretiens approfondis, les deux coureurs participant à deux entretiens distincts. Les expériences des coureurs ont été encadrées par des récits culturels autour de la nourriture, du corps et de l'exercice. Busanich et ses collègues ont découvert que les deux coureurs se sont appuyés sur un récit de performance pour construire des expériences de course à pied et des identités personnelles en tant qu'athlètes d'élite et que lorsque l'identité sportive d'élite est devenue menacée par des moments d'échec perçu, des pensées et des comportements alimentaires désordonnés ont émergé pour les deux coureurs.

E. Les méthodes mixtes et les études sur l'analyse narrative quantitative

Depuis les années 1980, les spécialistes des sciences sociales ont développé des méthodes d'analyse narrative qui intègrent des méthodes qualitatives dans les plans de recherche de méthodes mixtes (voir chapitre 4; Teddlie & Tashakkori, 2008; Tashakkori & Teddlie, 2010). Ces modèles de recherche impliquent l'analyse des schémas de mots dans les récits à l'aide de logiciels et d'outils statistiques.

En sociologie, l'une des approches de méthodes mixtes les plus importantes de l'analyse narrative est l'analyse des «grammaires narratives» développées par Franzosi (1987) et ses collègues, qui appellent leur approche l'analyse quantitative narrative. Franzosi et ses collaborateurs quantifient un élément structurel de base de ce que le psychologue Bruner a appelé le «mode narratif» de l'expérience de commande. Cet élément est un processus cognitif social par lequel les gens interprètent des situations de toutes sortes en termes de relations sociales de base des acteurs, des actions et des objets d'action.

Leur système identifie les acteurs clés dans un ensemble de nouvelles et les actions qu'ils effectuent en analysant leur position dans le réseau global d'acteurs et d'actions, les séries chronologiques associées à certaines des propriétés des acteurs et en

générant des diagrammes de dispersion décrivant le sujet ou l'objet biais de chaque acteur, puis en enquêtant sur les types d'actions associées à chaque acteur.

Le sociologue Roberts et ses collègues ont développé une autre méthode mixte d'analyse narrative qu'ils appellent analyse de modalité. Semblable à certains égards aux grammaires narratives de Franzosi, l'analyse des modalités est destinée à la recherche comparative interculturelle et multilingue (Roberts, 2009). L'analyse de modalité évalue les langues en analysant les clauses modales dans plusieurs grandes collections de texte à partir de plusieurs langues afin d'identifier quelles activités les utilisateurs de chaque langue considèrent comme possibles, impossibles, inévitables ou contingentes.

Roberts et ses collègues ont utilisé leur méthode pour analyser les caractéristiques de nombreuses cultures différentes sur la base d'études de journaux arabes et hindi (Roberts, Zuell, Landmann et Wang, 2010) et de journaux hongrois (Roberts, Popping et Pan, 2009). Plus récemment, Mische (2014) a analysé des documents en ligne de la Conférence des Nations Unies sur le développement durable et du Sommet des peuples qui l'accompagnait à Rio de Janeiro en 2012. Lors de la lecture des documents, Mische et son équipe d'étudiants diplômés ont remarqué que les différents groupes qui ont participé dans les délibérations en ligne, ont utilisé différents éléments grammaticaux et narratifs. Son équipe a ensuite développé un schéma de codage pour analyser les formes verbales prédictives, impératives et subjonctives dans ces documents, qu'ils ont codés à la main à l'aide de NVivo .

Les chercheurs en gestion Gorbatai et Nelson (2015) ont utilisé le dictionnaire Linguistic Inquiry and Word Count (LIWC) (Tausczik & Pennebaker, 2010; voir l'annexe C) et les modèles thématiques (voir le chapitre 4, titre 9) pour examiner le rôle de la langue dans le succès des collectes de fonds en ligne, qui est une nouvelle forme de financement de projets entrepreneuriaux. Gorbatai et Nelson ont évalué l'influence du contenu linguistique sur les résultats de la collecte de fonds sur la base des données du site Web Indiegogo. Ils ont émis l'hypothèse que l'inégalité entre les sexes en faveur des femmes par rapport aux hommes dans la collecte de fonds en ligne s'explique en partie par les différences linguistiques entre les hommes et les femmes par rapport aux langues qu'ils utilisent. Ils ont analysé quatre dimensions différentes du contenu linguistique dans les descriptions de campagne, y compris le langage positif, la vivacité, le langage inclusif et le langage commercial. Les résultats de leur analyse ont montré un lien entre les choix linguistiques et les résultats de la collecte de fonds.

2.2. Analyse des thèmes

Dans l'analyse thématique, l'écriture fait partie intégrante de chaque phase d'un projet de recherche. L'écriture commence généralement dans les étapes initiales d'un projet avec la rédaction des idées et des schémas de codage potentiels, puis se poursuit par les phases de codage et d'analyse. L'analyse thématique commence lorsque le chercheur remarque des schémas de signification dans une collection de textes, soit pendant le processus d'acquisition des textes, soit peu de temps après.

Le point final de l'analyse thématique est le rapport du contenu et de la signification des modèles de thèmes dans les textes. Bien que l'analyse thématique permet aux chercheurs d'interpréter des thèmes et sous-thèmes généraux dans les textes, elle ne leur permet pas de faire des allégations sur l'utilisation du langage ou la «fonctionnalité fine de la conversation» (Braun et Clarke, 2006). Pourtant, les thèmes sont compris pour capturer quelque chose d'important au sujet des données textuelles par rapport à une ou plusieurs questions de recherche.

Dans l'analyse thématique, les thèmes peuvent être identifiés de manière inductive ou déductive. Une approche inductive ascendante signifie que les thèmes identifiés sont directement liés aux textes analysés (Patton, 2014). Dans cette approche, si le corpus a été construit spécifiquement pour le projet de recherche - par exemple, en transcrivant des entretiens ou des interactions de groupes de discussion - les thèmes identifiés peuvent avoir peu de rapport avec les questions spécifiques qui ont été posées aux participants.

L'analyse thématique n'est pas une méthode linéaire mais plutôt une méthode récursive où l'analyste se déplace selon les besoins tout au long des nombreuses phases d'un projet. Les premières étapes d'une analyse thématique consistent pour le chercheur à acquérir une collection de textes et à s'immerger dans les textes par lecture répétée (Braun & Clarke, 2006). Une telle immersion et lecture répétée implique de rechercher des thèmes tout en lisant attentivement les textes et en prenant des notes détaillées.

L’étape qui suit est le codage formel qui commence par la recherche de codes initiaux basés sur les notes prises lors de la lecture active et répétée des textes. Une question importante à résoudre en termes de codage est ce qui compte comme thème. Il s'agit de la fréquence d'apparition d'un thème dans chaque texte ainsi que dans l'ensemble de la collection de textes analysés. Bien qu'idéalement, il y aura un certain nombre d'instances du thème dans la collection de textes. Un nombre plus important d'instances ne signifie pas nécessairement que le thème est plus crucial.

C’est ainsi qu’il n'y a pas de réponses convenues à la question de savoir quelle proportion d'une collection de textes ou de documents doit présenter des preuves d'un thème pour qu'il soit considéré comme un thème principal ou global. L'importance relative d'un thème n'est pas nécessairement basée sur des mesures quantifiables mais plutôt sur le fait qu'il capture quelque chose d'important par rapport à la question de recherche globale (Braun et Clarke, 2006).

Dans ce cadre, les codes identifient une caractéristique du corpus qui est intéressante pour le chercheur et se réfèrent au « segment ou élément le plus élémentaire des données ou informations brutes qui peuvent être évaluées de manière significative concernant le phénomène» (Boyatzis, 1998, p. 63). Le codage peut être effectué manuellement ou à l'aide d'un logiciel spécialisé. Si vous codez manuellement, vous pouvez écrire des notes sur les textes eux-mêmes en utilisant des surligneurs, des stylos de couleur ou des notes autocollantes pour indiquer des motifs potentiels. Par contre, in logiciel d'analyse de données qualitatives (QDAS) vous permet d'effectuer ces fonctions numériquement. Cependant, ces données codées diffèrent des unités d'analyse, qui sont les thèmes développés dans la prochaine phase de l'analyse thématique.

Après un premier codage manuel ou assisté par logiciel, la phase suivante de l'analyse thématique commence lorsque tous les textes ont été initialement codés et que vous disposez d'une liste de différents codes identifiés dans les données. À ce stade, vous vous recentrerez sur des thèmes plutôt que sur des codes. Les différents codes sont triés en thèmes potentiels et des extraits de données codées sont rassemblés au sein des thèmes identifiés.

Il existe ainsi plusieurs techniques d'observation disponibles pour trier le texte codé en thèmes. Une technique consiste à identifier les thèmes en reconnaissant les répétitions dans le texte codé. C'est la technique utilisée par Strauss (1992) dans son analyse thématique des entretiens avec un ouvrier retraité évoquée au début de ce chapitre. Une autre technique consiste à identifier les termes locaux qui sont utilisés de manière inconnue. Ces catégories autochtones (Patton, 1990) peuvent fournir des informations sur les typologies et les schémas de classification de la communauté étudiée.

Tableau 1.3 : Phases d'une recherche d'analyse de thèmes

Commencer avec une question de recherche (déductive) ou des données (iductives)
Acquisition des données
Répétition de lecture
Codage
Tri en fonction des thèmes
Considérer la révision du schéma de code

Les techniques d'analyse thématique sont utilisées par les chercheurs dans les domaines des affaires, du conseil, de l'éducation, de la psychologie et de nombreux autres domaines.

2.3. L'analyse des métaphores

A. Théorie de la métaphore cognitive

L'affirmation fondamentale de la CMT est que le langage est structuré par une métaphore au niveau neuronal et que les métaphores utilisées dans le langage naturel révèlent des modèles cognitifs (ou «schémas») partagés par des membres de groupes sociaux. Ainsi, la métaphore est une structure centrale et indispensable de la pensée et du langage, et tout langage naturel est caractérisé par la présence d'expressions métaphoriques conventionnelles organisées autour de métaphores prototypiques, que Lakoff et Johnson ont qualifiées de métaphores conceptuelles.

Selon la CMT, les métaphores trouvent leur origine dans un processus de «réalisation phénoménologique» (Lakoff & Johnson, 1999, p. 46). Ils se forment lorsque les expériences perceptives et sensorielles construisent un domaine source incarné, tel que " pushing, pulling, supporting, balance, straight–curved, near–far, front–back, and high–low " sont utilisés pour représenter des entités abstraites dans un domaine cible (Boroditsky, 2000; Lakoff, 1987; Richardson, Spivey, Barsalou, & McRae, 2003).

Concepts basiques :

- La métaphore est une figure de style dans laquelle un mot ou une phrase est appliqué à quelque chose auquel il n'est pas littéralement applicable pour faire référence à un sens parallèle ;
- La théorie de la métaphore cognitive (CMT) est la théorie développée par les linguistes cognitifs qui affirme que les métaphores structurent le langage à un niveau neuronal et que les métaphores utilisées dans le langage naturel révèlent des modèles cognitifs partagés par les membres des groupes sociaux ;
- Les métaphores conceptuelles sont des métaphores prototypiques qui font référence à des expressions linguistiques du schéma de pensée conventionnel d'un groupe ou d'une société ;
- Le domaine source d'une métaphore est l'ensemble des expériences perceptuelles et sensorielles (par exemple, l'expérience sensorielle qu'une roue tourne dans la boue) sont utilisées pour représenter des entités abstraites.

Le domaine cible d'une métaphore est l'ensemble des entités abstraites (par exemple, l'idée de faire des progrès inadéquats) qui est représenté métaphoriquement par de riches expériences perceptives et sensorielles d'un domaine source.

B. Les approches de l'analyse de métaphores

Influencés par la CMT, les chercheurs ont développé un certain nombre de stratégies pour l'analyse des métaphores. Malgré que les trois stratégies discutées ci-dessus n'épuisent pas les possibilités d'analyser la métaphore dans la recherche en sciences sociales, ces derniers peuvent vous aider à commencer le développement un projet de recherche gérable (voir tableau 1.4).

Tableau 1.4 : Stratégies de conception de recherches en analyse de métaphores

	Exemples
Commencer par une anomalie	Ignatow,2003 ; Ignatow and Williams, 2011
Comparer deux groupes	Rees, Knight, and Wilikinson, 2007
Analyser une sous-culture	Schmitt,2000,2005

Une première stratégie de recherche pour l'analyse des métaphores consiste à commencer votre analyse par un exemple anormal ou inattendu de langage

métaphorique. Il s'agit essentiellement d'une méthode inductive. Par exemple, le sociologue Ignatow et le chercheur en communication Williams ont remarqué que vers 2010, la métaphore «bébé ancre» était entrée dans la culture populaire (l'expression fait référence aux enfants nés aux États-Unis de parents migrants sans-papiers qui auraient été conçus pour améliorer les chances de leurs parents d'obtenir la citoyenneté).

Cela semblait anormal car l'expression était largement considérée comme raciste et déshumanisante et avant environ 2007, elle était utilisée presque exclusivement sur des sites Web anti-immigrés à faible audience. Ignatow et Williams (2011) ont analysé les taux d'utilisation de l'expression au fil du temps et sur toutes les plateformes médiatiques à l'aide de la recherche avancée de Google et d'autres outils logiciels.

Une deuxième stratégie de recherche productive consiste à démarrer votre projet en choisissant deux ou plusieurs groupes à comparer pour tenter de répondre à une question de recherche. Cette stratégie est essentiellement déductive en ce qu'elle vous oblige à sélectionner systématiquement les groupes à comparer afin que vos données puissent être utilisées pour tester les hypothèses dérivées de votre question de recherche.

Une troisième stratégie d'analyse des métaphores, étroitement liée à la deuxième stratégie déductive utilisée par Rees et ses collègues (2007) est la stratégie sous-culturelle du psychologue Rudolf Schmitt. Schmitt (2000, 2005) a développé une méthode qualitative d'analyse de texte centrée sur la métaphore, une « approche basée sur des règles et étape par étape » qui est à son tour basée sur une logique inductive. La stratégie de Schmitt opère à un niveau d'analyse sociologique qui implique de faire des inférences sur la communauté qui génère le texte analysé.

C. Méthodes qualitatives, quantitatives et mixtes

Les méthodes peuvent être classées en méthodes qualitatives, méthodes mixtes (qualitatives et quantitatives) et méthodes quantitatives.

Des études d'analyse de métaphore utilisant les stratégies examinées précédemment ont été menées dans de nombreux domaines des sciences sociales ainsi que dans des domaines de recherche appliquée tels que l'éducation et la gestion. Dans cette section, nous passons en revue des études exemplaires publiées récemment dans ces divers domaines, en nous concentrant sur leurs méthodes et modèles de recherche.

D. Études sur les méthodes qualitatives

Les linguistes cognitifs ont effectué des études qualitatives sur les métaphores utilisées dans le langage naturel et dans les documents officiels. Par exemple, Charteris-

Black (2009, 2012, 2013) a développé une approche rhétorique de la métaphore connue sous le nom d'analyse critique de métaphore qui s'appuie sur des méthodologies et des perspectives développées en linguistique cognitive, linguistique de corpus et linguistique critique. Il a utilisé cette approche pour examiner les métaphores des domaines de la rhétorique politique, de la presse, de la religion et des communications d'un large éventail de dirigeants politiques. Il a également travaillé conjointement avec des sociologues sur la relation entre les récits de genre, de langue et de maladie.

L'analyse métaphorique qualitative a été effectuée dans de nombreux domaines des sciences sociales. Par exemple, les anthropologues ont commencé à analyser le langage métaphorique dans les années 1970 (Sapir et Crocker, 1977). La collection éditée par Fernandez en 1991, Beyond Metaphor, donne un bon aperçu des premiers travaux ethnographiques sur la métaphore. Pour un aperçu des recherches anthropologiques linguistiques plus récentes.

Dans la recherche en gestion, les linguistes Sun et Jiang (2014) ont utilisé l'outil de corpus Wmatrix pour étudier les métaphores utilisées dans les énoncés de mission des entreprises chinoises et américaines. Ils ont trouvé des différences dans l'utilisation des domaines sources pour trois métaphores conceptuelles conventionnelles : (1) les marques sont des personnes, (2) les affaires sont la coopération et (3) les affaires sont la concurrence. Ils ont également constaté que différentes identités et idéologies d'entreprise étaient associées à différents modèles d'utilisation des métaphores, l'identité d'entreprise chinoise étant davantage axée sur la concurrence tandis que l'identité d'entreprise américaine était davantage orientée vers la coopération.

O'Mara-Shimek, Guillén-Parra et Ortega-Larrea (2015) ont présenté le concept de marketing de solutions de crise (CSM) pour explorer comment la métaphore peut être utilisée pour présenter des informations afin de proposer des « solutions » à des « problèmes » construits de manière discursive dans les médias. O'Mara-Shimek et ses collègues (2015) ont exploré la relation entre le positionnement éditorial et l'idéologie dans les nouvelles financières et en examinant les métaphores utilisées pour décrire la nature du marché boursier dans les rapports en ligne sur le krach boursier de 2008. Dans le New York Times et le Wall Street Journal, des métaphores animées et biologiques décrivent le marché boursier «in terms of a living being that must be ‘nurtured’ through intervention as opposed to being ‘left alone,’ which is more consistent with laissez-faire approaches to economic crisis scenario»(O'Mara-Shimek et al., 2015, p. 103).

Gatti et Catalano (2015) ont analysé le processus d'apprentissage de l'enseignement d'un enseignant novice, Rachael, inscrit dans une résidence d'enseignants urbains aux États-Unis. L'analyse métaphorique critique (voir Charteris-Black, 2009, 2012, 2013) révèle des cadres contradictoires d'apprentissage pour enseigner, y compris les cadres d'enseignement est un voyage et l'enseignement est une entreprise.

L'analyse indique que les métaphores de serre sont utilisées pour attribuer certaines caractéristiques des sciences naturelles au changement climatique, des métaphores de jeu pour lutter contre les effets positifs du changement climatique et des métaphores de guerre pour mettre en évidence les effets négatifs du changement climatique. Le document a conclu en discutant les représentations métaphoriques contrastées et complémentaires utilisées par les magazines agricoles pour conventionner le changement climatique.

Les chercheurs en sciences de l'information Puschmann et Burgess (2014) ont utilisé la CMT pour évaluer les valeurs et les hypothèses codées dans le cadrage du terme big data. Leur analyse métaphore des actualités en ligne sur le méga-données révèle qu'elles sont conçues soit comme une force naturelle à contrôler, soit comme une ressource à consommer.

Dans des études médiatiques, Bickes, Otten et Weymann (2014) ont analysé la présentation médiatique allemande de la crise financière grecque, qui a provoqué un tumulte inattendu en Allemagne. Bickes et ses collègues (2014) ont examiné le rôle des médias dans la formation d'une opinion publique négative en Allemagne envers la Grèce. Analysant l'utilisation de la métaphore dans 122 articles en ligne, les chercheurs ont trouvé des différences remarquables dans l'évaluation et la présentation de la crise dans les médias Spiegel (Allemagne), The Economist (Royaume-Uni) et Time magazine (États-Unis).

Le sociologue Santa Ana (2002) a combiné CDA et l’analyse de métaphore. Les données de Santa Ana sont des journaux qu'il a utilisés pour étudier les représentations médiatiques de Latino / as aux États-Unis.

Les politologues ont analysé le langage métaphorique dans les documents de politique, les discours et autres textes politiques pour explorer les façons dont les métaphores assurent la médiation des relations entre les pays et les autres acteurs politiques. Par exemple, une collection éditée en 2004 par Beer et De Landtsheer (2004) comprend des chapitres sur les métaphores qui ont guidé et façonné la politique étrangère américaine dans l'arène publique depuis le début de la guerre froide.

E. Études sur les méthodes mixtes

Les sociologues ont développé un certain nombre de stratégies de méthodes mixtes pour l'analyse des métaphores. Généralement, cela implique le codage humain des métaphores en combinaison avec des tests statistiques pour la fiabilité inter-évaluateurs et les différences de taux d'utilisation des métaphores dans plusieurs collections de documents. Les collections de documents sont généralement produites par des groupes sociaux ayant des antécédents sociaux ou culturels différents. Lorsque l'analyse métaphorique qualitative est principalement inductive, la recherche sur les méthodes mixtes est principalement déductive, même si elle implique souvent une inférence adductive.

La psychologue sociale Moser (2000) a développé une méthode d'analyse de texte basée sur la métaphore qu'elle a appliqué dans ses recherches sur la psychologie du travail et des organisations. L'approche de méthodes mixtes de Moser implique de catégoriser les métaphores pour soi pendant les transitions de l'école au travail. Le concept de soi est très complexe et abstrait et il est donc souvent représenté par des métaphores. Les sujets étudiés par Moser étaient des étudiants suisses allemands qui ont participé à une étude par questionnaire sur leur transition prévue de l'université au travail.

Les psychologues cliniciens ont analysé les métaphores utilisées par les sujets en thérapie psychanalytique (Buchholz et von Kleist, 1995 ; Roderburg, 1998), et les psychologues cognitifs et expérimentaux ont étudié les métaphores comme exemples de modèles mentaux (Johnson-Laird, 1983). Mais en psychologie, seul Schmitt (2000, 2005) a développé une méthode qualitative d'analyse de texte centrée sur la métaphore. « L'approche fondée sur des règles et étape par étape » de Schmitt (2000) est idiographique et qualitative et repose sur une logique inductive. Il opère à un niveau d'analyse sociologique qui implique de faire des inférences sur la communauté qui génère le texte analysé et implique une stratégie de sélection de données et de collecte de documents multiples.

L'objectif de la méthode d'analyse systématique des métaphores de Schmitt (2005) est de « découvrir des schémas de pensée sous-cultures », et sa méthode y parvient en plusieurs étapes (p. 365). La première étape consiste pour le chercheur à choisir un sujet d'analyse. Schmitt donne l'exemple de l'abstinence de ses propres travaux empiriques sur les métaphores de l'abstinence et de l'alcoolisme. L'étape suivante consiste à rassembler une « vaste collection de métaphores contextuelles » pour

le sujet (Schmitt, 2005, p. 370). Ces métaphores peuvent être collectées à partir de sources telles que des encyclopédies, des revues et des livres spécialisés et généralistes.

Les chercheurs en gestion Gibson et Zellmer-Bruhn (2001) ont utilisé des méthodes mixtes d'analyse des métaphores pour étudier les concepts du travail d'équipe à travers les cultures organisationnelles nationales. Leurs recherches théoriquement orientées ont utilisé une logique déductive et ont présenté une conception de recherche avec plusieurs collections de documents. Il a fonctionné à un niveau d'analyse sociologique qui a permis aux chercheurs d'analyser des textes afin de connaître les organisations et la société qui les ont produits.

L'objectif de ce projet était de tester une théorie bien connue de l'influence de la culture nationale sur les attitudes des employés (Hofstede, 1980) avec un plan de recherche qui incluait la sélection stratégique du premier des quatre pays (France, Philippines, Porto Rico et États-Unis) puis de quatre organisations (p. 281). Les chercheurs ont mené des entretiens qu'ils ont transcrits pour former leurs corpus, et qu'ils ont analysés à l'aide de QSR NUD * IST (voir l'annexe D sur NVivo) et de TACT (Bradley, 1989; Popping, 1997). Ces progiciels ont été utilisés pour organiser le codage qualitatif de cinq métaphores de travail d'équipe fréquemment utilisées, qui ont ensuite été utilisées pour créer des variables dépendantes pour le test d'hypothèse à l'aide de logit multinomial et de la régression logistique.

F. Étude des méthodes quantitatives

L'analyse des métaphores, tant qualitatives que mixtes, dépend en fin de compte de l'interprétation humaine et du codage des métaphores dans les textes. Un tel codage est un sujet relatif à la fatigue du codeur, à sa polarisation et aux problèmes de fiabilité inter-évaluateurs de celui-ci. Cela prend également beaucoup de temps, et le temps requis pour la formation et le codage a jusqu'à présent limité la capacité des chercheurs à étendre l'analyse de métaphore pour une utilisation avec les méga données. Mais aujourd'hui, la situation évolue rapidement, plusieurs équipes de recherche en informatique et dans des domaines connexes développent des méthodes assistées par ordinateur pour détecter automatiquement les métaphores dans les textes.

Hardie, Koller, Rayson et Semino (2007) ont réorienté les outils d'annotation sémantique afin d'extraire éventuellement des phrases métaphoriques des textes. Turney, Neuman, Assaf et Cohen (2011) ont identifié des expressions métaphoriques en supposant que ces expressions consistent à la fois en un terme plus concret et plus abstrait. Ils ont dérivé un algorithme pour définir l'abstraction d'un terme, puis ont utilisé

cet algorithme pour contraster l'abstraction des phrases adjectif-nom. Les phrases ont été étiquetées comme métaphoriques lorsque la différence entre l'abstraction du nom et l'abstrait de l'adjectif a dépassé un seuil prédéterminé.

Récemment, Gandy, Neuman et leurs collègues (Gandy et al., 2013 ; Neuman et al., 2013) ont développé un certain nombre d'algorithmes interdépendants qui ont pu identifier le langage métaphorique dans des textes avec un haut niveau de précision. Leur travail est basé sur les idées clés de Turney et ses collègues (2011) selon lesquelles une métaphore implique généralement une mise en correspondance d'un domaine concret vers un domaine plus abstrait.

3. Les six approches de l'analyse de texte

Ces approches incluent l'analyse des conversations, l'analyse des positions de discours ainsi que l'analyse critique des discours (CDA), l'analyse du contenu, l'analyse foucaldienne et finalement l'analyse des textes en tant qu'informations sociales. Ces approches utilisent différentes stratégies logiques basées sur différentes fondations théoriques ainsi que des hypothèses philosophiques (Thomas, 2015). Ces approches opèrent également sur différents niveaux d'analyse : micro, méso et macro (Grossetti, 2006) et emploient ainsi différentes sélections de stratégies.

3.1. Approche 1 : L'analyse de conversations

Les analystes de conversation étudient les conversations quotidiennes des individus en évaluant la façon avec laquelle les gens négocient le sens de la conversation à laquelle ils participent et en évaluant également le discours le plus large de la conversation. Les analystes de conversation se concentrent non seulement sur ce qui est dit dans les conversations quotidiennes, mais aussi sur la manière d'utilisation de langage afin de définir les situations dans lesquelles se trouvent les participants de la conversation. Ces processus passent généralement inaperçus jusqu'à ce qu'il y ait un désaccord sur le sens d'une situation particulière.

Dans le but de mieux comprendre et assimiler le langage unique du milieu universitaire et de disciplines académiques spécifiques, Evison a identifié six éléments qui ont une relation très forte avec la position d'ouverture (mhm, mm, yes, laughter, oh, no) comme caractéristiques clés de des discours d'enseignement. D'autres exemples de recherche sur l'analyse de la conversation comprennent des études sur la conversation en milieu éducatif. Ces recherches ont été réalisés par O'Keefe et Walsh (2012) ; en

milieu de soins de santé par Heath et Luff (2000), Heritage et Raymond (2005) et Silverman (2016) ; et dans des environnements en ligne par Danescu-Niculescu-Mizil, Lee, Pang et Kleinberg (2012) qui font partie des éditeurs de Wikipedia.

3.2. Approche 2 : L'analyse des positions de discours

L'analyse des positions du discours est une approche de l'analyse de texte qui permet aux chercheurs de reconstruire les interactions communicatives à travers lesquelles les textes sont produits. À partir de cette approche, on peut mieux comprendre le sens du point de vue de chaque auteur. Les positions de discours peuvent être définis comme étant des rôles discursifs typiques basés sur le raisonnement que les gens adoptent dans leurs pratiques de communication quotidiennes. Dans le même contexte, l'analyse des positions de discours est un moyen de relier les textes aux espaces sociaux dans lesquels ils ont été émergé.

3.3. Approche 3 : L'analyse critique des discours

L'analyse critique des discours consiste à rechercher la présence de caractéristiques d'autres discours dans le texte à analyser. Cette approche est basée sur le concept de Fairclough (1995) "d'intertextualité", qui réfère à l'idée que les gens s'approprient des discours faisant partie de leur espace social à chaque fois qu'ils parlent ou écrivent. Dans l'analyse critique des discours, la parole et l'écriture ordinaires de tous les jours sont comprises afin de combiner les éléments de discours dominants. Alors que le terme discours se réfère généralement à toutes les pratiques d'écriture et de parole, les discours de cette approche sont considérés comme des façons d'écriture et de parole qui excluent des façons de construction des connaissances sur des sujets donnés.

3.4. Approche 4 : L'analyse de contenu

L'analyse de contenu adopte une approche quantitative et scientifique de l'analyse de texte. Contrairement à l'analyse critique de discours, l'analyse de contenu est généralement axée sur les textes eux-mêmes plutôt que sur les relations des textes avec leurs contextes sociaux et historiques. L'une des définitions classiques de l'analyse de contenu est la suivante : « une technique de recherche pour la description objective, systématique et quantitative du contenu de la communication » (Berelson, 1952, p. 18).

À un niveau pratique, l'analyse de contenu implique le développement d'un cadre de codage appliqué aux données textuelles. Il consiste principalement à décomposer les

textes en unités d'information pertinentes afin de permettre un codage et une catégorisation ultérieurs. L'analyse de contenu de manuels classiques de Krippendorff (2013) est la référence standard pour les travaux dans ce domaine. Un grand nombre de principes de conception de la recherche et des techniques d'échantillonnage abordés dans le Chapitre 2 de ce rapport de thèse sont partagés avec l'analyse de contenu.

3.5. Approche 5 : L'analyse foucaldienne

Le philosophe et historien Foucault (1973) a développé une conceptualisation influente de l'intertextualité qui diffère considérablement de la conceptualisation de Fairclough dans l'analyse critique de discours. Au lieu d'identifier l'influence des discours externes au sein d'un texte, pour Foucault le sens d'un texte émerge en référence aux discours avec lesquels il s'engage dans le dialogue. Ces engagements peuvent être explicites ou implicites. Dans l'analyse intertextuelle foucaldienne, l'analyste doit interroger chaque texte sur ses présupposés et avec quels discours il dialogue.

Par ailleurs, le sens d'un texte découle de ses similitudes et différences par rapport à d'autres textes, discours et présuppositions implicites dans le texte qui peuvent être reconnus par une lecture attentive historiquement informée. L'analyse foucaldienne des textes est réalisée dans de nombreux domaines de recherche théorique et appliquée. Par exemple, un certain nombre d'études ont utilisé l'analyse intertextuelle foucaldienne pour analyser la politique forestière (voir Winkel, 2012, pour un aperçu). Des chercheurs travaillant en Europe (par exemple, Berglund, 2001 ; Franklin, 2002 ; Van Herzele, 2006), en Amérique du Nord et dans les pays en développement (par exemple, Asher et Ojeda, 2009 ; Mathews, 2005) ont utilisé l'analyse Foucaldienne pour étudier les discours politiques concernant la gestion des forêts, les incendies des forêts et la responsabilité des entreprises.

3.6. Approche 6 : L'analyse des textes en tant qu'informations sociales

Cette catégorie d'analyse de texte traite les textes comme étant des reflets des connaissances pratiques de leurs auteurs. Ce type d'analyse est répandu dans les études théoriques fondées (voir les chapitres 2 et 3) ainsi que dans les études appliquées des discours d'experts. L'intérêt pour l'analyse informative des textes est dû en partie à sa valeur pratique, car les textes générés par les utilisateurs peuvent potentiellement fournir aux analystes des informations fiables sur la réalité sociale. Naturellement, la qualité des informations sur la réalité sociale contenue dans les textes varie selon le niveau de connaissance de chaque individu qui a participé à la création du texte, et les informations

fournies par les sujets sont partielles dans la mesure où elles sont filtrées par leur propre point de vue.

4. Les défis et limites de l’utilisation des données en ligne

Après avoir introduit les notions de fouille et analyse de texte, nous passerons en revue quelques leçons tirées d'autres domaines sur la meilleure façon d'adapter les méthodes de recherche en sciences sociales aux données provenant d'environnements en ligne. Cette section est d'une importance extrême pour les étudiants qui envisagent d'effectuer des recherches avec des données provenant de plateformes de médias sociaux et de sites Web. Les méthodologies telles que la fouille de texte qui analysent les données des environnements numériques offrent des avantages potentiels en termes de coûts et de temps par rapport aux méthodes plus anciennes (Hewson et Laurent, 2012 ; Hewson, Yule, Laurent et Vogel, 2003), car Internet offre un accès rapide à une panoplie de participants potentiellement vaste et géographiquement diversifié.

4.1. Les enquêtes sociales

Les enquêtes sociales sont l'une des méthodes les plus couramment utilisées en sciences sociales. Dans ce contexte, les chercheurs travaillent avec des versions en ligne des enquêtes depuis les années 1990. Les enquêtes téléphoniques et papier traditionnelles ont tendance à coûter cher, même en utilisant des échantillons relativement petits, et les coûts d'une enquête traditionnelle à grande échelle utilisant des questionnaires postés peuvent être énormes.

Bien que les coûts des logiciels de création d'enquêtes en ligne et des services d'enquêtes Web varient considérablement, en éliminant le besoin de papier, d'affranchissement et de saisie de données, les enquêtes en ligne sont généralement moins chères que leurs équivalents papier et téléphonique (Couper, 2000 ; Ilieva, Baron et Healey, 2002; Yun et Trumbo, 2000). Les sondages en ligne peuvent également faire gagner du temps aux chercheurs en leur permettant d'atteindre rapidement des milliers de personnes, même s'ils sont séparés par de grandes distances géographiques (Garton, Haythornthwaite et Wellman, 2007). Grâce à une enquête en ligne, un chercheur peut rapidement accéder à de vastes populations en publiant des invitations à participer à l'enquête dans des groupes de discussion ou des forums de discussion.

En plus de leurs économies de temps et d'argent et de leur commodité globale, les sondages en ligne présentent un autre avantage : ils exploitent la capacité d'Internet

à fournir un accès à des groupes et des individus qui seraient difficiles à atteindre autrement (Garton et al., 1997).

Bien que les enquêtes en ligne présentent des avantages significatifs par rapport aux enquêtes sur papier et par téléphone, elles entraînent de nouveaux défis en termes d'application des méthodes traditionnelles de recherche par sondage à l'étude du comportement en ligne. Les enquêteurs en ligne rencontrent souvent des problèmes concernant l'échantillonnage, car relativement peu de choses peuvent être connues sur les caractéristiques des personnes dans les communautés en ligne en dehors de certaines -variables démographiques de base, et même ces informations peuvent être discutables (Walejko, 2009).

4.2. L'ethnographie

Dans les années 1990, les chercheurs ont commencé à adapter les méthodes ethnographiques conçues pour étudier les communautés géographiquement situées à des environnements en ligne qui se caractérisent par des relations à médiation technologique plutôt qu'immédiate (Salmons, 2014). Le résultat est une ethnographie virtuelle (Hine, 2000) ou une netnographie (Kozinets, 2009), qui est l'étude ethnographique de personnes interagissant dans un large éventail d'environnements en ligne. Kozinets, un pionnier netnographique, soutient que la réussite de la netnographie oblige les chercheurs à reconnaître les caractéristiques uniques de ces environnements et à opérer un « changement radical » de l'ethnographie hors ligne, qui observe les gens, vers un mode d'analyse impliquant la recontextualisation des actes conversationnels (Kozinets, 2002, p. 64).

Parce que la netnographie fournit un accès plus limité aux marqueurs démographiques fixes que l'ethnographie, l'identité des participants est beaucoup plus difficile à discerner. Pourtant, les netnographes doivent en apprendre autant que possible sur les forums, les groupes et les individus qu'ils cherchent à comprendre. Contrairement aux ethnographies traditionnelles, dans l'identification des communautés pertinentes, les moteurs de recherche en ligne se sont révélés inestimables pour la tâche de se renseigner sur les populations de recherche (Kozinets, 2002, p. 63).

4.3. Les méthodes de recherche historiques

Les méthodes de recherche historiques sont parmi les méthodes les plus anciennes des sciences sociales. Les pères fondateurs de la sociologie - Marx, Weber et

Durkheim - ont effectué des études historiques basées sur la recherche archivistique, et aujourd'hui, les méthodes de recherche historiques sont largement utilisées par les historiens, les politologues et les sociologues.

Comme toutes les méthodes des sciences sociales, les méthodes de fouille de texte présentent des avantages et des inconvénients qui doivent être reconnus dès le départ et pris en considération à chaque étape du processus de recherche. Ainsi, les chercheurs en fouille de texte doivent être conscients des préoccupations des historiens concernant la qualité des données stockées dans les archives numériques et la possibilité pour les archives numériques d'encourager la passivité des chercheurs dans la phase de collecte des données de la recherche.

5. Conclusion

Ce chapitre a présenté les diverses méthodologies de fouille et d’analyse de texte en donnant un aperçu général sur les principales approches de l'analyse de texte et en discutant également les risques associés à l'analyse des données provenant de sources en ligne. Malgré ces risques, les chercheurs en sciences sociales et en informatique développent de nouveaux outils de fouille de texte et d'analyse de texte pour répondre à un large éventail de questions de recherche appliquée et théorique, dans le monde universitaire ainsi que dans les secteurs privé et public.

Chapitre 2 : Conception d'un projet de recherche de fouille du web

1. Introduction

Le design de recherche est l'un des plus importants et plus difficiles sujets en sciences sociales (Creswell, 2014 ; Gorard, 2013). Il concerne essentiellement l'architecture de base des projets de recherche, avec le design des projets comme des systèmes qui permettent à la théorie, aux données et aux méthodes de recherche de s'interfacer de manière à maximiser la capacité d'un projet à atteindre ses objectifs (voir Figure 2.1 : La triade de design de recherche).

Le design de recherche implique une séquence de décisions qui doivent être prises en considération au début d'un projet de recherche, lorsqu'une erreur ou une mauvaise décision peut conduire à des résultats non significatifs et non fiables.

Ainsi, il est extrêmement important de réfléchir prudemment et systématiquement au design de recherche avant de consacrer du temps et des ressources à l'acquisition de textes, à la maîtrise de logiciels ou de langages de programmation pour un projet de la fouille de texte web (Web Content Mining).

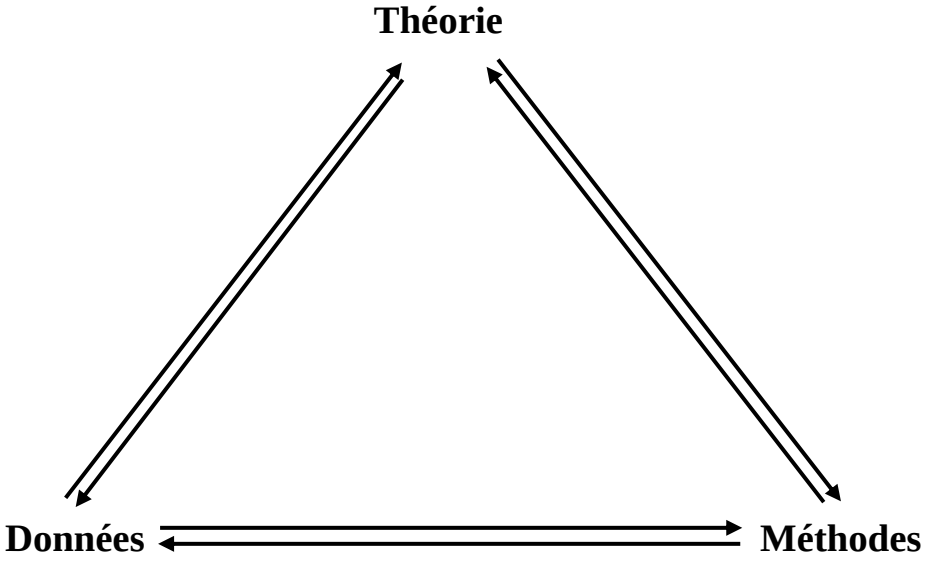


Figure 2.1 : La triade de design de recherche

Nous commençons ce chapitre par une revue des principes majeurs du design de recherche en sciences sociales applicables à la fouille de contenu Web (Web Content Mining) des projets de recherche. Nous discutons les différences entre les **approches idiographiques** et les **approches nomothétiques** de la recherche, entre la recherche effectuée en différents **niveaux d'analyse** et entre les méthodes de recherche qualitatives, quantitatives et mixtes.

Nous couvrons aussi les stratégies de sélection des données et puis examinons des exemplaires des projets d'analyse de texte qui utilisent diverses approches, qualitatives, quantitatives et mixtes, en combinaison avec diverses stratégies de sélection de cas et stratégies d'échantillonnage de données.

2. Décisions critiques

Le design de recherche s'intéresse à l'architecture de base des projets de recherche, avec la façon où un projet de recherche rassemble systématiquement la théorie, les données et une ou plusieurs méthodes d'analyse.

Le processus du design de recherche commence généralement par une question de recherche concernant le domaine social (Ravitch et Riggan, 2016).

Pour tenter de contribuer à la compréhension d'un phénomène social, les spécialistes en sciences sociales s'efforcent de générer deux formes principales de la connaissance fondée sur les preuves (Evidence-Based Knowledge) : les connaissances idiographiques et nomothétiques.

Les chercheurs travaillant avec des méthodes de la fouille de texte web conçoivent des projets qui opèrent à un niveau d'analyse textuel ou à un niveau contextuel, ou qui tentent d'identifier les relations sociologiques entre les textes et les contextes sociaux dans lesquels ils sont produits et reçus (Ruiz Ruiz, 2009).

Les designs de recherche sont souvent classés en utilisant des méthodes qualitatives, quantitatives ou mixtes, en utilisant aussi la collection des documents soit unique ou multiple, en employant une ou plusieurs méthodes spécifiques de sélection de cas et/ou d'échantillonnage de données, et en utilisant des méthodes inductives, déductives et/ou abductive.

La Figure 2.2 : Les décisions d’un design de recherche ci-dessous est une représentation des décisions d’un design de recherche qui sont généralement réalisés dans le cadre des projets de la fouille de texte en science sociale. Les décisions les plus

essentielles et les plus abstraites sont au sommet de la figure, et les décisions deviennent plus axées sur les données que l'on procède de 1 à 5. Mais, en fait, l'ordre des décisions présentées dans la Figure 2.2 : Les décisions d'un design de recherche n'est pas crucial parce que la figure est une simplification qui cache diverses relations de compatibilité, d'incompatibilité et d'interdépendance entre les décisions prises.

Néanmoins, bien que l'ordre dans lequel on prend ces décisions puisse être imprévisible et que nous devions réviser certains de nos figures à plusieurs reprises, chaque décision de la Figure 2.2 : Les décisions d'un design de recherche est d'une importance cruciale si notre projet de recherche veut atteindre ses objectifs.

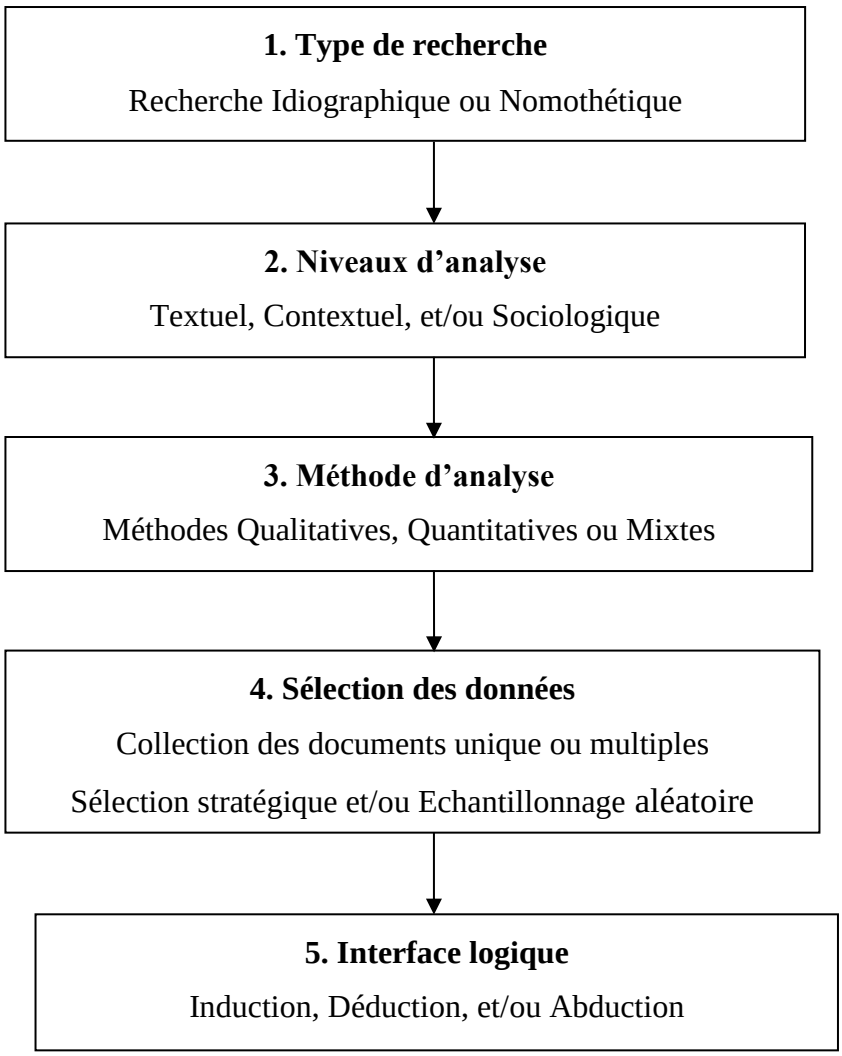


Figure 2.2 : Les décisions d'un design de recherche

3. Recherche idiographique et nomothétique

À la fin du 19ème siècle, le philosophe allemand Windelband (1894/1998) a inventé les termes **idiographiques** et **nomothétiques** pour désigner différentes formes de connaissances fondées sur des preuves. Pour Windelband (1901/2001), la connaissance idiographique impliquait la description et l'explication de phénomènes particuliers, tandis que la connaissance nomothétique consistait à trouver des généralités communes à une classe de phénomènes et à dériver des théories ou des lois pour rendre compte de ces généralités. Les formes de connaissances idiographiques et nomothétiques ne s'excluent pas nécessairement mutuellement. Et pourtant, au cours du siècle dernier, des méthodologies de recherche hautement spécialisées et parfois mutuellement antagonistes, se sont développées pour produire des connaissances sur les deux formes.

Les méthodes nomothétiques (ou positivistes) des sciences sociales tentent d'analyser de grands ensembles de données à l'aide de statistiques inférentiels, tandis que les méthodes idiographiques (post-positivistes telles que l'ethnographie se focalisent sur les détails des cas ou des événements.

Les méthodes de la fouille de texte web ont été utilisées efficacement à la fois pour les formes de recherche idiographiques (Kuckartz, 2014) et pour les recherches nomothétiques, bien qu'il existe des méthodes de la fouille de texte spécifiques, des progiciels et des langages de programmation qui conviennent un peu mieux à l'un ou l'autre.

4. Niveaux d'analyse

Bien qu'il puisse être difficile d'appliquer des méthodes d'exploration de texte à des applications de base, leur utilisation dans des projets de recherche en sciences sociales conduit inévitablement à de nouvelles sources de complexité. Une partie de cette complexité supplémentaire émane des multiples niveaux d'analyse qui sont endémiques à la recherche en sciences sociales. Ruiz Ruiz (2009) a utilement suggéré trois principaux niveaux d'analyse sur lesquels la recherche textuelle pourrait éventuellement fonctionner : le niveau du texte lui-même, le niveau du contexte social, et le niveau sociologique qui cherche à identifier les relations causales entre les textes et les contextes sociaux dans lesquels ils sont produits et reçus. La recherche sur la fouille de texte menée dans diverses disciplines fonctionne généralement à l'un ou l'autre de ces trois niveaux, et l'intéressement à ces niveaux vous permet de mieux comprendre les similitudes et les différences entre ces approches de recherche.

4.1. Le niveau textuel

L'analyse du texte en sciences sociales au niveau textuel « Concerne la caractérisation ou la détermination de la composition et de la structure du discours » (Ruiz Ruiz, 2009). La plupart des méthodes introduites dans ce rapport de thèse concernent l'analyse des textes au niveau textuel en termes de structures narratives et de thèmes de textes, de métaphores, de sujets et d'autres aspects de la composition et de la structure des textes eux-mêmes.

4.2. Le niveau contextuel

En plus de révéler des modèles dans les textes eux-mêmes, la fouille de textes peut également révéler des éléments du contexte social dans lequel les textes sont produits et reçus, y compris les contextes situationnels dans lesquels le discours a été produit et les caractéristiques des textes des auteurs. Les spécialistes des sciences sociales ont développé plusieurs méthodes d'analyse du texte situationnel, y compris l'analyse des positions du discours et l'analyse des conversations.

4.3. Le niveau sociologique

L'analyse des textes au niveau sociologique implique d'établir des liens entre les textes analysés et les espaces sociaux dans lesquels ils sont produits et reçus. Les textes ne peuvent être analysés au niveau sociologique que s'ils sont d'abord analysés au niveau textuel et au niveau contextuel. Bien que les liens entre les textes et leurs espaces sociaux puissent être très divers selon les questions de recherche de l'analyste et son orientation théorique, ils peuvent être classés en deux catégories : la recherche qui analyse les textes comme une information sociale et la recherche qui analyse les textes comme des reflets de l'idéologie de leurs auteurs et de leur public (Pour plus de détail, voir Ruiz Ruiz, 2009).

- **Textes comme une information sociale**

La première forme de la fouille de texte sociologique traite les textes comme des reflets de la connaissance pratique de leurs auteurs. Il est commun d'interpréter les textes comme des informations, ce qui peut être très utile pour les projets de recherche appliqués et académiques. Ce type d'analyse est répandu dans les études de la théorie ancrée (Grounded Theory) et dans les études appliquées des discours d'experts. L'intérêt général de l'analyse informative des textes consiste en sa valeur pratique, parce que les textes générés par l'utilisateur peuvent fournir aux chercheurs des informations valides

et pertinentes sur la réalité sociale. Parmi les exemples récents d'analyse informative de texte, nous pouvons citer des études menées par Trappey, Wu et leurs collègues sur la fouille de données des utilisateurs à partir des pages Facebook (Trappey, Wu, Liu et Lin, 2013 ; Wu, Liu et Trappey, 2014).

- **Textes comme des produits idéologiques**

En plus de leurs composants informatifs, les textes incluent des composants idéologiques. Dans la fouille de texte à orientation idéologique, ce qui intéresse le chercheur est le point de vue particulier de l'auteur, qui est généralement compris comme une indication de l'idéologie populaire. L'analyse idéologique des textes est une caractéristique de l'analyse du discours critique (CDA ; Van Dijk, 1993), qui vise à montrer comment les discours sociaux sont formés par des discours dominants projetés par des groupes et des institutions puissants. Des exemples d'études récentes du CDA comprennent des travaux sur les télé-soins et les appelants de Hakimnia, Holmström, Carlsson et Höglund (2014) et sur les communications des organisations par Merkl-Davies et Koller (2012).

Un autre programme de la fouille de texte à orientation idéologique est centré sur l'analyse de Bourdieu et Thompson (1991) de ce qu'ils appelaient les marchés linguistiques. Bourdieu analyse les textes comme des produits sociaux résultant de la position de leur auteur dans la société. Selon Bourdieu, les textes reflètent *l'habitus* de leurs auteurs, où *l'habitus* est compris dans ce contexte comme la compétence linguistique de l'auteur, qui découle de l'habitation d'une position spécifique dans la société et des expériences sociales qui sont conduits par cette position. Différentes manières de communiquer ont une valeur sociale différente sur les marchés linguistiques. La diversité des styles de communication impliquant les accents, la diction, la grammaire et le vocabulaire est considérée à la fois comme le reflet de l'inégalité sociale et comme un moyen de préserver et de reproduire l'inégalité sociale. Des exemples de la recherche de Bourdieusian sur la fouille de texte comprennent des études sur la production de statuts sur les plates-formes de contenu générées par les utilisateurs (User Generated Content) par Levina et Arriaga (2012) et des groupes de soutien outremangeurs en ligne par Ignatow (2009).

5. Les méthodes de recherche qualitative, quantitative et mixte

Dans tous les projets de la fouille de texte, le choix des outils méthodologiques est étroitement lié au type de connaissance fondée sur les preuves que le chercheur espère produire. Les chercheurs en sciences sociales qui souhaitent écrire des récits richement détaillés de phénomènes sociaux spécifiques s'appuient essentiellement sur des méthodes de recherche qualitatives telles que l'ethnographie.

Les chercheurs qui s'intéresses à produire des connaissances qui sont généralisables à des phénomènes autres que ceux qui font l'objet d'une enquête font généralement appel à des méthodes de recherche quantitatives et mixtes (Creswell, 2014 ; Tashakkori et Teddlie, 2010) qui incluent à la fois des éléments interprétatifs et quantitatifs.

Au sein des sciences sociales, les principales traditions de recherche en analyse de texte se divisent en méthodes qualitative et quantitative. **L'analyse du discours** est généralement considérée comme la tradition de recherche qualitative la plus influente en analyse de texte, tandis que **l'analyse de contenu** est la principale tradition de recherche basée sur le texte qui utilise des méthodes quantitatives et mixtes (voir Herrera et Braumoeller, 2004).

5.1. Analyse du discours

L'analyse du discours est une méthodologie d'analyse des phénomènes sociaux qualitative, idiographique et constructionniste. L'analyse du discours diffère des autres méthodologies qualitatives qui tentent de comprendre les processus de construction du sens en ce qu'elle cherche de découvrir les processus sociaux qui contribuent à la construction du sens partagé (Hardy, 2001 ; Phillips et Hardy, 2002). L'analyse du discours suppose que les discours n'ont pas de sens inhérent en eux-mêmes et que pour comprendre leurs effets constructifs, les chercheurs doivent les situer socialement et historiquement (Fairclough, 1995). Ainsi, l'analyse du discours découle de la conviction que le sens, et, par conséquent, la réalité sociale, découle de collections interdépendantes de textes qui apportent de nouvelles idées, objets et pratiques dans le monde. De cet perspective, les sciences sociales deviennent l'étude qualitative du développement de discours qui rendent la réalité sociale significative.

Parmi les exemples récents de recherches publiées sur l'analyse du discours, mentionnons des études sur le discours de l'après-guerre sur les droits de l'homme, qui a donné naissance à l'idée contemporaine d'un réfugié ayant des droits d'asile (Phillips et Hardy, 2002), et le discours sur le Sida, qui a renforcé les groupes de patients-activistes (Maguire, Hardy et Lawrence, 2004).

5.2. Analyse de contenu

Alors que l'analyse du discours utilise des méthodes qualitatives pour comprendre les relations entre les textes et leurs contextes sociaux et historiques, l'analyse de contenu adopte une approche plus orientée quantitative et positiviste (voir chapitres 1 et 2) impliquant le test d'hypothèse et l'analyse statistique (Schwandt, 2001).

Par conséquent, l'analyse de contenu est généralement focalisée sur les textes eux-mêmes plutôt que sur leurs relations avec leurs contextes sociaux et historiques. Berelson (1952) définit l'analyse de contenu comme "une technique de recherche pour la description objective, systématique et quantitative du contenu manifeste de la communication". Cette orientation quantitative a accompagné l'analyse de contenu jusqu'à aujourd'hui, bien que les méthodes statistiques utilisées soient devenues beaucoup plus compliquées. L'analyse de contenu se caractérise par un souci d'être objectif, systématique et de quantitatif (Kassarjian, 1977). Cette préoccupation repose sur la conviction que le sens d'un texte est constant et peut être connu avec précision et avec constance par différents chercheurs tant qu'ils utilisent des procédures analytiques rigoureuses et correctes (Silverman, 2016). Sur le plan pratique, l'analyse de contenu implique le développement de catégories analytiques qui sont utilisées pour construire un cadre de codage qui est ensuite appliqué aux données textuelles. Il consiste principalement à décomposer les textes en unités d'information pertinentes afin de permettre un codage et une catégorisation ultérieurs. Lorsque l'analyse du discours est hautement théorique, l'analyse de contenu est souvent considérée comme une méthode inductive.

Roberts (1997) a classé les techniques d'analyse de contenu dans les sciences sociales en trois catégories : techniques thématiques, techniques de réseau ou techniques sémantiques.

Les techniques thématiques de d'analyse de contenu se focalisent sur les significations manifestes dans les textes et incluent les méthodes couramment utilisées dans les affaires ainsi que dans les sciences sociales, telles que la modélisation de sujet.

Les techniques de réseau de l'analyse de texte modélisent les associations statistiques entre les mots pour déduire l'existence de modèles mentaux partagés par les membres d'une communauté. Les techniques sémantiques de l'analyse de texte, parfois appelées techniques herméneutiques ou herméneutiques structuralistes, comprennent une variété de méthodes conçues pour reconnaître les significations latentes dans les textes.

5.3. Méthodes mixtes

Si l'analyse du discours implique une analyse qualitative de la façon dont les contextes sociaux conditionnent la production de textes et l'analyse du contenu est une analyse quantitative des textes eux-mêmes sans tenir compte du contexte (Bauer, Bicquelet et Suerdem, 2014 ; Ruiz Ruiz, 2009), la plupart des recherches en sciences sociales qui utilisent des outils de la fouille de texte ne convient pas parfaitement dans l'une de ces catégories. De nombreux projets quantitatifs fonctionnent à des niveaux d'analyse contextuelles et sociologiques, tandis que de nombreux chercheurs soutiennent que toutes les recherches quantitatives sur la fouille de textes sont imprégnées par des évaluations qualitatives :

La prétendue dichotomique entre qualitatif et quantitatif est fausse parce que, premièrement, aucune quantification n'est possible sans qualification a priori et, deuxièmement, aucune explication quantifiable n'est possible sans une analyse qualitative postérieure. Dès le départ, tout projet de recherche sociale exige une notion de distinction qualitative entre les catégories sociales (ou, dans l'analyse textuelle, sémantique) avant que le chercheur puisse mesurer combien de mots appartiennent à l'une ou l'autre catégorie. De même, dans l'étape finale et qui est l'étape cruciale de toute analyse, c'est l'interprétation des extrants qui est la clé pour donner un sens à tout cela — et ici, plus un modèle statistique est complexe, plus l'interprétation des résultats est difficile. (Bauer, Gaskell et Allum, 2000)

Bauer et ses collègues (2000) ont clairement raison ici, et ils sont allés plus loin en problématisant la distinction qualitative-quantitative dans une publication de 2014, suggérant que cette distinction est superficielle (voir aussi Ruiz Ruiz, 2009) et n'a de valeur que pour des utilisations pédagogiques :

Elle permet aux programmes de cours et aux manuels de délimiter avec précision de nombreuses compétences et techniques qualitatives et quantitatives très spécifiques pour traiter divers types de données et de questions de recherche. Un étudiant confronté à un problème devrait ainsi le résoudre en connaissant et en appliquant simplement la

bonne technique quantitative ou qualitative ; et l'attente est que cela fonctionnera, comme la magie.

Les projets de recherche complexes sont souvent qualifiés soit qualitatifs ou quantitatifs pour revendiquer la supériorité putative d'une approche sur l'autre. Mais ces qualifications sont trompeuses dans la mesure où elles supposent que les deux approches de la recherche sont incompatibles. Bien que des méthodes particulières de la fouille de texte reposent sur une quantification à des degrés plus ou moins élevés, la plupart des méthodes de la fouille de texte sont mieux comprises comme des méthodes mixtes (Creswell, 2014 ; Teddlie et Tashakkori, 2008). La distinction qualitative-quantitative a sa place dans les manuels et les programmes et comme un raccourci pour catégoriser la recherche, mais les chercheurs ayant une expérience de travail avec les outils de la fouille de texte et d'analyse de texte savent que cette distinction peut être trompeuse.

6. Choix des données

Dans le chapitre suivant, nous investiguons les techniques disponibles pour l'acquisition de données textuelles à partir d'internet (voir aussi annexes **Erreur ! Source du renvoi introuvable.** et **Erreur ! Source du renvoi introuvable.** sur les sources de données et les logiciels pour la préparation et le nettoyage des données).

Mais avant l'acquisition des données, dans la plupart des recherches sur la fouille de texte (à l'exception des approches strictement inductives), l'analyste doit prendre en considération des décisions réfléchies concernant le choix des données. Le choix des données exige l'examen des préoccupations pratiques concernant le temps et le coût, ainsi que des préoccupations plus théoriques.

Là où les ethnographes et les chercheurs antécédents ont développé des procédures de sélection stratégique de cas pour les aider à choisir des données, et des chercheurs plus orientés quantitativement utilisent des méthodes d'échantillonnage statistique, la plupart des études de la fouille de texte et des projets de recherche d'analyse de texte nécessitent un mélange de procédures de sélection stratégiques et de méthodes d'échantillonnage.

Dans la pratique, de nombreux projets de la fouille de texte ne commencent pas par une question de recherche clairement formulée, mais plutôt par un ensemble de données. Les chercheurs rencontrent souvent des collections intéressantes ou uniques de documents qu'ils veulent utiliser pour un projet de recherche, et ils peuvent ne pas se

préoccuper du « problème potentiel d'avoir à annuler ou à compenser les biais de ces données » (Krippendorff, 2013, p. 122).

Le terme technique pour cette pratique est l'échantillonnage de commodité. Dans l'échantillonnage de commodité, la sélection des données telle qu'elle est discutée ci-après n'est pas applicable, le chercheur utilisera une logique inductive ou abductive, et l'échantillonnage des données est la prochaine étape cruciale. Inversement, les chercheurs commencent par une question de recherche - qu'il s'agisse d'une question théorique ou de fond- ils doivent examiner attentivement les stratégies de sélection des données afin de mettre en place un design de recherche qui peut potentiellement répondre à la question/les questions de recherche avant de passer au choix d'une stratégie d'échantillonnage de données (ou peut-être de stratégies multiples).

6.1. Sélection des données

La sélection et l'échantillonnage des données sont d'une importance centrale pour relier la théorie et les données dans tout projet de recherche empirique. Dans ce contexte, le terme *sélection de cas* est utilisé dans des projets de recherche avec des méthodes qualitatives et mixtes impliquant un petit nombre de cas où la généralisation des résultats de ces projets peut être augmentée par la sélection stratégique des cas. Dans cette section, nous empruntons le langage de la sélection des cas à partir des méthodes qualitatives et mixtes de recherche pour discuter la sélection des données (parfois appelée sélection de documents).

La sélection de cas critiques permet aux spécialistes des sciences sociales d'économiser du temps et de l'argent en étudiant un sujet donné en formulant une généralisation du formulaire : « Si c'est valable pour ce cas, il est valable pour tous (ou beaucoup) de cas. » Dans sa forme négative, la généralisation serait : « Si elle n'est pas valide pour ce cas, elle n'est pas valide pour les cas (ou seulement quelques cas) » (Denzin et Lincoln, 2011, p. 307). Un exemple de l'utilisation de cas critiques dans la fouille de texte est l'analyse de la métaphore de la méthode mixte de Gibson et Zellmer-Bruhn en 2001 sur les attitudes des employés dans quatre pays. Les quatre pays de cette étude ont été choisis stratégiquement pour la comparaison afin de maximiser les variations géographiques et culturelles pour que les résultats puissent être généralisés.

6.2. Échantillonnage de données

Une fois les sources de données pour le projet de recherche sont selectionnées, il peut être nécessaire de mettre en œuvre une stratégie d'échantillonnage de données. L'échantillonnage est souvent effectué dans le but de créer des échantillons représentatifs, d'une large population à partir de laquelle ils sont tirés. L'échantillonnage représentatif permet de généraliser les résultats de l'échantillon à cette population, et permet aux chercheurs de généraliser leurs résultats à la population par des inférences statistiques. Cependant, dans la recherche de la fouille de texte, il existe d'importants obstacles à l'obtention d'échantillons de probabilité représentatifs.

Krippendorff (2013) est allé jusqu'à faire valoir que le défi auquel sont confrontés les chercheurs d'échantillonnage d'une population en vue d'une autre « diffère radicalement des problèmes abordés par la théorie de l'échantillonnage statistique » (p. 114). Par exemple, un chercheur peut échantillonner les commentaires des utilisateurs sur une plate-forme de médias sociaux comme Facebook ou Twitter, mais il est presque impossible d'échantillonner de manière à pouvoir généraliser l'échantillon à l'univers entier des utilisateurs de Facebook ou Twitter. Il existe également des différences fondamentales entre les données textuelles et le type de données individuelles qui sont généralement utilisées dans des projets de recherche en sciences sociales à grande échelle tels que les enquêtes sociales.

En outre, dans la recherche de la fouille de texte, les unités échantillonnées sont rarement les unités comptées. Par exemple, un chercheur peut échantillonner des articles de journaux à partir d'une archive, mais compter des mots ou des co-occurrences de mots (plutôt que des articles). Même lorsque le chercheur s'installe sur des unités d'analyse pour l'échantillonnage et le comptage, et adopte des techniques d'échantillonnage de probabilité, des biais peuvent se produire. Certains groupes peuvent être systématiquement exclus de la collecte de données ou être sous-représentés en raison de biais d'auto-sélection. La collecte d'échantillons représentatifs à l'aide d'Internet est considérée très problématique depuis un certain temps (voir Hewson et Laurent, 2012).

Il est aussi possible d'utiliser l'échantillonnage stratifié, qui implique l'échantillonnage à l'intérieur des sous-unités, ou des strates, d'une population. Par exemple, un chercheur intéressé par les commentaires des lecteurs sur les articles de journaux pourrait créer un échantillon stratifié à partir des sections de journaux les plus populaires (p. ex., nouvelles mondiales, affaires, arts et divertissement, sports) et ensuite échantillonner au hasard ou systématiquement de chacune de ces strates. Bien sûr, la sélection des strates dépendrait des questions de recherche qui guideraient le projet. Une autre option possible, est l'utilisation des échantillons de probabilité variables pour

échantillonner proportionnellement à partir de sources de données ayant des tailles ou des niveaux d'importance différents, comme les journaux ayant des niveaux de diffusion différents (voir Krippendorff, 2013, p. 117).

L'échantillonnage de pertinence ou l'échantillonnage raisonné est une technique d'échantillonnage non probabiliste plus axée sur la question de recherche, dans laquelle le chercheur se renseigne sur une population d'intérêt, puis réduit progressivement le nombre de textes à analyser en fonction de leur pertinence pour la question de recherche. L'échantillonnage de pertinence est « si naturel qu'il est rarement discuté comme une catégorie à part » (Krippendorff, 2013, p. 121).

Lors de la conception d'un projet de recherche, il est tout naturel d'éliminer les données qui ne sont pas directement pertinentes à la question de la recherche. Par exemple, si le chercheur s'intéressie à des groupes ou des pages de femmes sur Facebook, il peut échantillonner les groupes de femmes et éliminer les groupes et les pages qui se concentrent sur les hommes. Un tel échantillonnage raisonné limite la représentativité de l'échantillon et la généralisation des résultats, mais dans de nombreux projets, il s'agira de préoccupations secondaires.

7. Agencement de vos données

Une fois que les données sont organisées sous forme de tableau, il est possible d'utiliser des statistiques descriptives (fréquences, moyennes et médianes) et de simples commandes de tableau croisé pour avoir une idée de la façon dont les données des documents sont liées aux questions de recherche. Ensuite, il faut appliquer des analyses de variance (ANOVA), des tests du *khi carré,* des régressions et d'autres tests et procédures statistiques qui permettront de tester les hypothèses dans les designs de recherche déductive.

8. Conclusion

Le design de recherche en sciences sociales ressemble beaucoup à l'architecture. Tout comme une planification approfondie est nécessaire avant le début des travaux de construction d'un bâtiment, de la même manière, il est important de réfléchir soigneusement et stratégiquement au design du projet de recherche dès ses premiers stades. Ce chapitre a eu pour but d'aider le chercheur à savoir où les décisions les plus critiques sont prises dans la recherche de la fouille de texte et à se familiariser avec divers modèles de recherche pour la recherche qualitative, quantitative et mixte.

Chapitre 3 : Méthodes de fouille de texte web

1. Introduction

Les méthodes de fouille de texte web aident à extraire des informations utiles et importantes à partir de formats de documents hétérogènes, tels que des pages Web, des courriels, des médias sociaux, des articles de magazines, etc. Cela se fait en identifiant les modèles dans les textes, tels que les tendances dans l'utilisation des mots, la structure syntaxique, etc.

Ces méthodes sont utiles a bien des égards. En effet, la fouille de texte peut aider à trouver des technologies nouvelles et innovantes dans certains domaines. Il s'agit d'une méthode très efficace pour générer de nouvelles informations et connaissances. Aussi, cette pratique permet de réduire le temps consacré à la lecture de longs textes et extraits littéraires. Cela permet également aux utilisateurs d'obtenir de nouvelles informations qui seraient autrement difficiles à trouver.

2. Collection des textes

2.1. Introduction

Tandis que les spécialistes des sciences sociales utilisent depuis des décennies les données des enquêtes d'attitude, les chercheurs tentent aujourd'hui de tirer parti du volume croissant de données non structurées naturelles générées par des personnes, telles que du texte ou des images. Certaines de ces données non structurées sont appelées « méga données ». Naturellement, l'utilisation d'ensembles de données textuelles permet de se renseigner sur les groupes sociaux et les communautés.

Sans doute, il existe des avantages et des inconvénients pour chacun, et il existe également des moyens de tirer parti des enquêtes et des méga données. Les enquêtes sont les mécanismes traditionnels de collecte d'informations sur les personnes, et il existe des domaines entiers qui se sont développés autour de ces instruments de collecte de données. Les enquêtes peuvent collecter des informations claires et ciblées, ainsi, les informations obtenues à partir des enquêtes sont beaucoup plus « propres » et beaucoup plus faciles à traiter par rapport aux informations extraites de sources de données non structurées.

La principale difficulté associée aux instruments d'enquête est le fait qu'ils sont coûteux à utiliser, à la fois en termes de temps et en termes de coûts financiers. L'alternative aux enquêtes qui a été largement explorée ces dernières années est l'extraction d'informations à partir de sources non structurées. Par exemple, plutôt que d'interroger un groupe de personnes sur leur optimisme ou leur pessimisme, en plus de demander leur emplacement, comme moyen de créer des cartes « d'optimisme », on pourrait atteindre le même objectif en collectant des données Twitter ou blog, en extrayant l'emplacement des écrivains à partir de leur profil, et en utilisant des outils de classification de texte automatique pour déduire leur niveau d'optimisme (Ruan, Wilson et Mihalcea, 2016).

2.2. Les sources de données en ligne

Les chercheurs préfèrent souvent utiliser des données prédéfinies plutôt que de construire leurs propres ensembles de données à l'aide d'outils d'exploration et de grattage. Alors que de nombreuses sources de données sont dans le domaine public, certaines nécessitent un accès via un abonnement universitaire. Par exemple, les sources de données d'actualités comprennent les sites Web des agences de presse locales et régionales ainsi que les bases de données privées telles qu'EBSCO, Factiva et LexisNexis, qui donnent accès à des dizaines de milliers de sources d'actualités mondiales, notamment des blogs, des transcriptions télévisées et radiophoniques, et les nouvelles imprimées traditionnelles.

Un exemple de l'utilisation de ces bases de données est une étude de la recherche universitaire sur l'entrepreneuriat international par les chercheurs en gestion Jones, Coviello et Tang (2011). Jones et ses collègues ont utilisé les outils de recherche EBSCO et ABI / INFORM pour sélectionner leur ensemble de données final de 323 articles de revues sur l'entrepreneuriat international publiés entre 1989 et 2009. Ils ont ensuite utilisé une analyse thématique (voir le chapitre 2, titre 2.2) pour identifier les thèmes et sous-thèmes dans leurs données.

2.3. Avantages et limites des ressources numériques en ligne pour les sciences sociales de recherche

L'utilisation des ressources numériques en ligne, et en particulier des médias sociaux, s'accompagne de ses avantages et inconvénients. Salganik (sous presse) a présenté un bon résumé des caractéristiques des méga données en général, dont beaucoup s'appliquent aux médias sociaux en particulier. Il a regroupé les

caractéristiques entre celles qui sont efficaces pour la recherche et celles qui ne le sont pas. Parmi les caractéristiques qui rendent les méga données efficaces pour la recherche, on peut citer :

a) Sa taille, qui peut permettre l'observation d'événements rares, des inférences causales et généralement un traitement statistique plus avancé qui n'est pas possible autrement lorsque les données sont petites :
b) Sa propriété « toujours active », qui fournit une dimension temporelle aux données et les rend aptes à étudier des événements inattendus et à produire des mesures en temps réel (par exemple, capturer les réactions des gens pendant une tornade, en analysant les tweets de la zone touchée) ;
c) Sa nature non réactive, ce qui implique que les répondants se comportent plus naturellement du fait qu'ils ne sont pas au courant de la saisie de leurs données (comme c'est le cas pour les enquêtes).

Ensuite, il existe également des caractéristiques qui rendent les méga données moins attrayantes pour la recherche, telles que :

a) Leur caractère incomplet – c'est-à-dire que souvent les collectes de données numériques manquent de données démographiques ou d'autres informations importantes pour les études sociales ;
b) Morceaux biais inhérent, dans la mesure où les contributeurs à ces ressources en ligne ne sont pas un échantillon aléatoire de personnes – par exemple, les personnes qui tweetent de nombreux tweets par jour par rapport à celles qui choisissent de ne jamais tweeter ; ils représentent différents types de populations ayant des intérêts, des personnalités et des valeurs différents, et même la plus grande collection de tweets ne capturera pas les comportements de ceux qui n'utilisent pas Twitter ;
c) Son évolution dans le temps, en termes d'utilisateurs (qui génère les données des médias sociaux et comment elles les génèrent) et de plates-formes (comment les données des médias sociaux sont-elles capturées), ce qui rend difficile la réalisation d'études longitudinales ;
d) Enfin sa sensibilité aux confusions algorithmiques, qui sont des propriétés qui semblent appartenir aux données étudiées et qui sont en fait causées par le système sous-jacent utilisé pour collecter les données – comme dans le nombre apparemment magique de 20 amis que beaucoup de gens semblent avoir sur Facebook, ce qui s'avère être un effet de la plateforme Facebook qui encourage

activement les gens à se faire des amis jusqu'à ce qu'ils atteignent 20 amis (Salganik, sous presse).

De plus, certains types de données numériques sont inaccessibles, par exemple les courriels, les requêtes envoyées aux moteurs de recherche, les appels téléphoniques, etc., ce qui rend difficile la recherche de comportements associés à ces types de données.

2.4. Exemples de recherche en sciences sociales utilisant des données numériques

Il existe des exemples d'études de recherche en sciences sociales qui utilisent les données des médias sociaux. Pour utiliser les données Facebook pour un propre projet, il est important de passer en revue les études sur la controverse éthique sur Facebook. De plus, la recherche de la sociologue Hanna (2013) sur l'utilisation de Facebook pour étudier les mouvements sociaux peut être un point de départ utile. Hanna a examiné les procédures d'analyse des mouvements sociaux tels que les mouvements du printemps arabe et d'occupation en appliquant des méthodes de fouille de texte aux données Facebook. Hanna utilise le Natural Language Toolkit (NLTK ; www.nltk.org) et le package R ReadMe (http ://gking.harvard.edu/readme) pour analyser les schémas de mobilisation du mouvement de jeunesse du 6 avril en Égypte. Il a corroboré les résultats de ses méthodes de fouille de texte avec des entretiens approfondis avec les participants au mouvement.

Pour utiliser les données Twitter, deux études d'analyse thématique basées sur Twitter (voir chapitre 2, titre 2.2) sont de bons points de départ. Le premier est une étude du chat en direct sur Twitter des Centers for Disease Control and Prevention menée par Lazard, Scheinfeld, Bernhardt, Wilcox et Suran (2015). L'équipe de Lazard a collecté, trié et analysé les tweets des utilisateurs pour révéler les principaux thèmes de préoccupation du public concernant les symptômes et la durée de vie du virus, le transfert et la contraction de la maladie, les voyages en toute sécurité et la protection de son corps.

Lazard et son équipe ont utilisé SAS Text Miner (www.sas.com /en_us/software/analytics/text-miner.html) pour organiser et analyser les données Twitter. Une deuxième étude d'analyse thématique qui utilise des données Twitter a été réalisée par les chercheurs en santé mentale Shepherd, Sanders, Doyle et Shaw (2015). Les chercheurs ont évalué la façon dont Twitter est utilisé par des personnes ayant des problèmes de santé mentale en suivant l'hashtag #dearmentalhealthprofessionals et en

effectuant une analyse thématique pour identifier les thèmes de discussion communs. Ils ont trouvé 515 communications uniques liées à la conversation spécifiée.

Ils ont constaté comme résultat 515 communications liées à la conversation spécifiée. La majorité du matériel se rapportait à quatre thèmes principaux :

1) L'impact du diagnostic sur l'identité personnelle et en tant que facilitateur pour l'accès aux soins,
2) L'équilibre des pouvoirs entre le professionnel et l'utilisateur de services,
3) La relation thérapeutique et le développement de la communication professionnelle.
4) Et soutenir la prestation par le biais des médicaments, de la planification de crise, de la prestation de services et de la société en général.

3. Pré-traitement

3.1. Introduction

L'analyse de texte nécessite une certaine forme de traitement adapté aux traitements de texte. Prenons l'exemple d'un tweet : "Aujourd'hui est le grand jour, mesdames et messieurs. M. K atterrira aux États-Unis :)". Si l'on veut utiliser les informations de ce morceau de texte pour toute forme d'exploration de texte ou autre analyse de texte, il est important de déterminer quels sont les jetons dans ce texte - aujourd'hui, est, le, grand, jour, mesdames et messieurs, Mr., K, atterrira, aux, États-Unis , :) - ce qui implique un processus qui assimile que les périodes dans les abréviations (par exemple, Mr.) et les acronymes (par exemple, US) doivent être préservées telles qu'elles sont, mais il y a aussi la ponctuation qui doit être séparée des jetons voisins (virgule après jour ou période après messieurs). De plus, un préprocesseur de texte normalise souvent le texte, il peut essayer d'identifier la racine ou la racine des mots (par exemple, "lady" for "ladies", or "be" for " 's"), et il peut même tenter d'identifier et étiqueter des symboles spéciaux tels que cet émoticône : ":)".

Le traitement de texte consiste généralement en des étapes de base telles que la suppression des balises HTML d'une collection de documents collectés sur le Web, la séparation de la ponctuation des mots, la suppression des mots de fonction et l'application de la racine ou de la lemmatisation. Le traitement peut également prendre d'autres formes plus avancées telles que l'annotation du texte avec des balises de partie du discours, des arbres de dépendances syntaxiques ou d'autres couches d'annotations

telles que le mappage de mots aux sens dans un dictionnaire et la recherche de marqueurs de discours.

Dans ce contexte, ce chapitre couvrira certaines des étapes de base du traitement de texte tout en abordant les bases des modèles de langage et fournissant des pointeurs pour une lecture supplémentaire dans l'objectif de réaliser un traitement de texte plus avancé.

Le traitement de texte de base est généralement la première étape de toute étude de recherche impliquant une entrée linguistique. Il suffit alors de supprimer les balises superflues (par exemple, HTML, XML) et de symboliser la ponctuation, ce qui se traduira par un ensemble de jetons qui peuvent être utilisés pour collecter des statistiques ou utiliser comme entrée pour d'autres applications telles que l'analyse de sentiments ou la classification de texte.

Concepts de base

- **La tokenisation** se réfère au processus de séparation de la ponctuation des mots tout en conservant leur signification prévue.
- **La suppression des stop word** est le processus d'élimination des mots de fonction tels que le, le, le, etc.
- **La racinisation** est une étape de traitement qui utilise un ensemble de règles pour supprimer les inflexions.
- **La lemmatisation** est le processus d'identification de la forme de base (ou forme racine) d'un mot.
- **Les modèles de langage** sont des représentations probabilistes du langage.
- **La répartition des mots** dans la langue est effectivement modelée par deux lois : la loi de Zipf et la loi de Heaps.

D'autres étapes de traitement de texte plus avancées qui sont souvent utilisées dans l'exploration de texte sont le balisage de partie de la parole, l'identification de la colocalisation, l'analyse syntaxique, la reconnaissance d'entité nommée (NER), la désambiguïsation du sens du mot et la similitude des mots.

3.2. Le traitement de texte de base :

Le traitement de texte de base est souvent la première étape de toute application d'exploration de texte et se compose de plusieurs couches de traitement simple telles que la tokenisation, la lemmatisation et la normalisation ou des processus plus avancés tels

que le balisage d'une partie du discours, l'analyse syntaxique et autres. L'annexe H cite quelques logiciels populaires de traitement de texte.

- **La tokenisation**

La tokenisation est le processus d'identification des mots dans la séquence d'entrée de caractères, principalement en séparant les signes de ponctuation mais également en identifiant les contractions, les abréviations, etc. Par exemple, étant donné le texte « M. Forgeron n'aime pas les pommes ", nous aimerions une sortie qui a chaque mot comme un jeton distinct, comme dans « M. Smith n'aime pas les pommes. » La tokenisation peut sembler un processus trivial dans un premier temps, mais certains cas nécessitent une attention particulière. Par exemple, pour une période, nous devons faire la distinction entre la période de fin de phrase et les marqueurs d'acronymes (par exemple, U.S) ou d'abréviations (par exemple, M., Dr.). Bien que nous voulions séparer la fin de la phrase du mot précédent, il est préférable de conserver la période attachée aux acronymes ou aux abréviations, car ce sont des mots qui nécessitent que la période soit bien formée. La période a également une signification particulière et doit être conservée telle quelle à l'intérieur des chiffres (par exemple, 12.4) ou des dates (par exemple, 12.05.2015) ou des adresses IP (100.2.34.58).

- Pour une apostrophe, nous cherchons souvent à identifier les contractions et les séparer de sorte qu'elles forment des mots individuels significatifs. Par exemple, le mot possessif "book's" doit former deux mots "book" et "is". Les contractions "aren't" et "he's" doivent être séparés en "are not" et "he is".
- Les citations doivent également être séparées du texte, comme dans, par exemple, « Let it be » qui devrait être transformé à « let it be ».
- Pour les césures, nous les laissons souvent en place, pour indiquer une collocation comme dans, par exemple, "state-of-the-art", bien qu'il soit parfois utile de la séparer, pour permettre l'accès à des mots individuels, par exemple, séparer « Hewlett-Packard » à « Hewlett-Packard ».

Bien que la tokenisation soit largement indépendante de la langue, plusieurs cas particuliers qui doivent être traités pour une tokenisation correcte peuvent être spécifiques à la langue. Par exemple, les abréviations et les contractions dépendent souvent de la langue utilisée, et donc, il faut compiler une liste de ces mots pour s'assurer que la tokenisation de la période est gérée correctement. Il en va de même pour l'apostrophe et la césure.

Parfois, le processus de tokenisation comprend également d'autres étapes de normalisation du texte, telles que la minuscule ou la mise en forme plus avancée (par exemple, en sélectionnant la casse correcte pour le mot « apple » dans la phrase suivante « There is an apple symbol on my Apple Macbook ») ou la suppression des balises HTML, si le texte est obtenu à partir d'une page Web. Notez que le processus de tokenisation suppose que les espaces blancs et la ponctuation sont utilisés comme limites explicites des mots. C'est le cas de nombreuses langues utilisant l'alphabet latin, ainsi que de plusieurs autres familles de langues, mais ce n'est pas le cas de la plupart des langues asiatiques. Le traitement d'une source d'entrée de caractères dans une langue asiatique nécessite également une étape distincte de détection des limites de mots, qui est souvent effectuée à l'aide d'algorithmes d'apprentissage supervisé. Il existe également des langues qui font un usage intensif de mots composés tels que l'allemand (*Computerlinguistik* means "computational linguistics") ou des langues agglutinantes telles que l'inuktitut (Tusaatsiarunnanngittualuujunga signifie « je ne peux pas très bien entendre").

- **La suppression des Stop Word**

Les stop words, également appelés mots de fonction ou mots de classe fermée, sont constitués de mots à haute fréquence comprenant des pronoms (par exemple, je, nous), des déterminants (par exemple, le, la), des prépositions (par exemple, dans, sur), et d'autres. Pour certaines tâches, les stop words peuvent être utiles : par exemple, il a été constaté qu'ils peuvent donner un aperçu significatif de la personnalité et des comportements des personnes (Mihalcea & Strapparava, 2009; Pennebaker & King, 1999).

Mais il y a aussi des tâches quand il est utile de les supprimer et de concentrer l'attention sur les mots de contenu tels que les noms et les verbes. Dans tous les cas, il est important d'avoir les moyens d'identifier les stop words dans un texte saisi. En général, cela se fait à l'aide d'une liste précompilée de stop words avec un algorithme de recherche efficace.

Les stop words dépendent clairement de la langue ; ainsi, une question importante est de savoir comment créer une liste de stop words pour la langue d'intérêt. Les langues bien étudiées, comme l'anglais, l'espagnol ou le chinois, ont plusieurs de ces listes accessibles au public. Si une liste de stop words n'est pas disponible pour une langue donnée, en se basant sur le fait que les stop words sont des mots de haute fréquence (voir la section suivante), on peut rassembler des statistiques de mots sur un très grand corpus

de textes écrits dans cette langue et par conséquent, obtenir les N premiers mots les plus fréquents comme candidats stop words.

Idéalement, le corpus devrait contenir un mélange de textes de différents domaines, pour éviter la fréquence élevée de certains mots en raison de leur spécificité de domaine (par exemple, une collection de textes sur l'informatique inclura évidemment le mot ordinateur à haute fréquence). Il est recommandé d'obtenir également les commentaires d'un locuteur natif sur la liste des stop words candidats, car il peut parfois inclure des mots fréquents mais pas des mots vides (par exemple, avoir, obtenir, aujourd'hui).

- **La racinisation et la lemmatisation**

De nombreux mots en langage naturel sont liés, mais ils ont différentes formes de surface, ce qui rend leur reconnaissance non triviale. Bien que certaines de ces relations soient de nature sémantique et nécessitent des connaissances de dictionnaire, comme dans le cas, par exemple de “sick“ et “ill“, il existe également de nombreuses relations qui peuvent être plus facilement capturées par des formes plus simples d'analyse de chaînes, comme le cas avec “construction“ et “construct“ ou même “water“ et “watered“.

La façon la plus simple d'identifier la racine commune de plusieurs mots est d'utiliser le processus appelé la racinisation. Autrement dit, la racinisation applique un ensemble de règles à un mot d'entrée pour supprimer les suffixes et les préfixes et obtenir sa racine, qui sera désormais partagée avec d'autres mots apparentés. Par exemple, “computer“, “computational“ et “computation“ qui seront tous réduits à la même racine : compute.

La racinisation produit souvent des racines qui ne sont pas des mots valides, ce qui n'est pas pertinent si le « consommateur » de ces racines est un système et non un être humain. Par exemple, la racinisation est utilisée dans la recherche d'informations (voir chapitre 4, titre 3), où les racines sont introduites dans le processus d'indexation et améliorent la qualité du système de recherche d'informations, sans jamais être lues par les utilisateurs de ce système. La racine, cependant, ne doit pas être utilisée si le texte dérivé doit être lu par un humain, car il est souvent difficile à comprendre. Considérez, par exemple, le texte « for example compressed and compression are both accepted as equivalent to compress », qu'un “stemmer“transformera en « for example compres and compres are both accept as equival to compres.” ».

Il existe de nombreux “stemmers“, le plus populaire étant “the Porter Stremmer“, qui est une procédure simple pour supprimer les affixes connus en anglais sans utiliser de dictionnaire. Le “Porter Stemmer“ consiste en un ensemble de règles de transformation, telles que “sses → ss, ies → i, ational → ate, tional → tion“, qui sont appliquées de manière répétée sur un mot jusqu'à ce qu'aucune transformation ne soit obtenue. Il a été constaté que le “stemmer“ fonctionnait bien dans les évaluations effectuées dans les systèmes de recherche d'informations, où la qualité du système de recherche a été améliorée lorsqu'il était appliqué sur du texte en racine. Il fait également des erreurs, y compris des erreurs de « commission », telles que “organization“ et “organ“ qui sont tous deux à l'origine du mot “organ“ ou meme les termes “police“ et “policy“ ayant en commun la racine “polic“, ou des erreurs « d’omission », comme “cylinder“ et “cylindrical“ ou “Europe“ et “European“. Le stemmer dépend clairement de la langue, mais il existe également des versions du stemmer pour plusieurs langues autres que l'anglais.

L'alternative de la racinisation est la lemmatisation, qui réduit les formes flexionnelles d'un mot à sa forme racine. Par exemple, la lemmatisation transformera “boys“ à “boy“, “children“ à “child“, et “am“, “are“, ou “is“ à “be“. Contrairement à la racinisation, la sortie obtenue de la lemmatisation est une forme de mot valide, qui est la forme de base d'un mot telle que trouvée dans un dictionnaire. Ainsi, la lemmatisation a l'avantage que sa sortie est lisible par l’homme ; cependant, cela se fait au prix d'un processus plus intensif en calcul, car il nécessite une liste de formes grammaticales pour gérer les inflexions régulières ainsi qu'une longue liste de mots irréguliers.

3.3. Modèles de langage et statistiques de texte

- **Modèles de langage**

Les modèles de langage sont des représentations dites probabilistes du langage naturel, qui peuvent être utilisées comme étant des modèles prédictifs ou explicatifs. Un modèle de langage est destiné à capturer la probabilité de voir certaines séquences de mots ou de caractères. À titre d’exemple, étant donné que nous avons déjà vu le mot “chien“, est-il plus probable que nous verrons le mot “aboiements“ ou “écritures“ ?

Les modèles de langage peuvent ainsi être utilisés pour proposer des alternatives possibles en se référant à l’historique des mots. En prenant toujours l’exemple du mot “chien“, Quelles seront les suites possibles ? Peuvent- ils être utilisées pour évaluer le

taux de probabilité ? Par exemple, quelle est la probabilité de voir la séquence “aboiement des chiens“ ?

Les modèles de langage ont diverses applications, nous pouvons énumérer les exemples d’applications suivants :

- Les corrections d'orthographe (Quelles sont les corrections les plus probables ? en se référant à un texte incorrect observé).
- La reconnaissance vocale (Quel est le texte le plus probable parmi tous les textes possibles qui pourraient résulter d'un énoncé d'entrée ?).
- La traduction automatique (Quelle est la traduction la plus probable, compte tenu de toutes les alternatives possibles ?).
- La reconnaissance de l'écriture manuscrite.
- La reconnaissance linguistiques, etc…

Les modèles de langage sont construits sur de larges corpus (appelés corpus de formation) à partir du calcul des probabilités de mots ou de séquences de mots sur une collection de texte donnée. Évidemment, plus la collection est grande, plus les modèles linguistiques sont précis. Dans ce contexte, on peut penser à prédire la probabilité de voir la séquence “un chien mange“ en traitant le texte de ce chapitre par rapport au traitement du texte disponible dans quelques millions de pages Web.

Les modèles de langage les plus simples sont basés sur des uni-grammes, où nous avons des probabilités associées à des mots individuels. Dans ces modèles, nous comptons principalement la fréquence des mots individuels et ensuite, nous cherchons à calculer leur probabilité par rapport à l'ensemble des mots du corpus d'apprentissage. À titre d’exemple, si nous avons une collection de 100 mots, dont deux sont des “chien“, la probabilité de voir ce mot individuel (ou uni-gramme) est ***P (chien) = 2/100*.**

L'étape suivante consiste à créer des modèles de langage bi-gramme. Nous mesurons dans cette étape la probabilité d'un mot donné par rapport au mot précèdent :

P (Wi | Wi − 1), qui peut être calculé comme ***Count (Wi − 1 Wi) / Count (Wi − 1)***. Par exemple, nous pourrions calculer ***P (aboiements | chien)*** comme le nombre de fois où nous avons vu des “aboiements de chien“ divisé par le nombre de fois où nous avons vu des “chien“.

Nous identifions ensuite les modèles de tri-grammes, où nous estimons

P (Wi | Wi − 1 Wi − 2), puis quatre-grammes, où nous estimons ***P (Wi | Wi − 1 Wi − 2 Wi − 3)*** ainsi de suite. Plus l'ordre des modèles à n-grammes est

élevé, plus la précision des modèles est meilleure et plus efficace, telle que mesurée par leur pouvoir prédictif ou explicatif. Cependant, le compromis et la complexité résident dans la quantité de données nécessaire pour former de tels modèles avec des n-grammes d'ordre supérieur qui nécessitent des ensembles de formation beaucoup plus grands afin d'éviter la rareté des données.

Par exemple, on peut calculer les probabilités de mots individuels à partir d'un corpus de 1 million de mots par rapport au calcul de la probabilité de séquences de six mots à la fois (six- grammes) à partir du même corpus. Il est probable que nous aurions vu beaucoup de mots individuels dans ce corpus, mais il est également probable que nous n'avons pas vu la plupart des séquences possibles de six mots dans ce même corpus. L'effet de cela est que nous aurons beaucoup de comptes nuls dans nos estimations de probabilité de six-grammes, et par conséquent, le modèle finira par ne pas être précis.

Compte tenu de ces modèles de langage, nous pouvons les combiner pour faire des prédictions pour des textes entiers. Par exemple, si nous avons le texte « Je veux manger de la nourriture chinoise », en supposant un modèle bi-gramme, nous pouvons calculer P (je veux manger de la nourriture britannique) = ***P (je | start) P (veux | je) P (manger| veux) P (de | manger) P (la | de) P (nourriture | la) P (britannique | nourriture).*** De même, avec un modèle de trigramme, nous calculerions ***P (je | start start) P (veux| start je) P (manger | je veux) P (de| veux manger) P (la | manger de) P (nourriture | de la), P (britannique | la nourriture)*** et ainsi de suite …

- **Statistiques de texte**

L'une des analyses les plus simples que l'on puisse faire sur n'importe quelle collection de texte est de compter le nombre de mots et de déterminer quels mots apparaissent avec une fréquence plus élevée. Le processus est intéressant mais assez complexe compte tenu que le langage naturel est très prévisible : on peut, par exemple, faire une bonne supposition quant aux mots qui seront les plus fréquents dans toute collection de texte ou faire des prédictions par rapport à la taille du vocabulaire (c.-à-d., nombre de mots uniques) dans une nouvelle collection. Par exemple le Tableau 3.1 montre les 10 mots les plus fréquemment utilisés dans l'une des collections de référence de la ***« Text Retrieval Conference (TREC 3) »***. Comme indiqué dans ce tableau, les mots les plus fréquents sont généralement des stop words tel qu'on les a définis dans les paragraphes précédents. Si nous traçons la fréquence des mots dans un corpus, nous obtenons généralement une courbe qui ressemble à celle montrée dans la Figure 3.1.

Figure 3.1 : La distribution des fréquences de mots dans un corpus

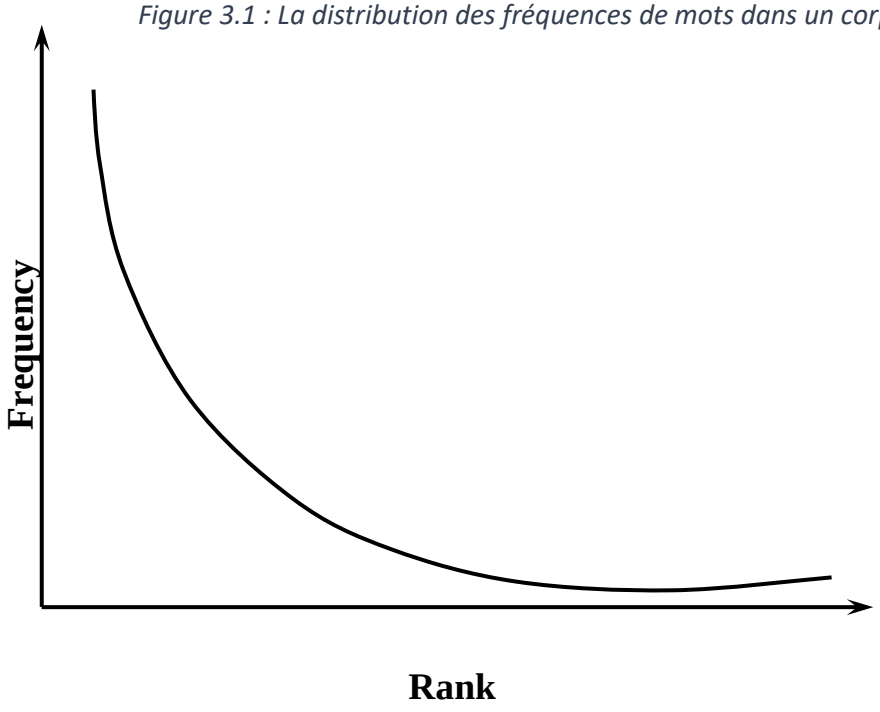


Tableau 3.1: Fréquence de mots selon TREC 3 (125,720,891 total de mots, 508,209 mots uniques)

Mot fréquent	**Nombre d'occurrences**	**Total pourcentage**
The	7,398,934	5.9
Of	3,893,790	3.1
To	3,364,653	2.7
And	3,320,687	2.6
In	2,311,785	1,8
Is	1,599,147	1,2
For	1,313,561	1,0
The	1,144,860	0,9
That	1,066,503	0.8
Said	1,027,713	0.8

Cette courbe montre qu'il existe des mots très courants. Par exemple, des mots tels qui représentent jusqu'à 10% de toutes les occurrences de mots dans une collection.

À l'autre extrémité de la courbe, nous identifions les mots rares, qui font la « masse » des mots dans un corpus. En fait, il a été démontré que la moitié des mots d'une collection de textes n'apparaissent qu'une seule fois (ces mots sont appelés *hapax legomena*, ce qui signifie « lire une seule fois » en grec).

Comme nous l'avons cité dans le chapitre précédent, deux lois ont été établies sur les distributions de mots : la loi de Zipf et la loi de Heaps. La loi de Zipf modélise la distribution des termes dans un corpus et fournit une manière mathématique de répondre à cette question : combien de fois le "r-ème" mot le plus fréquent apparaît-il dans un corpus de N mots ? Plus précisément, en supposant que f est la fréquence d'un mot, et r est son rang reflétant la position du mot dans une liste triée par fréquence décroissante (par exemple, dans le Tableau 4.1 le mot « of » a une fréquence de 3 893 790 et possède le rang 2 dans la liste triée par fréquence).

Zipf a trouvé ce qui suit en 1949, où k est une constante qui dépend du corpus : ***f r = k (pour la constante k)***. La loi de Zipf peut être utilisée pour faire des prédictions concernant le nombre de mots qui ont une certaine plage de fréquences et généralement pour montrer la distribution des mots dans un corpus.

La deuxième loi, appelée loi de Heaps, modélise le nombre de mots du vocabulaire en fonction de la taille du corpus. Le nombre de mots uniques (vocabulaire) dans une collection n'augmente pas linéairement avec le nombre de mots dans cette collection. En effet, les mots que nous avons déjà vus vont se répéter à fur et à mesure que le corpus se développera - ainsi, la forme du vocabulaire par rapport à la taille qui est généralement comme le montre la Figure 3.2.

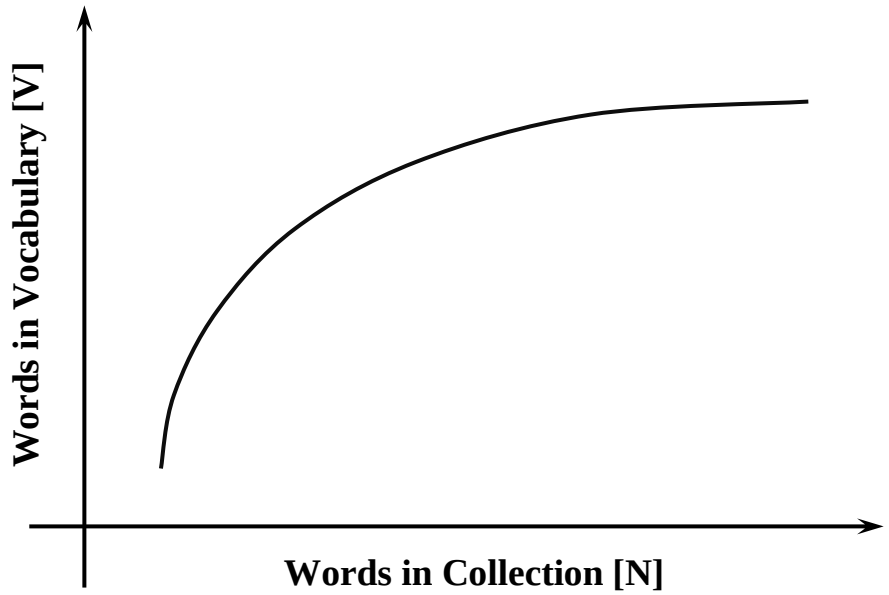


Figure 3.2 : Croissance du vocabulaire avec la taille du corpus

On remarque d'après la courbe que peu importe la taille d'un corpus (et donc son vocabulaire correspondant), le fait d'y ajouter plus de texte apportera quelques nouveaux mots. Dans ce contexte, la loi de Heaps' peut être utilisée ainsi pour répondre à des questions comme celles-ci :

- Quel est le nombre de mots uniques apparaissant dans un corpus de N mots ?
- Combien de mots vais-je avoir dans une collection quand j'aurai un vocabulaire de mots V ?

Étant donné un corpus de N mots avec un vocabulaire de V mots, la loi de Heaps' énonce ce qui suit (où K et β sont des paramètres qui peuvent être déterminés sur une collection donnée de textes) :

$V = KN^{\beta}$ avec les constantes K, $0 < \beta < 1$

La plage typique des constantes est $K \approx 10-100$ et $\beta \approx 0{,}4-0{,}6$. En utilisant cette loi, on peut faire des prédictions sur ce que sera le vocabulaire d'un corpus à mesure qu'il va accroitre. Par exemple, nous pouvons savoir qu'une collection de textes juridiques comprend 1 000 000 de mots et un vocabulaire de 135 000 mots. La loi de Heaps' nous aidera à prédire la taille du vocabulaire dans un contexte futur lorsque le corpus atteindra, par exemple, 5 000 000 mots.

3.4. Traitement de texte avancé

Comme indiqué au début de ce chapitre, en plus des étapes de traitement de base, il existe d'autres étapes de traitement de texte qui peuvent être appliquées à n'importe quel texte donné. Certains des plus couramment utilisés sont brièvement décrits ci-dessous.

- **Le balisage d'une partie du discours**

Le balisage de partie du discours consiste à attribuer à chaque mot dans un texte donné son rôle syntaxique correct, tel que nom, verbe, etc. La plupart des algorithmes sont basés sur un apprentissage supervisé et s'appuient donc sur des données précédemment annotées pour apprendre à attribuer des parties du discours à un nouveau texte. Il existe également différents ensembles de balises qui ont été proposés, allant d'une poignée d'étiquettes (ex : nom, verbe) à de plus grands ensembles de balises tels que Penn Treebank (Marcus, Marcinkiewicz et Santorini, 1993), qui contient près de 40

différentes balises (par exemple, NN = nom commun singulier, NNS = nom commun pluriel, NNP = nom propre, NNPS = nom propre pluriel) …

Lorsque des données annotées sont disponibles, l'approche la plus élémentaire et courante qu'on pourrait utiliser consiste à attribuer à chaque mot dans un texte sa balise de partie de discours la plus fréquente à partir des données annotées. Si, par exemple, le mot à annoter est race et qu'il a été trouvé 118 fois avec une annotation de verbe et 206 fois avec une annotation de nom, alors cette ligne de base choisira l'étiquette de nom simplement en fonction de sa fréquence précédente.

Cependant, les méthodes de marquage les plus avancées vont au-delà de la fréquence du tag et prennent également en compte le contexte. Par exemple, si une fois de plus nous devons annoter le mot race et nous savons que le contexte est de faire la course, intuitivement sa balise verbale sera plus probable que dans un contexte tel que la race, où la balise substantielle est plus probable.

Pour tenir compte du contexte, un marqueur de partie du discours collectera les probabilités de séquence à partir des données annotées. Dans leur forme la plus simple, ces probabilités vont considérer la probabilité d'une balise à partir de la balise précédente - par exemple***, P (VB | TO)*** (probabilité du verbe de balise étant donné que la balise précédente est "to") - mais elles peuvent également approfondir l'historique et envisager des séquences de balises plus longues. La question qui se pose dans ce cadre est la dépendance entre les balises - dans notre exemple qui vient d'être donné, comment savoir que la balise précédente est "to", car le balisage de ce mot précédent est lui-même un processus probabiliste. Les modèles de Markov cachés (HMM) peuvent répondre à de telles situations.

Par exemple, l'algorithme de Viterbi est un algorithme récursif qui calcule la probabilité de la balise actuelle en considérant toutes les différentes balises précédentes ainsi que leurs probabilités. Les autres variantes des HMM sont les algorithmes avant et arrière, qui sont également des méthodes récursives qui permettent de calculer de manière itérative la probabilité d'une séquence de balises.

En plus des HMM, de nombreuses autres méthodes d'étiquetage des parties de la parole ont été introduites. Par exemple, le balisage basé sur la transformation tel que celui de Brill (1992) commence par annoter tous les mots avec leur balise la plus fréquente, puis il utilise des données annotées pour apprendre des règles de transformation telles que « si le mot précédent est le changement de balise de VB en NN » ou « si le mot suivant a la balise NN, changez la balise de RB à JJ »., 2003).

Ainsi, des méthodes d'apprentissage automatique telles que celles décrites au chapitre 4, titre 5 ont également été utilisées efficacement pour le balisage d'une partie de la parole.

- **L’identification de la collocation**

Le but de l'identification de la collocation est d'identifier automatiquement les séquences de mots qui ont une signification particulière lorsqu'ils sont pris sous forme de phrase, comme « mother-in-law » ou « kick the bucket ». L'identification de ces collocations peut être utile à de nombreuses fins tel que :

- La récupération d'informations, sachant que la requête inclut le mot “mother-in-law“ plutôt que les mots individuels “mother“, “in“ et “law“, cela évitera les correspondances erronées avec des documents qui renvoient à des documents juridiques.
- La traduction automatique, les collocations ont souvent un correspondant dans l'autre langue qui n'est pas une simple composition des traductions des mots individuels - conformément à l'exemple de “mother-in-law“, la traduction correcte en italien est un mot suocera, plutôt que le séquence insensée « mama in legge » obtenue à partir d'une traduction littérale.

Les méthodes les plus fréquemment utilisées pour l'identification des collocations consistent principalement en des mesures théoriques de l'information, qui tentent d'identifier les co-occurrentes significatives dans le texte. Ceux-ci incluent des informations mutuelles ponctuelles (PMI), le coefficient de Jaccard, le chi-square et bien d'autres (Church & Hanks, 1990). L'essentiel de ces méthodes est qu'elles calculent une mesure de cooccurrence à la fois en supposant que les mots d'une collocation se produisent ensemble et en supposant l'indépendance. Ces deux mesures sont ensuite combinées ensemble dans une métrique finale qui reflète la probabilité que les mots considérés se produisent ensemble.

Par exemple, étant donné les mots “wood“ and “desk“, afin de déterminer si l'expression “wood desk“ est une collocation ou non, on pourrait, par exemple, utiliser PMI et calculer la probabilité ***P (wood)*** de voir le mot “wood“ dans un corpus donné (le nombre de fois où le “wood“ apparaît dans le corpus divisé par le nombre total de mots dans le corpus), la probabilité ***P (desk)*** de voir le mot “desk“ dans le même corpus (en utilisant le même comptage de division) et la probabilité ***P (wood desk)*** de voir l’expression “wood desk“ dans le corpus (le nombre de fois que les mots “wood“ et “desk“ apparaissent côte à côte dans le corpus divisé par le nombre total de mots dans

le corpus). Avec ces entrées, le PMI sera calculé comme ***PMI (wood, desk) = log (P (wood, desk) / (P (wood) × P (desk)))***, et selon l'endroit où cette valeur tombe par rapport à une donnée seuil, un wood desk serait considéré comme une collocation ou non (pour cet exemple particulier, le PMI sera probablement très petit, ce qui suggérera que le bureau en bois n'est pas une collocation).

- **L'analyse syntaxique**

L'analyse syntaxique s'appuie souvent sur le texte balisé d'une partie du discours et vise à identifier les relations syntaxiques entre les constituants du langage. Certains analyseurs produisent des arbres syntaxiques de circonscription (Collins, 2003), souvent exprimés en notation arborescente ou entre parenthèses, qui peuvent avoir plusieurs éléments dans un constituant syntaxique (par exemple, identifier une expression nominale comme étant formée par un déterminant suivi d'un adjectif puis d'un nom). Par exemple, étant donné le texte « I am happy », un analyseur de circonscription produira la sortie suivante: (ROOT (S (NP (PRP I)) (VP (VBP am) (ADJP (JJ happy))) (..)), qui reflète entre autres qu'il existe une phrase (S) formée par une phrase substantielle (NP) et une phrase verbale (VP) et que la VP est formée par un verbe (VBP) et une phrase adjectivale (ADJP), etc.

D'autres analyseurs génèrent principalement des dépendances (Klein et Manning, 2004), c'est-à-dire des relations binaires entre les éléments du texte (par exemple, un adjectif qui a une relation modificatrice avec son nom). Par exemple, pour le même texte considéré précédemment, un analyseur de dépendances peut générer les dépendances suivantes **nsubj (happy-3, I-1) et cop (happy-3, am-2)**, ce qui indique une dépendance de sujet **(nsubj)** entre « happy » et « I », et une dépendance copulative (flic) entre « happy » et « am ».

Les analyseurs les plus précis fonctionnent en entraînant des systèmes supervisés sur des données analysées manuellement, comme le Penn Treebank (Marcus et al., 1993). Étant donné un grand corpus de données annotées, ces analyseurs apprendront les probabilités de relations entre les mots, par exemple, la probabilité qu'une phrase soit formée par une phrase nominale suivie d'une phrase verbale par rapport à une phrase formée uniquement par une phrase verbale.

Compte tenu de toutes ces probabilités apprises au cours de la formation, les analyseurs utiliseront un algorithme de programmation dynamique pour trouver l'ensemble de règles qui s'appliquent à une entrée donnée. En plus des analyseurs

supervisés, il y a également eu des travaux sur la construction d'analyseurs non supervisé.

- **Reconnaissance d'entité nommée**

Le balisage d'entité nommée - parfois considéré comme un cas spécial d'extraction d'informations (voir le chapitre 4, titre 8) - vise à identifier les entités nommées à partir d'un ensemble prédéfini, comme une personne, un emplacement ou une organisation. Le balisage d'entité nommée est une étape nécessaire pour de nombreuses autres applications, telles que

- La réponse aux questions ;
- Les technologies conversationnelles ;
- La géolocalisation de texte ;
- Les applications d'exploration de texte où il est important de connaître les personnes, les emplacements ou les organisations dans un texte.

On pourrait, par exemple, chercher à savoir dans quelle mesure un texte est centré sur la personne, ce qui peut être partiellement résolu en trouvant toutes les mentions de la personne dans le texte, ou à trouver le sentiment envers certaines organisations, auquel cas l'identification des organisations est importante.

Les techniques les plus courantes pour le balisage d'entités nommées combinent des règles disponibles via des lexiques étendus qui spécifient les valeurs possibles que les entités peuvent prendre (par exemple, les répertoires géographiques) ou des jetons qui peuvent se produire avant ou après une entité (par exemple, M. avant une mention de personne) ou Inc. après mention d'une organisation) avec apprentissage supervisé, qui vise à apprendre automatiquement les propriétés des entités nommées à partir de texte précédemment annoté.

Plus précisément, en utilisant un cadre d'apprentissage supervisé tel que décrit au chapitre 4, titre 5, chaque entité dans un texte annoté (d'apprentissage) sera transformée en une instance d'apprentissage, avec des caractéristiques (ou attributs) qui reflètent les propriétés de l'entité tout en incluant, par exemple, la valeur des mots juste avant et après l'entité, la position dans la phrase, la casse (majuscule, minuscule), etc.

Comme il s'agit d'un nouveau texte, tous les mots candidats (par exemple, tous les noms) sont transformés en un vecteur d'entité en utilisant un ensemble d'attributs similaires et par conséquent étiquetés comme une certaine entité nommée en utilisant un algorithme d'apprentissage automatique (par exemple, Person / NotPerson).

Alors que de nombreux algorithmes supervisés ont été explorés pour le balisage d'entités nommées, les champs aléatoires conditionnels (CRF) restent parmi les plus efficaces. Les données annotées utilisées pour construire un tagger d'entité nommé peuvent être obtenues automatiquement en appliquant un ensemble de règles comme décrit précédemment (Collins et Singer, 1999), un processus souvent appelé bootstrapping, ou peuvent être obtenus également par des annotations manuelles (Collins, 2002).

- **La désambiguïsation lexicale**

La désambiguïsation lexicale mappe les mots d'entrée à leurs sens du dictionnaire. Elle est utilisée pour identifier la signification d'un mot en fonction de son contexte. Les annotations de signification de mots peuvent être utilisées pour la traduction automatique basée sur des règles, pour l'expansion des requêtes dans la récupération d'informations, pour l'apprentissage des langues, etc. Les deux approches les plus courantes pour les désambiguïsations lexicales sont

- Les approches supervisées basées sur des données annotées.
- Les approches non supervisées qui reposent sur des sources de connaissances.

Lorsque des instances étiquetées pour un mot sont disponibles - par exemple, un certain nombre d'exemples annotés manuellement pour le nom “plante“, nous pouvons utiliser des méthodes supervisées traditionnelles telles que celles décrites au chapitre 4, titre 5 pour construire un système qui peut prédire automatiquement la signification d'un mot dans le nouveau texte (Yarowsky, 2000).

L'alternative consiste à utiliser exclusivement les informations obtenues à partir de ressources lexicales, telles que les définitions de sens des mots, les synonymes, les hyperonymes, etc. (voir le chapitre 4, titre 4). Dans ce cas, on pourrait utiliser, par exemple, le chevauchement entre les définitions des significations des mots dans une phrase, pour trouver l'ensemble des sens des mots qui maximisent ce chevauchement (Lesk, 1986).

Il convient de noter que cette méthode peut faire face à une explosion combinatoire, car elle vise à considérer simultanément l’ensemble des significations des mots dans un texte. Une version simplifiée et plus efficace de cette approche consiste à lever l'ambiguïté d’un mot à la fois, en mesurant le chevauchement entre les définitions du sens de mot et le contexte où le mot apparaît, et choisir ainsi la signification qui a le plus grand chevauchement avec le contexte (Banerjee et Pedersen, 2002).

Des travaux récents sur la désambiguïsation lexicale ont également utilisé des ressources lexicales autres que des dictionnaires, tels que Wikipedia (Mihalcea, 2007).

- **La similarité lexicale**

Mesurer la similarité lexicale des mots ou séquences plus longues telles que des phrases ou des documents entiers est l'une des principales tâches du domaine du traitement du langage naturel (NLP) et se trouve au cœur d'un grand nombre d'applications telles que :

- La recherche d'informations ;
- La détection de plagiat ;
- La notation à réponse courte ;
- L'implication textuelle et autres ;
- Etc.

Un nombre relativement important de mesures de similarité de mots et de textes ont été proposées dans le passé, allant de mesures basées sur la distance calculée sur des réseaux sémantiques ou des taxonomies, à des métriques basées sur des modèles de similarité distributionnelle tirés de grandes collections de textes.

Les mesures de similarité sémantique des mots basées sur les corpus tentent d'identifier le degré de similarité entre les mots en utilisant des informations provenant exclusivement de grands corpus. Dans la similarité de distribution, les mots sont représentés par leur distribution dans un grand corpus (par exemple, présence – absence ou poids à l'intérieur des documents dans une collection ; position à l'intérieur des dépendances syntaxiques), et par conséquent, la similarité de deux mots est mesurée par la similarité de leurs représentations vectorielles.

L'analyse sémantique latente (LSA ; Landauer, Foltz et Laham, 1998) tente d'apporter dans ce sens des améliorations en réduisant ces représentations à un espace de faible dimension, qui capture également les relations sémantiques entre les mots. Une autre méthode connexe est la méthode explicite d'analyse sémantique (Gabrilovich et Markovitch, 2007), qui représente chaque mot comme un vecteur qui reflète la présence-absence du mot dans les articles de Wikipédia.

Une autre méthode de mesurer la similarité de deux mots est de calculer leur PMI dans un large dataset (Turney, 2001) tout en en mesurant la probabilité que les mots coexistent ensemble. Elle est mesurée sous forme de ***log (P (w1, w2) /(P(w1) P(w2)), où***

P (w1, w2) est la probabilité que les deux mots apparaissent dans une petite fenêtre, tandis que ***P(w1) et P(w2)*** reflètent les probabilités de chacun des deux mots.

Les mots similaires ont tendance à avoir des informations mutuelles plus élevées, tandis que les mots qui ne sont pas liés auront une faible information mutuelle. Plus récemment, des approches d'apprentissage approfondi ont été utilisées pour créer des incorporations de mots, qui sont des représentations vectorielles de mots, obtenues en tant que sortie d'un réseau de neurones formé sur de très grands corpus textuels (Mikolov, Sutskever, Chen, Corrado et Dean, 2013). Par exemple, Google a publié l'outil word2vec avec l'intégration de mots préformé sur de grandes données de nouvelles, où la similarité de deux mots peut être mesurée en calculant la similarité de leurs vecteurs d'intégration.

4. Ressources lexicales

4.1. Introduction

Les ressources lexicales peuvent être monolingues, bilingues ou multilingues. Leur développement demande un travail minutieux et acharné car il nécessite beaucoup de temps de la part d'experts tels que des lexicographes, des spécialistes en linguistes ou des psychologues. Certaines ressources ont pris de nombreuses années à compléter. Par exemple, “l'Oxford English Dictionary“ considéré comme étant le premier dictionnaire complet de la langue anglaise. Ce dernier a été compilé sur une période de 27 ans. De plus, de nombreuses ressources lexicales ont connu plusieurs éditions au fil du temps. Dans ce contexte, nous identifions par exemple “WordNet“ qui est désormais le dictionnaire électronique anglais le plus populaire qui a vu sa première version publiée en 1991, et depuis lors, il a évolué vers la version 3.1, publiée en 2012. Cependant, il existe des ressources lexicales plus récemment construites, telles que Wikipédia ou Wiktionnaire bénéficiant ainsi du “crowdsourcing“ en ligne, qui présente l'avantage d'un nombre significativement plus important de contributeurs au détriment de la cohérence ou de la qualité.

4.2. La base de données lexicales : WordNet

WordNet (Fellbaum, 1998 ; Miller, 1995) est un dictionnaire créé en 1985 par un groupe dirigé par Miller à l'Université de Princeton. WordNet couvre la majorité des noms, verbes, adjectifs et adverbes en langue anglaise ainsi qu'un riche ensemble de

relations sémantiques qui relient ces concepts. Les mots dans WordNet sont organisés en ensembles de synonymes. La version "WordNet 3.1" est la dernière version de WordNet (en décembre 2016) et dispose d'un vaste réseau de 155 000 mots regroupés en 117 000 synonymes.

De nombreux synonymes dans WordNet sont connectés à plusieurs autres synonymes via des relations sémantiques, telles que l'hyperonymie (« est un »), l'holonymie (« partie de »), etc. Le Tableau 3.2*Tableau 3.2 : Relations sémantique* répertorie les relations sémantiques disponibles dans WordNet, ainsi que des exemples.

Tableau 3.2 : Relations sémantique

Relation	**Description**	**Exemple**
Hyperonymie (nom, verbe)	A est un hyperonyme de B ➔ B est dans A.	Le canine est une hyperonyme du chien.
Hyponymie (nom, verbe)	A est un hyponyme de B ➔ A est dans B.	Le dalmatien est un hyponyme du chien.
Holonymie (nom)	A est un holonyme de B ➔ B est une partie de A.	La forêt est un holonyme de l'arbre.
La méronymie (nom)	A est un méronyme de B ➔ A est une partie de B.	L'arbre est un méronyme de la forêt.
Les coordonnées (nom, verbe)	A est une coordonnée de B ➔ A et B possèdent le même hyperonyme.	Le dalmatien est une coordonnée du caniche. (Les deux possèdent le même hyperonyme chien)
Troponyme (verbe)	A est un troponyme de B ➔ Faire A est une manière de faire B.	Marcher est un troponyme de se promener.
Implication (verbe)	A implique B ➔ Faire A implique également l'exécution de B.	Ronfler implique dormir.
Noms apparentés (adjectifs)	A a un nom connexe B➔ A est dérivé de B.	Studieux est relié au nom étude.
Antonymes (adjectifs, adverbes)	A est un antonyme de B ➔ A et B ont des sens opposés.	Joli est n antonyme de laid.
Similaire à (adjectif)	A est similaire à B ➔ A et B ont le même sens.	Joli est similaire à charmant.

Concepts de base :

- Une ressource lexicale est une base de données qui contient des informations sur les mots du langage. Les ressources lexicales peuvent être monolingues, bilingues ou multilingues.
- Un dictionnaire est une liste alphabétique des mots dans une langue, qui peut inclure des informations telles que des définitions, des exemples d'utilisation, des étymologies, des traductions, etc.
- Un thésaurus est une base de données qui regroupe les mots dans une langue en fonction de leur similitude.
- Un réseau sémantique est un réseau qui définit les relations sémantiques entre les mots.
- Une concordance est une liste alphabétique des mots utilisés dans un texte, ainsi que les contextes immédiats où chaque mot apparaît.

Les noms et les verbes sont organisés en hiérarchies basées sur la relation d'hyperonymie et d'hyponymie.

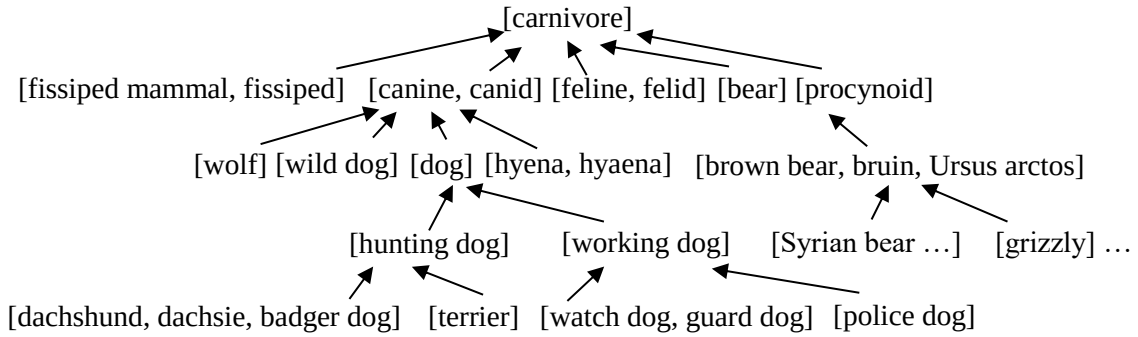


Figure 3.3 Exemple d'une hiérarchie de noms WordNet

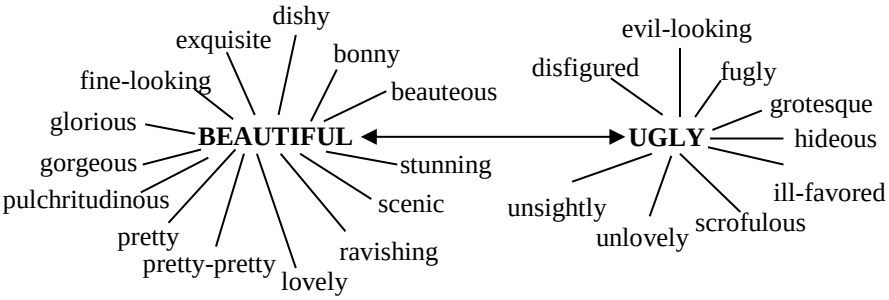


Figure 3.4 : Exemple d'un cluster d'adjectifs WordNet

Dans le paragraphe suivant, nous examinons de plus près deux ressources qui ont été créées directement sur WordNet : WordNet-Affect et WordNet Domains.

- **L’extension Wordnet Domains**

WordNet Domains (Magnini & Cavaglia, 2000) est une extension de WordNet avec des étiquettes de domaine. Chaque synonyme dans WordNet a été étiqueté de façon semi-automatique avec une ou plusieurs des 200 étiquettes de domaine, qui reflètent l'appartenance des mots du synonyme à des domaines tels que l'archéologie, l'art, le sport, etc. En plus de l’enrichissement des synonymes du WordNet avec des informations supplémentaires sur leur catégorie sémantique, les domaines WordNet ont également l'avantage supplémentaire de :

- Connecter des mots de différentes parties du discours, car un domaine peut englober des verbes, des noms, des adjectifs ou des adverbes.
- Permettre des méthodes fondées sur des principes pour regrouper des sens similaires, en fusionnant les synonymes dans lesquels un mot apparaît et qui appartiennent au même domaine.

Dans ce contexte, WordNet Domains a été utilisé dans un grand nombre d'applications, notamment l’explication du sens des mots, la classification des textes et la génération de blagues. Le Tableau 3.3 : Exemples de domaines du WordNet Domains montre plusieurs exemples de domaines ainsi que des synonymes dans WordNet.

Tableau 3.3 : Exemples de domaines du WordNet Domains

Domaines	**Synonymes**
Sport	{athlète, jock} , {jeu, équipement}
Industrie	{hôpital, infirmerie} , {station de chemin de fer, station de train, terminal de chemin de fer, dépôt de train)
Littérature	{poésie, vers} , {vers, rime}, {manuscrit, livre de feuilles}
Religion	{service de chapelle, chapelle } , {religieux, monastique}, {couvent}

- **WordNet- Affect**

WordNet-Affect (Strapparava & Valitutti, 2004) est une autre ressource qui a été créée à partir du WordNet, en annotant des synonymes avec plusieurs émotions. Il utilise plusieurs ressources pour l'information affective, y compris la classification des émotions d'Ortony, Clore et Collins (1990). WordNet-Affect a été construit en deux étapes :

- Tout d'abord, une ressource de base a été construite sur la base d'un certain nombre d'heuristiques et de traitement semi-automatique.
- Dans la deuxième étape les synonymes de base ont été automatiquement développés à l'aide des relations sémantiques disponibles dans WordNet.

Le Tableau 3.4 : Exemples de mots de WordNet-Affect montre plusieurs exemples de mots pour chacune des six émotions de base d'Ortony et ses collègues (1990) : **colère, dégoût, peur, joie, tristesse et surprise**.

4.3. Thésaurus de Roget

Le Thésaurus de Roget (Roget, 1911/1987) est un thésaurus de la langue anglaise, avec des mots et des phrases regroupées en classes hiérarchiques. Une classe de mots comprend généralement des synonymes, ainsi que d'autres mots sémantiquement liés. Les classes sont divisées en sections, sous-sections, en-têtes et paragraphes, qui sont, à leur tour, divisés par une partie du discours : noms, verbes, adjectifs et adverbes.

Enfin, chaque paragraphe est regroupé en plusieurs ensembles de mots sémantiquement liés. La version la plus récente du Thésaurus de Roget (Roget, 1911/1987) comprend 250 000 mots, regroupés en huit grandes classes qui se divisent en 39 sections, 79 sous-sections, 596 groupes de têtes et enfin 990 têtes.

Tableau 3.4 : Exemples de mots de WordNet-Affect

Émotion	**Exemples de mots**
Colère	colère, ombrage, offense, humeur, irritation, lividité, irascibilité, fureur, rage.
Dégout	horreur, horrible, dégoûtant, abominable, hideux, malade, fatigué, méchant.
Peur	terrible, laid, incertain, méchant, timide, effrayé, scandaleux, panique, hystérique.
Joie	Adorable, sympathie, tendresse, égard, respect, fierté, préférence, amour.
Tristesse	Misère, oppressif, pathétique, apeuré, désolé, sombre, consterné.
Surprise	Merveille, admiration, étonnement, stupéfiant.

Le Tableau 3.5 montre quatre exemples de jeux de mots trouvés sous chacune des quatre parties du discours sous l'informalité 408.

Tableau 3.5 : Exemples de jeux de mots dans Roget

Partie du discours	**Échantillons d'œuvres liées à la sémantique**
Nom	informalité, manque de formalité, manque de cérémonie, manque de convention, indifférence, non-conformité.
Verbe	être informel, ne pas participer à la cérémonie, être soi-même, être naturel, se détendre, se sentir chez soi, ne pas insister, renoncer aux règles, venir comme vous êtes.
Adjectif	familier, naturel, simple , chaleureux, folklorique, commun, non affecté.
Adverbe	librement, indulgemment, toléramment, permissivement, lâchement, irrégulièrement

4.4. Linguistic Inquiry and Word Count (LIWC)

Linguistic Inquiry and Word Count (LIWC) a été développé comme ressource pour l'analyse psycholinguistique par Pennebaker, Francis et Booth (2001). Cette ressource a été utilisée dans un grand nombre d'études en recherche sociale, psychologique et linguistique, portant sur des tâches telles que :

- L'analyse des traits psychologiques (Mairesse et al., 2007 ; Pennebaker & King, 1999)
- La tromperie (Mihalcea & Strapparava, 2009; Ott, Choi, Cardie et Hancock, 2011)
- L'analyse sociale des conversations (Stark, Shafran et Kaye, 2012)
- La prédiction de la dépression (Resnik, Garron et Resnik, 2013).
- L'identification du sarcasme (González-Ibánez, Muresan , & Wacholder, 2011), et bien d'autres…

La version 2007 de LIWC comprend plus de 2 000 mots et tiges de mots regroupés en 80 catégories de mots. Les catégories de mots sont regroupées en quatre grandes classes :

- Les processus linguistiques couvrant principalement les mots fonctionnels et autres mots courants.
- Les processus psychologiques, y compris les processus sociaux, affectifs, cognitifs, perceptuels, biologiques et temporels.
- Les préoccupations personnelles, telles que le travail, l'argent et la religion.
- Les catégories parlées, y compris les fillers et autres mots parlés.

Le lexique LIWC a été validé en montrant une corrélation significative entre les notations humaines d'un grand nombre de textes écrits et la notation obtenue par des analyses basées sur LIWC des mêmes textes.

Le Tableau 3.6 : Exemples de classes LIWC avec des exemples de mots montre quatre catégories LIWC ainsi qu'un ensemble d'exemples de mots inclus dans ces classes.

Tableau 3.6 : Exemples de classes LIWC avec des exemples de mots

Catégorie	**Classe**	**Exemple de mots**
Nous	Processus linguistique	Notre, nous-mêmes, nous allons…
Optimisme	Processus psychologique	Accepter, meilleur, audacieux, certain, confiance, déterminé, glorieux, espoir
Réussite	Préoccupations personnelles	Mieux, récompenser, avancer, réaliser, motiver, perdre, honorer, grimper
Non-fluidité	Catégories communicationnelles	Er, euhm, zz …

4.5. General Inquirer

The General Inquirer (Stone & Hunt, 1963) est un programme de dictionnaire d'environ 10 000 mots regroupés en une centaine de catégories. Deux des catégories de mots les plus utilisées dans l'Enquêteur général sont les catégories de sentiments positifs et négatifs.. En plus de ces deux classes de mots, The General Inquirer comprend de nombreuses autres classes de mots qui ont des motivations sociales et psycholinguistiques. Ces derniers ont été largement utilisés pour l'analyse de contenu. Le Tableau 3.7 présente trois classes de questions générales, ainsi que des exemples de mots de ces classes.

Tableau 3.7: Exemple de mots et catégories dans le général inquirer

Catégorie	**Exemple de mots**
Académique	Académique, astronomie, biologie, chimie, crédit, degré, physicien, librairie
Rituel	Embuscade, rendez-vous, affaire, pont, recensement, démonstration
Femelle	Tante, féminin, fille, déesse, elle, héroïne, grand-mère, mère, reine.

4.6. Wikipédia

Wikipédia est une encyclopédie en ligne gratuite, qui représente le résultat d'un effort continu de collaboration d'un grand nombre de contributeurs bénévoles. Tout internaute peut créer ou éditer une page Web Wikipédia, et cette « liberté de contribution » a un impact positif à la fois sur la quantité (nombre croissant d'articles) et sur la qualité (les erreurs potentielles sont rapidement corrigées dans l'environnement collaboratif) de cette ressource en ligne.

Les éditions Wikipédia sont ainsi disponibles dans plus de 280 langues, avec un nombre d'entrées variant de quelques pages à 4 millions d'articles ou plus pour chaque langue. Le Tableau 3.8 : Nombre d'articles et d'utilisateurs pour le top 10 des éditions Wkipedia présente les 10 plus grandes Wikipédia (en avril 2017), ainsi que le nombre d'articles et le nombre approximatif de contributeurs.

Tableau 3.8 : Nombre d'articles et d'utilisateurs pour le top 10 des éditions Wkipedia

Langue	**Wiki**	**Articles**	**Utilisateurs**
Anglais	En	5,391,707	30,771,377
Cebuano	Ceb	4,138,135	33,996
Suédois	Sv	3,785,049	543,852
Allemand	De	2,048,714	2,616,328
Néerlandais	Nl	1,897,908	830,046
Français	Fr	1,858,259	2,757,540

Russe	Ru	1,384,908	2,078,886
Italien	It	1,346,331	1,448,417
Espagnol	Es	1,327,066	4,564,582
Waray	war	1,262,379	31,594

- **Wiktionnaire**

Wiktionnaire est un projet jumeau de Wikipédia, géré par la même fondation Wikimedia. Wiktionnaire est un dictionnaire bénévole couvrant un grand nombre de langues. Les mots du Wiktionnaire incluent des synonymes et des définitions, des liens vers des traductions dans d'autres langues et un certain nombre de relations telles que des hyponymes et des termes dérivés. L'étymologie des mots, qui relie la forme actuelle d'un mot à des versions antérieures dans d'autres langues, est une information utile disponible pour de nombreux mots du Wiktionnaire.

- **BabelNet**

BabelNet (Navigli & Ponzetto, 2012) est une très grande ressource multilingue qui a été créée en connectant automatiquement Wikipédia en plusieurs langues, ainsi que des dictionnaires WordNet, des Wiktionnaires et plus récemment d'autres ressources telles que ImageNet et FrameNet. Il couvre actuellement 271 langues et comprend environ 14 millions de synonymes Babel, chacun de ces synonymes contenant toutes les disponibilités pour un concept en plusieurs langues. Dans l'**Erreur ! Source du renvoi introuvable.** de ce rapport nous citons quelques ressources lexicales et interfaces de programmation d'applications.

5. L'apprentissage supervisé

En termes simples, la tâche de l'apprentissage supervisé (également appelé apprentissage machine supervisé ou parfois simplement apprentissage) consiste à utiliser un système automatique pour tirer des enseignements de l'historique des occurrences d'un certain « événement » et, par conséquent, faire des prédictions sur les occurrences futures de cet événement. À titre d'exemple, considérons l'une des tâches mentionnées précédemment concernant les prévisions météorologiques et en particulier le problème de savoir s'il pleuvra ou non. Nous pourrions imaginer un ensemble

d'occurrences antérieures de l'événement « pluie ou pas », ainsi que certaines caractéristiques (ou attributs) représentatifs, comme l'illustrent les lignes 1 à 4 du tableau 5.9. Compte tenu de ces occurrences précédentes de l'événement de pluie, nous pourrions imaginer un système qui pourrait identifier une association entre le ciel = couvert, l'humidité = élevé, le vent = fort et la pluie. Par conséquent, ce système serait capable de prédire la probabilité de pleuvoir compte tenu des observations de la cinquième instance du tableau 39.

Il existe de nombreux algorithmes d'apprentissage supervisé qui peuvent être utilisés suivant la représentation vectorielle d'un élément relatif à un événement donné, ainsi que des instances spécifiques de cet événement représentées comme des vecteurs de valeurs d'entité et de classe. Ces algorithmes sont généralement classés en deux catégories :

1. ***Les algorithmes désireux*** : ils effectuent l'apprentissage lorsqu'ils sont présentés avec les instances de formation et construisent un modèle qui peut ensuite être rapidement appliqué pour faire des prédictions pour les instances de test. La plupart des algorithmes d'apprentissage supervisé entrent dans cette catégorie. Des exemples de tels algorithmes sont les arbres de décision, les réseaux de neurones, les machines à vecteurs de support (SVM), les Naive Bayesien, etc.
2. ***Les algorithmes paresseux*** ne font pas de travail intensif au moment de la formation et réservent plutôt la majeure partie du processus d'apprentissage au moment où les instances de test deviennent disponibles. Le plus proche voisin est un exemple d'algorithme paresseux.

Tableau 3.9 : Exemple d'occurrences pour l'évènement "pluie"

Instance	Ciel	Température	Humidité	Vent	Pluie ?
1	Ensolleilé	Chaude	Élevé	None	Non
2	Sombre	Chaude	Élevé	Fort	Oui
3	Sombre	Froide	Non	None	Non
4	Sombre	Chaude	Élevé	Fort	Oui
5	Sombre	Froide	Élevé	Fort	?

Bien que la quantité de données d'entrainement soit un aspect important de toute tâche d'apprentissage mais la représentation de données demeure ainsi un critère d'importance extrême. En d'autres termes, quelles caractéristiques sont sélectionnées pour décrire les instances de d'entrainement et de test ? Les sections suivantes traitent le processus de la représentation et de la pondération des caractéristiques et décrivent ainsi trois des algorithmes d'apprentissage supervisé les plus performants.

Concepts de base

L'apprentissage automatique est un domaine de l'intelligence artificielle qui traite le développement de programmes qui donnent aux machines la possibilité de développer leur apprentissage à partir des expériences passées.

- L'apprentissage supervisé est la tâche d'apprendre à partir d'un ensemble d'instances étiquetées (également appelées données d'entrainement) et d'utiliser les modèles appris pour faire des inférences sur de nouvelles données invisibles ;
- L'apprentissage non supervisé consiste à tirer des inférences sans données étiquetées explicites. L'un des types les plus courants d'apprentissage non supervisé est le clustering ;
- Les caractéristiques (ou attributs) sont des propriétés mesurables d'un événement observé ;
- Les vecteurs d'entités sont des collections de ces propriétés, utilisées pour représenter une instance d'un événement. Le même type de vecteurs de caractéristiques est utilisé pour représenter les instances précédentes (l'historique de l'événement) ainsi que les instances futures (pour lesquelles des prédictions doivent être faites);
- Les données d'entrainement sont la collection d'instances d'un événement utilisée pour former un algorithme d'apprentissage automatique. Les instances sont également associées à une classification, qui est souvent attribuée manuellement.

Les données de test sont la collection d'instances d'un événement utilisée pour tester un algorithme d'apprentissage automatique. L'algorithme fera automatiquement une prédiction des instances dans les données de test.

5.1. La représentation et pondération des caractéristiques

Les caractéristiques utilisées pour représenter les instances d'un problème d'apprentissage peuvent appartenir à deux types différents. Il peut s'agir de caractéristiques discrètes, retirant des valeurs d'un ensemble fini. Par exemple, la valeur de la fonction ciel dans l'exemple illustré à la Figure 3.5 peut-être ensoleillée ou sombre. Les valeurs d'une fonction discrète n'ont pas à être fixé a priori. Il est plutôt bénéfique de déduire l'ensemble de valeurs en fonction des instances d'entrainement et de test observées. Les caractéristiques peuvent également être continues, c'est-à-dire qu'elles prennent des valeurs numériques, qui peuvent être des valeurs entières ou réelles, positives ou négatives. Il est également possible que la valeur d'une fonctionnalité ne soit pas observable pour une certaine instance, dans ce cas, il est typique d'utiliser un point d'interrogation pour refléter le manque d'informations pour cet attribut d'instance.

Les caractéristiques sélectionnées pour décrire un problème d'apprentissage ne sont pas toutes également utiles. Il est donc important d'avoir des moyens de mesurer le poids de chaque entité. Cela se fait généralement automatiquement par les algorithmes d'apprentissage eux-mêmes, qui passent les instances d'entrainement en cible et calculent le degré de discrimination de chaque fonctionnalité (en d'autres termes, dans quelle mesure cette fonctionnalité aide à trouver la bonne classe pour une instance ?).

- **La pondération des caractéristiques**

Il existe différentes métriques de pondération qui peuvent être utilisées pour pondérer les entités. L'une des mesures les plus couramment utilisées est le gain d'informations. Bien que le calcul du gain d'informations constitue une partie intégrante de la plupart des algorithmes d'apprentissage, nous le décrivons brièvement ici, car il est simple à comprendre et peut être utilisé comme un outil d'analyse de données.

Étant donné une collection S d'instances positives et négatives, soit p la probabilité qu'une instance soit positive et q la probabilité qu'une instance soit négative. Nous définissons son entropie comme **Entropie (S) = - p log p - q log q**. L'entropie est à son maximum lorsque **p = q = 1/2** et à son minimum lorsque **p = 1 et q = 0** (nous utilisons l'hypothèse que **log 0 = 0**). Par exemple, si **S** contient **14** exemples : **9 positifs et 5 négatifs**, **Entropie (S) = - (9/14) log (9/14) - (5/14) log (5/14) = 0,94.**

Nous pouvons maintenant définir le gain d'informations comme la réduction attendue de l'entropie lorsque nous divisons un ensemble de données S en fonction d'une certaine caractéristique A.

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

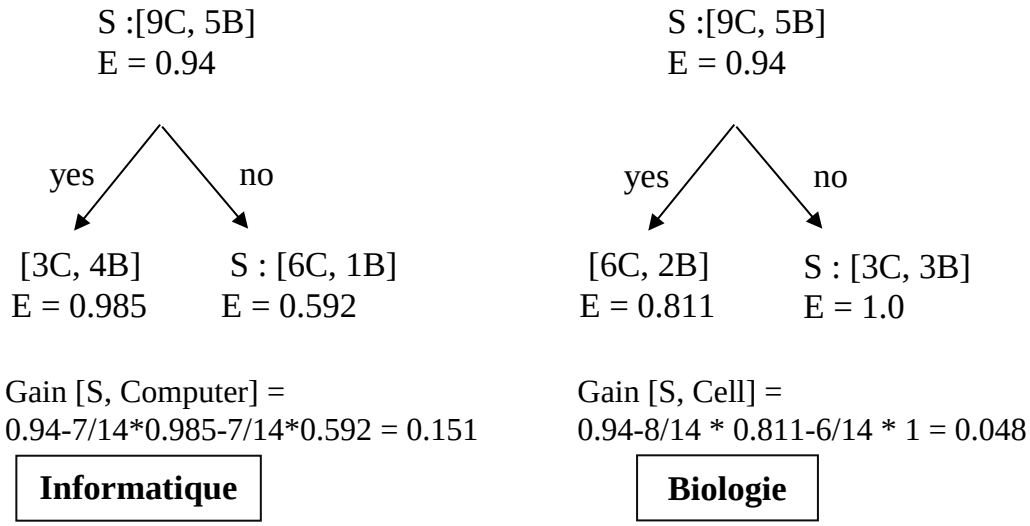
La figure 5.6 montre un exemple de calcul du gain d'informations pour deux entités binaires. Dans cet exemple, nous avons le problème de classer un texte, représenté comme un ensemble de mots, en informatique (étiqueté comme C) ou en biologie (étiqueté comme B). Nous supposons un ensemble de 14 instances d'entrainement, composé de 9 instances appartenant à la classe C, et 5 appartenant à la classe B. Nous avons deux caractéristiques à l'étude : une caractéristique, nommée ordinateur, avec des valeurs oui ou non (selon si le mot ordinateur est présent dans une instance) et une deuxième caractéristique nommée cellule - encore une fois, avec des valeurs oui ou non. Nous pouvons ensuite calculer l'entropie de cet ensemble de données avant de diviser les données en fonction de l'une ou l'autre de ces deux caractéristiques ; comme précédemment : **Entropie (S) = - (9/14) log (9/14) - (5/14) log (5/14) = 0,94**.

Nous pouvons également calculer l'entropie des données après avoir suivi chacune des branches pour une certaine caractéristique. Par exemple, si nous suivons la branche oui pour la caractéristique : ordinateur, ce qui signifie que nous ne regardons que les instances où cette caractéristique a une valeur de oui, nous trouverons trois instances en classe C et 4 en classe B, ce qui correspond à une entropie d'entropie **(ordinateur S = oui) = - (3/7) log (3/7) - (4/7) log (4/7) = 0,985**.

De même, si nous suivons la branche non pour la même caractéristique, l'entropie de l'ensemble résultant sera **Entropie (S ordinateur = non) = 0,592**. Nous pouvons maintenant combiner toutes ces valeurs d'entropie pour calculer le gain d'informations, en pondérant chacune des entropies correspondant aux branches de la caractéristique ordinateur.

Il y a sept instances qui tombent sous la branche oui pour la caractéristique ordinatrice, et donc **Entropie (S Ordinateur = oui)** sera pondérée comme **7/14**. Tout comme une coïncidence dans cet exemple, l'entropie correspondant à l'autre branche aura également un poids de **7/14**. Une fois tous les calculs terminés, nous conclurons que, parmi les deux caractéristiques considérées, l'ordinateur a un pouvoir discriminatoire plus élevé pour ce problème de classification spécifique (avec un gain d'informations de **0,151**, par rapport au gain d'informations de cellule, qui est de **0,048**).

Figure 3.5 Exemple de calcul de gain d'information pour 2 caractéristiques



5.2. Les algorithmes d'apprentissage supervisé

Il existe un très grand nombre d'algorithmes d'apprentissage supervisé qui ont été proposés à ce jour, dont beaucoup ont une ou plusieurs implémentations accessibles au public, en tant que code autonome ou partie de packages d'apprentissage automatique. Certains des algorithmes d'apprentissage supervisé les plus populaires sont Naive Bayes, les régressions, les arbres de décision, l'apprentissage basé sur les instances, les SVM, l'apprentissage en profondeur avec les réseaux de neurones (ces cinq algorithmes sont abordés ci-dessous), les perceptrons et les forets d'arbres décisionnels.

- **Apprentissage basé sur les instances**

L'apprentissage basé sur les instances est une forme d'apprentissage paresseux et comprend des algorithmes tels que **k- le plus proche voisin** (ou KNN, pour faire court) et les machines du noyau. L'idée principale qui sous-tend l'apprentissage basé sur les instances, en général, et KNN, en particulier, est qu'une instance de test peut être classée en trouvant les instances d'entrainement les plus similaires et en utilisant leur classe comme étiquette.

- **L'apprentissage en profondeur avec les réseaux de neurones**

L'apprentissage en profondeur (Goodfellow, Bengio et Courville, 2016) est l'une des plus récentes branches de l'apprentissage automatique et se compose d'algorithmes qui visent à apprendre des représentations de haut niveau des données qui peuvent être utilisées pour un apprentissage efficace. D'une certaine manière, l'apprentissage profond peut être considéré comme un processus dans lequel les caractéristiques elles-mêmes

sont apprises. Plutôt que de faire une ingénierie minutieuse des caractéristiques, qui est requise par la plupart des autres algorithmes d'apprentissage, le principe de l'apprentissage en profondeur est le suivant : avec suffisamment de données, les caractéristiques peuvent être automatiquement apprises.

Diverses architectures d'apprentissage en profondeur ont été proposées, y compris les **réseaux de neurones profonds, les réseaux de neurones récursifs et les réseaux de neurones convolutifs**, avec un grand nombre d'applications en vision par ordinateur, en bio-informatique et en traitement du langage naturel (PNL).

5.3. Évaluation de l'apprentissage supervisé

Pour tout système automatique, il est important d'avoir des moyens d'évaluer les systèmes d'apprentissage supervisé. Habituellement, cela se fait sur des données de test qui sont indépendantes des données d'entrainement, en utilisant des mesures telles que l'exactitude, la précision ou le rappel. La précision est définie comme le nombre total d'instances de test correctement classées, sur le nombre total d'instances de test. La précision et le rappel sont définis par rapport à une classe Ci donnée, la précision étant le nombre total d'instances correctement étiquetées comme Ci par le système sur le nombre total d'instances étiquetées comme Ci par le système, et le rappel étant le nombre total d'instances correctement étiquetés comme Ci par le système sur le nombre total d'instances étiquetées comme Ci dans l'ensemble des données de test.

Pour des résultats plus robustes, les expériences sont généralement exécutées sur plusieurs divisions de données d'apprentissage ou de test, et les résultats obtenus pour différentes divisions sont moyennés. Autrement dit, étant donné un ensemble d'instances étiquetées, disons 1000, nous pourrions prendre 90% de cet ensemble et l'utiliser pour la formation, et les 10% restants pour le test, puis diviser les 1000 instances en une autre division de 90% à 10% et répéter l'évaluation, etc.

Cela conduit également à N-fois des évaluations de validation croisée, lorsque l'ensemble des instances étiquetées est divisé en N sous-ensembles, puis un sous-ensemble est utilisé pour le test et les N-1 sous-ensembles restants sont utilisés pour l'entrainement, puis un autre sous-ensemble est utilisé pour le test et le N restant −1 sous-ensembles pour la l'entrainement, et ainsi de suite N fois, suivis d'une moyenne sur l'ensemble des N résultats. Une alternative à cela est la **N-fold-cross-validation**, où l'ensemble de test se compose d'une seule instance, et l'ensemble d'entrainement se

compose des instances restantes ; ce processus est à nouveau répété plusieurs fois, pour chaque instance de l'ensemble de données.

6. La classification de textes

La classification de textes (appelée également catégorisation de textes) consiste à affecter des textes à une ou plusieurs catégories prédéfinies. Formellement, étant donné une représentation “R“ d'un texte “T“ et étant donné un ensemble fixe de **catégories C = {C1, C2, . . . , Cn}**, la tâche de la classification de texte est de déterminer un mappage de “R“ à une catégorie en “C“. Cela se fait généralement en apprenant à faire de tels mappages à partir d'un ensemble de textes qui ont déjà été mappés à des catégories en “C“.

Dans l'exemple de spam par e-mail que nous avons donné dans la section précédente, en supposant que la représentation “R“ utilisée pour les textes se compose des mots dans ces textes, une association possible qu'un classificateur apprendrait pourrait être par exemple que ***l'hypothèque*** et ***les intérêts*** sont plus souvent associé au courrier indésirable que le courrier électronique légitime, alors que ***le dîner*** et ***le bébé*** se produisent plus fréquemment dans les courriers électroniques non indésirables.

Les catégories peuvent souvent être hiérarchisées, comme, par exemple, dans le cas des catégories de la figure 3.10, qui représentent une classification possible d'un certain nombre de domaines de l'intelligence artificielle.

Dans ce cadre, il est important de faire la différence entre **la classification de texte** et **le clustering de texte**. La classification de texte se réfère à la tâche de regrouper les textes en catégories, cependant, dans le clustering de textes, les catégories ne sont pas connues a priori. Étant donné un ensemble de textes, un système de clustering identifiera que certains textes sont plus similaires les uns aux autres et doivent être affectés au même cluster, mais il ne donnera pas de nom à ce cluster.

De plus, le nombre de clusters dans lesquels une collection de textes sera divisée est souvent inconnu. Ainsi, la classification de texte est souvent considérée comme une tâche supervisée, tandis que le clustering de texte est souvent non supervisé (ainsi, il existe des exceptions, car il existe des méthodes de classification de texte non supervisées et des méthodes de clustering de texte qui sont supervisées).

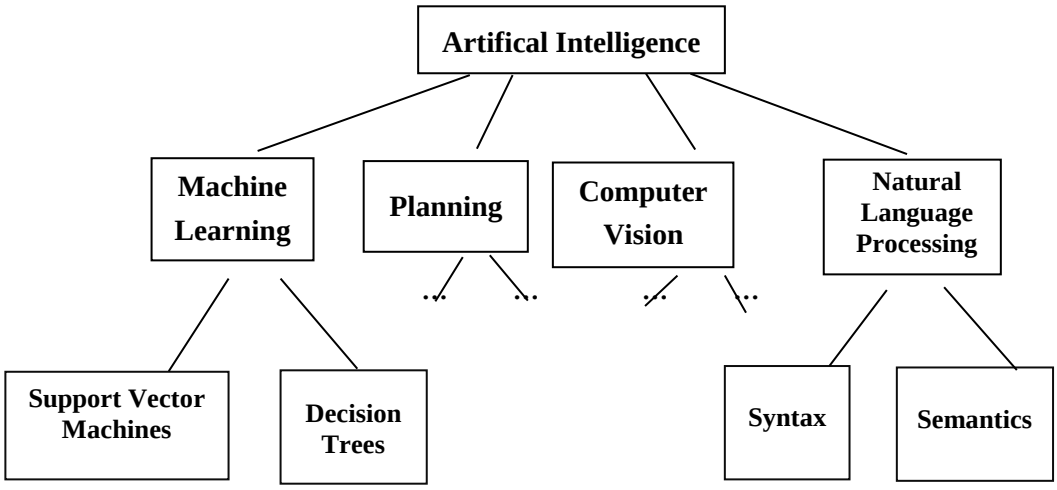


Figure 3.6 : Exemple de catégories hiérarchiques

6.1. Applications de la classification de textes

La classification de textes est l'une des applications les plus utilisées de la fouille de texte ou de l'intelligence artificielle. Ci-dessous quelques exemples de la classification de textes appliquée à des problèmes réels.

- **La classification des sujets**

La classification des sujets est utilisée pour classer les documents en fonction de leur sujet (l'informatique, la musique, la biologie (McCallum et Nigam, 1998)…). Cette tâche est souvent utilisée pour organiser des textes Web (par exemple, le projet Open Directory [http://dmoztools.net.] où des millions de pages Web sont organisées en une hiérarchie de catégories telles que la remise en forme, les logiciels et l'immobilier).

- **La détection des e-mails indésirables**

La détection des e-mails indésirables est l'application la plus omniprésente de la classification de textes. Elle est utilisée quotidiennement par tous les utilisateurs d'un service de messagerie. Les détecteurs de spam s'exécutent généralement « en arrière-plan » sur n'importe quel serveur de messagerie et s'exécutent au-dessus du flux d'e-mails entrants pour déterminer si un e-mail doit être envoyé dans la boîte de réception de l'utilisateur ou dans le dossier de spam.

Il s'agit d'une application où il est très important de maintenir le taux de faux positifs très bas même si cela se fait au détriment de certains faux négatifs. En d'autres termes, on voudrait idéalement éviter de perdre tout e-mail légitime dans le dossier spam, même si certains e-mails de spam parviennent à la boîte de réception de

l'utilisateur. Cette contrainte dicte également certaines caractéristiques des paramètres utilisées pour la classification de texte afin de détecter du spam, seul le spam de haute confidentialité étant filtré.

- **Analyse des sentiments / fouille d'opinions**

L'analyse des sentiments est une application de classification de textes qui a reçu une attention croissante au cours des dernières années (également appelée fouille d'opinion ; Mihalcea, Banea et Wiebe, 2007; Pang et Lee, 2008; Wiebe, Wilson et Cardie, 2005). La classification est effectuée entre un sentiment positif et négatif et utilisée pour détecter le sentiment des consommateurs pour des produits donnés (par exemple, un examen positif ou négatif d'un iPhone) afin de surveiller la marque d'une entreprise et détecter ainsi les problèmes de produits ou de services pour un service client ciblé.

- **Classification de sexe**

La classification des textes est également utilisée pour les tâches visant à « profiler les auteurs », c'est-à-dire déterminer l'âge, le sexe ou l'orientation politique de l'auteur d'un texte (Koppel, Argamon et Shimoni, 2002 ; Liu et Mihalcea, 2007). La classification des textes a également été appliquée à des textes de type traditionnels tels que les livres ou autres œuvres de fiction, c'est dans ce cadre que l'intérêt pour le profilage des auteurs s'est accru avec l'explosion des médias sociaux.

C'est une tâche intéressante non seulement pour ce qu'elle accomplit, mais aussi parce que c'est l'une des rares exemples en intelligence artificielle où un ordinateur fonctionne beaucoup mieux qu'un humain. Prenons, par exemple, l'exemple de textes écrits par un homme ou une femme. Il s'avère que les gens ont généralement du mal à déterminer si un texte a été écrit par un homme ou une femme, principalement parce que certains des indices les plus utiles pour la détection du genre consistent en des mots de fonction (tels que "we" ou "of"), auquel nous attribuons attention.

Les ordinateurs réussissent mieux dans cette tâche, car ils ne sont pas utilisés, comme nous, pour concentrer leur « attention » uniquement sur les mots de contenu et peuvent compter rapidement les mots de fonction qui sont utiles pour cette tâche de classification.

- **Détection de déception**

La détection de déception est une tâche difficile qui consiste à identifier les tromperies dans le texte (Mihalcea et Strapparava, 2009 ; Newman, Pennebaker, Berry et Richards, 2003; Ott, Choi, Cardie et Hancock, 2011). Cela a trouvé des applications non seulement dans le domaine juridique, mais aussi dans la détection de fausses critiques et de publications trompeuses sur les réseaux sociaux.

Comme pour le profilage des auteurs, les indices linguistiques les plus utiles pour la détection de la tromperie consistent en des mots fonctionnels (par exemple, les trompeurs utilisent des expressions auto-référentes telles que “I“ ou moins souvent “we“); ainsi, les humains ont souvent de très faibles performances dans cette tâche.

- **Autres applications**

Outre les exemples illustrés précédemment, il existe de nombreuses autres applications de la classification de texte. Quelques exemples supplémentaires incluent la classification des textes par les axes suivants :

- Leur langue (par exemple, anglais vs chinois vs roumain).
- La classification du genre d'un texte (par exemple, éditoriaux versus critiques de films versus news).
- La détection de contenu émotionnel (par exemple, heureux versus triste versus en colère).
- La classification par rapport à une dimension spécifique du lecteur (par exemple, intéressant pour moi par rapport à pas intéressant pour moi), et bien d'autres.

6.2. Approches de la classification de texte

Les méthodes de classification de texte les plus efficaces sont basées sur les données, ce qui signifie qu'elles s'appuient sur des collections de textes annotées manuellement (ou semestriellement) pour apprendre automatiquement les modèles d'associations entre les mots (ou d'autres indices textuels) et les classes de texte (ou catégories).

Une première étape requise pour activer de tels classificateurs automatiques est la représentation des textes, répondant à des questions telles que celles-ci :

- Quels sont les indices (ou caractéristiques ou attributs) qui sont utiles pour la tâche ?
- Quel devrait être le poids attribué à ces indices ?

La deuxième étape consiste en l'attribution d'un mécanisme d'apprentissage réel, avec un large choix d'algorithmes d'apprentissage automatique. Enfin, lorsque nous parlons de classification de texte, nous devons également parler d'évaluation :

- Comment savons-nous qu'une certaine méthode de classification fonctionne mieux qu'une autre ?

C'est ainsi que cette section fournit des réponses à toutes les questions citées ci-dessus.

- **Représentation des textes pour la classification supervisée**

La plupart des systèmes de classification de texte utilisent un espace de caractéristiques de très grande dimension composé des mots dans les textes. Autrement dit, étant donné une collection de textes, nous pouvons extraire le vocabulaire de cette collection en identifiant tous les mots uniques. Les mots du vocabulaire constitueront alors l'espace caractéristique ; par conséquent, chaque texte de la collection sera représenté comme un vecteur dans cet espace, en utilisant des poids qui représentent l'importance d'un mot dans un texte donné.

Prenons, par exemple, l'exemple des deux textes suivants : « aujourd'hui est un beau jour » et « aujourd'hui est le grand jour ». Le vocabulaire se compose de sept mots (est, un, beau, jour, le, grand, aujourd'hui); ainsi, les vecteurs utilisés pour représenter ces textes auront tous une longueur de sept. En supposant un schéma de pondération très simple, qui ne recherche que la présence d'un mot dans un texte, le vecteur de caractéristiques pour le premier texte sera (1,1,1,1,0,0,1) et pour le second texte sera (1,0,0,1,1,1,1).

Alors que les mots individuels (également appelés uni-grammes) sont les fonctionnalités les plus fréquemment utilisées pour la classification de texte, d'autres fonctionnalités peuvent également être utilisées. Par exemple, on peut également utiliser des séquences de deux mots à la fois : bi-grammes, trigrammes, etc. Dans l'exemple donné précédemment, le vocabulaire des bi-grammes serait (aujourd'hui_est, est_un, un_beau, beau_jour, est_le, le_grand, grand_jour) et les vecteurs caractéristiques des deux textes seraient (1, 1, 1, 1, 0, 0,0) et (1, 0, 0, 0, 1, 1, 1). Bien sûr, plus l'ordre des n-grammes utilisés pour générer les entités est élevé, plus les représentations seront plus claires.

En plus des fonctionnalités basées sur des mots, on peut également utiliser des classes de mots pour créer des fonctionnalités pour la classification de texte. Il existe de

nombreuses ressources lexicales qui peuvent être utilisées comme source de classes de mots, comme **"WordNet, Roget's Thesaurus"** ou "**Linguistic Inquiry and Word Count"** (LIWC; voir le chapitre 4, titre 4). Dans cette représentation, plutôt que d'utiliser des mots individuels, nous utilisons des classes de mots pour créer chaque entité. Par exemple, en supposant la classe de mots WE, y compris des mots tels que "*we", "us", "ourselves", "our"*, et "so on", et en supposant un schéma de pondération simple qui utilise la fréquence des mots, la valeur de cette fonctionnalité sera le nombre total d'occurrences de mots "WE" à l'intérieur du texte. En utilisant cette technique de création d'un poids cumulatif pour chaque classe de mots, nous pouvons générer des vecteurs d'entités qui incluent une entité pour chaque classe - par exemple, 80 entités si nous utilisons LIWC, et 700 à 1 000 entités si nous utilisons Roget, etc.

- **Pondération et sélection des fonctionnalités**

Compte tenu de l'espace d'entités de grande dimension généralement utilisé dans la classification de texte, la pondération et la sélection d'entités jouent un rôle important. La question est la suivante : comment pouvons-nous pondérer les fonctionnalités afin de donner plus de poids aux fonctionnalités qui sont plus importantes pour un texte donné que pour d'autres ?

Intuitivement, nous aimerions donner un poids faible à des mots tels que "is, a, have" et donner un poids élevé à des mots tels que "minage, salle de classe ou histoire". Même parmi ces trois derniers mots, si nous pensons, par exemple, aux sujets abordés dans ce rapport de thèse, "l'histoire" et "la classe" devraient avoir un poids inférieur à celui de l'exploitation minière.

Il existe plusieurs façons de créer des pondérations de fonction. La méthode la plus simple consiste à utiliser des poids binaires, qui sont 0 ou 1 selon qu'un mot (ou un autre gramme) apparaît ou non dans un texte. Une autre méthode consiste à utiliser le terme fréquence, qui compte le nombre d'occurrences d'un mot dans un texte. Une autre méthode encore est le terme fréquence inverse de la fréquence du document (tf idf), qui détermine le terme fréquence d'un mot, comme précédemment, puis il le divise par le nombre total de textes où ce mot apparaît.

- **Les algorithmes de classification de texte**

Une fois que les textes sont représentés comme des vecteurs de caractéristiques, ils peuvent être exécutés à travers n'importe quel algorithme de classification supervisé

pour classer automatiquement les nouveaux textes entrants appartenant à une ou plusieurs catégories. Certains de ces algorithmes produiront également une mesure de confiance associée à la classification, qui indiquera dans quelle mesure un certain élément de test peut être classé avec précision par le chargé de catégorisation automatique.

Dans ce chapitre, nous présentons deux classificateurs qui ont été plus largement utilisés en conjonction avec la classification de texte.

- **Algorithme Naive Bayes**

Bien que Naive Bayes soit l'un des premiers algorithmes de classification de texte, il reste l'une des méthodes de classification les plus utilisées. Naive Bayes est basé sur le théorème de Bayes de la théorie des probabilités, qui énonce la probabilité conditionnelle d'un événement C étant donné un événement T.

$$P(C \setminus T) = \frac{P(T \setminus C)P(C)}{P(T)}$$

Le théorème de Bayes peut être très facilement déduit de la probabilité d'événements conjoints. La probabilité que C et T se produisent ensemble peut s'écrire ainsi :

P(C,T)* = *P(C\T)P(T) ou alors ***P(C,T)* = *P(T\C)P(C)***.

Le théorème de Bayes découle directement de l'équation de ces deux manières différentes d'écrire P (C, T).

En supposant la représentation vectorielle caractéristique d'un texte T dont nous avons discuté précédemment, nous pouvons dire :

$$\langle t_1, t_2, ..., t_n \rangle$$

Dans la classification de texte, nous voulons trouver la catégorie Ci en C telle qu'elle ait la probabilité maximale donnée par le texte T. En d'autres termes, nous voulons trouver la catégorie qui satisfait :

$$C = \arg\max_{C_i \in C} P(C_i \setminus t_1, t_2, ..., t_n)$$

En utilisant le théorème de Bayes, nous cherchons à trouver :

$$C = \arg\max_{C_i \in C} \frac{P(t_1, t_2, ..., t_n \setminus C_i) P(C_i)}{P(t_1, t_2, ..., t_n)}$$

Et ceci, en tenant compte que le numérateur est le même quelle que soit la catégorie Ci considérée, nous pouvons réécrire ceci tel que :

$$C = \arg\max_{C_i \in C} P(t_1, t_2, ..., t_n) \mid C_i) P(C_i)$$

Une hypothèse importante faite dans l'algorithme Naive Bayes (et la raison de la « Naïf » dans son nom) est l'indépendance conditionnelle des caractéristiques d'une représentation textuelle. Nous supposons que les caractéristiques ti à l'intérieur d'un texte sont indépendantes les unes des autres, ce qui nous permet de réécrire la dernière équation ainsi :

$$C = \arg\max_{C_i \in C} P(C_i) \prod_{t_k \in T} P(T_k \mid C_i)$$

Cette réécriture finale rend l'algorithme traitable et facilement calculable. Nous pouvons calculer P (Ci) en comptant le nombre de textes dans les données d'apprentissage qui sont étiquetés avec la catégorie Ci et en divisant cela par le nombre total de textes dans les données d'apprentissage. Nous pouvons calculer P (tk | Ci) en comptant le nombre de textes dans les données d'apprentissage qui sont étiquetés avec la catégorie Ci et incluent la fonction tk, parmi tous les textes dans les données d'apprentissage qui sont étiquetées avec la catégorie Ci nous identifions :

$$P(C_i) = \frac{N(C = C_i)}{N} \quad et \quad P(t_k \mid C_i) = \frac{N(T_k = t_k, C = C_i)}{N(C = C_i)}$$

La dernière étape est le lissage, elle est souvent nécessaire et se réfère au processus de traitement des cas où il n'y a pas d'observations pour un certain événement (c.-à-d. zéro comptage). Pour y remédier, la dernière probabilité est souvent réécrite ainsi :

$$P(t_k \mid C_i) = \frac{N(T_k = t_k, C = C_i) + 1}{N(C = C_i) + k}$$

Avec k est la taille du vocabulaire (mots uniques) dans les données d‘entrainement.

- **Le bootstrapping dans la classification de textes**

Il existe un autre aspect intéressant de la classification automatique des textes est le fait de savoir comment utiliser efficacement les données textuelles brutes, qui sont souvent gratuites. Supposons, par exemple, que nous ayons 100 textes étiquetés pour leur sujet, et que nous ayons 1 000 000 de textes non étiquetés. La question est la suivante : ***comment utiliser les exemples non étiquetés pour améliorer la précision d'un classificateur de texte?***

Il existe plusieurs réponses à cette question, mais celle qui est souvent utilisée dans la classification de texte est le bootstrap. Il s'agit d'une méthode où nous formons un ou plusieurs classificateurs sur les données de formation existantes et étiquetons automatiquement les textes bruts. Nous sélectionnons ensuite les instances étiquetées avec un niveau de confiance élevé ; les déplacer vers l'ensemble de données de formation, augmentant ainsi sa taille; puis répétez le processus de formation et d'annotation. L'ensemble de données d'entrainement augmentera donc avec le temps et augmentera donc la précision du système.

Il existe différentes façons de fournir un score de confiance pour la classification. Les plus importantes sont les suivantes :

- La **méthode de l'auto-formation** : un seul classificateur est utilisé et le score de confiance de l'algorithme d'apprentissage lui-même est utilisé pour le processus d'amorçage.
- Le **Co-entrainement** : deux classificateurs sont utilisés en combinaison, et l'accord entre le classificateur est utilisé comme une indication de confiance. Dans cette méthode, seuls les cas où les deux classificateurs sont d'accord sont sélectionnés pour être ajoutés à l'ensemble de formation croissant.

Il faut dire qu'il y a plusieurs problèmes qui doivent être résolus avant d'appliquer le bootstrap dans le processus. Par exemple, une question importante est de savoir combien d'éléments à ajouter à l'ensemble d'entrainement à chaque itération. Une autre question concerne le nombre d'itérations que le classificateur doit exécuter, car la tendance générale au bootstrap est que la précision du ou des classificateurs augmente généralement pendant quelques itérations, suivie d'une baisse en raison du nombre croissant d'erreurs dans l'ensemble des données d'entrainement.

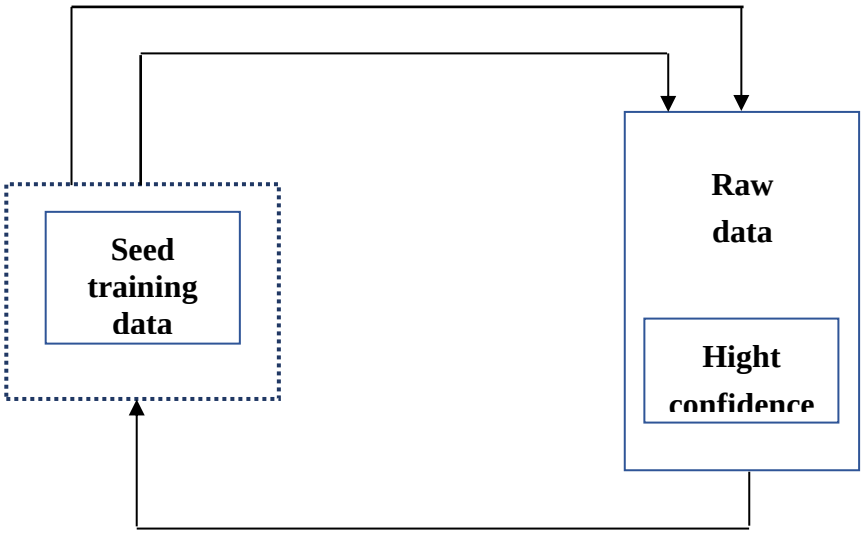


Figure 3.7 : La classification de textes en utilisant le bootstrapping

Dans ce contexte, la figure 3.7 illustre le processus d'amorçage général. Commençant par un ensemble de données d'entrainement, un ou plusieurs classificateurs sont formés et utilisés pour identifier les éléments qui peuvent être étiquetés en toute confiance dans un grand ensemble de données brutes. Ces éléments sont ajoutés à l'ensemble de données d'entrainement et le processus est répété pendant plusieurs itérations.

7. La fouille d'opinion

7.1. Que ce que la fouille d'opinion ?

Un type important d'informations véhiculées dans de nombreux types de discours écrits et parlés est l'état mental ou émotionnel de l'écrivain ou du locuteur ou de toute autre entité référencée dans le discours. Les articles de presse, par exemple, rapportent souvent des réponses émotionnelles à une histoire en plus des faits. Les éditoriaux, les critiques, les blogs et les discours politiques expriment les opinions, les croyances ou les intentions de l'écrivain ou du conférencier.

Les opinions sont des expressions d'états privés, tels que les émotions, les sentiments, les évaluations, les croyances et les spéculations en langage naturel. Les opinions ont des attributs, notamment qui exprime l'opinion, le (s) type (s) d'attitude exprimée, à propos de qui ou de quoi l'opinion est exprimée, le sentiment (ou la polarité) de l'opinion (c.-à-d. Qu'elle soit positive ou négative), etc.

La fouille d'opinion est définie comme la tâche d'identifier ces états privés dans la langue. Elle est généralement divisée en deux sous-tâches principales :

- L’analyse de la subjectivité, qui identifie si un texte contient une opinion, et étiquette en conséquence le texte comme étant subjectif ou objectif.
- L'analyse des sentiments, qui classe davantage une opinion (ou texte subjectif) comme positif, négatif ou neutre. Par exemple, considérez la phrase suivante :

‘The choice of Miers was praised by the Senate’s top Democrat, Harry Reid of Nevada‘

Dans cette phrase, l’expression « was praised by » indique qu’une opinion a été exprimée. L'opinion, selon l'auteur de la phrase, est exprimée par Reid, et il s'agit du choix de Miers, qui a été nommé à la Cour suprême par le président Bush en octobre 2005. Le type d'attitude est un sentiment (une évaluation, une émotion ou un jugement), et la polarité est positive (Wilson, 2008).

On peut juger de la subjectivité et du sentiment (ou polarité) des textes à plusieurs niveaux différents. Au niveau du document, on peut se demander si un texte comprend une opinion et, si oui, si cette opinion est principalement positive ou négative. Nous pouvons effectuer une analyse plus fine et demander si une phrase contient une subjectivité. Par exemple, considérons les exemples suivants de Wilson (2008). La première phrase qui suit est subjective (et a une polarité positive), mais la seconde est objective, car elle ne contient aucune expression subjective :

“He spins a riveting plot, which grabs and holds the reader’s interest. The notes do not pay interest. “

De plus, les expressions individuelles peuvent être jugées - par exemple, que les rotations, le rivetage et l'intérêt pour la première phrase qui vient d'être donnée sont des expressions subjectives. Un exemple plus intéressant apparaît dans cette phrase : “Bravo à Timothy Whitfield pour les visuels merveilleusement horribles “. Alors que le mot “horrible“ serait répertorié comme ayant une polarité négative dans un lexique de subjectivité au niveau des mots, dans ce contexte, il est utilisé positivement : les mots “merveilleusement horribles“ expriment un sentiment positif envers les visuels (de même, “bravo“ exprime un sentiment positif envers Timothy Whitfield).

On peut aussi classer les sens des mots selon leur subjectivité et leur polarité. Prenons, par exemple, les deux sens d'intérêt suivants de WordNet (Miller, 1995) :

“Interest, involvement— (a sense of concern with and curiosity about someone or something; ‘an interest in music’)”

“Interest—a fixed charge for borrowing money; usually a percentage of the amount borrowed; ‘how much interest do you pay on your mortgage?’”

Le premier sens est subjectif, avec une polarité positive. Mais le second sens ne l'est pas (les sens non subjectifs sont appelés sens objectifs) - il ne fait pas référence à un état privé. Les lexiques de subjectivité au niveau des mots et des sens sont importants car ils sont des ressources utiles pour l'analyse contextuelle de la subjectivité (Wilson, 2008) - reconnaître et extraire des expressions d'état privé dans un texte ou un dialogue réel.

Concepts basiques :

Une opinion est l'expression d'un état privé, tel qu'une émotion, un sentiment, une évaluation, une croyance ou une spéculation. La fouille d'opinion consiste à identifier les opinions dans le texte.

- *La fouille d'opinion* se déroule généralement en deux étapes :
- *L’analyse de la subjectivité,* dans laquelle les opinions sont identifiées, et l'analyse des sentiments, dans laquelle une opinion est classée en fonction de sa polarité (positive, négative, neutre).
- *L'analyse de la subjectivité et des sentiments* peut être effectuée à différents *niveaux de granularité* : documents, phrases, mots ou phrases et sens des mots.

7.2. Ressources pour la fouille d’opinions

Il existe deux principaux types de ressources utilisées dans l'exploration d’opinion :

- Les lexiques, qui consistent en de vastes listes de mots et de phrases annotées avec une étiquette de subjectivité, de sentiment et / ou d'émotion.
- Les corpus, qui sont des collections de phrases ou de courts documents étiquetés pour la subjectivité ou le sentiment.

Ces ressources constituent les bases de méthodes automatiques supervisées ou non supervisées pour identifier les opinions dans le texte, comme décrit dans la section suivante.

- **Les lexiques**

L'un des lexiques les plus fréquemment utilisés est peut-être le lexique de subjectivité et de sentiment fourni avec la distribution OpinionFinder (Wiebe, Wilson et Cardie, 2005). Le lexique a été compilé à partir de ressources développées

manuellement augmentées d'entrées tirées des corpus. Il contient 6 856 entrées uniques, dont 990 sont des expressions multi-mots. Les entrées du lexique ont été étiquetées pour une partie du discours ainsi que pour la fiabilité - celles qui apparaissent le plus souvent dans des contextes subjectifs sont de “forts“ indices de subjectivité, tandis que celles qui apparaissent moins souvent, sont étiquetées “faible“.

Chaque entrée est également associée à une étiquette de polarité, indiquant si le mot ou la phrase correspondant est positif, négatif ou neutre. Pour illustrer cela, considérons l'entrée suivante du lexique OpinionFinder *type=strongsubj word1=agree pos1=verb mpqapolarity=weakpos*, qui indique que le mot d'accord lorsqu'il est utilisé comme verbe est un indice fort de subjectivité et a une polarité faiblement positive.

Un autre lexique souvent utilisé dans l'analyse de polarité est le projet General Inquirer (Stone, 1968). Il s'agit d'un dictionnaire d'environ 10 000 mots regroupés en environ 180 catégories, qui ont été largement utilisés pour l'analyse de contenu. Il comprend des classes sémantiques (par exemple, animées, humaines), des classes de verbes (par exemple, des négatifs, devenir des verbes), des classes d'orientation cognitive (par exemple, causales, sachant, perception) et autres. Les classes de valence, qui forment un lexique de 1 915 mots positifs et 2 291 mots négatifs, sont deux des catégories les plus importantes du General Inquirer.

SentiWordNet (Esuli & Sebastiani, 2006b) est une ressource pour la fouille d'opinion construite au-dessus de WordNet, qui attribue à chaque synset dans WordNet un triplet de score (positif, négatif et objectif), indiquant la force de chacune de ces trois propriétés pour le mot dans le synset. Les annotations SentiWordNet ont été générées automatiquement, en commençant par un ensemble de synsets étiquetés manuellement. Actuellement, SentiWordNet inclut une annotation automatique pour tous les synsets dans WordNet, totalisant plus de 100 000 mots.

- **Les corpus**

Les corpus annotés de subjectivité et de sentiment sont utiles non seulement comme moyen de former des classificateurs automatiques, mais aussi comme ressources pour extraire des lexiques de fouille de l'opinion.

Le corpus MPQA (Wiebe et al., 2005) a été collecté et annoté dans le cadre d'un atelier de 2002 sur la réponse aux questions multiperspectives. Il s'agit d'une collection de 535 articles de presse en anglais provenant de diverses sources d'information annotées manuellement pour des opinions et d'autres états privés (c.-à-d. Croyances,

émotions, sentiments, spéculations). Le corpus était à l'origine annoté au niveau des clauses et des phrases, mais les annotations au niveau des phrases associées à l'ensemble de données peuvent également être dérivées via de simples heuristiques (Wiebe et al., 2005).

Un autre corpus annoté manuellement est la collection de titres de journaux créés et utilisés lors de la récente tâche SemEval sur « Texte affectif » (Strapparava et Mihalcea, 2007). L'ensemble de données comprend 1 000 titres de test et 200 titres de développement, chacun d'eux annoté avec les six émotions d'Ekman (colère, dégoût, peur, joie, tristesse, surprise) et leur orientation de polarité (positive, négative).

Deux autres ensembles de données, couvrant tous les deux le domaine des critiques de films, sont un ensemble de données de polarité composé de 1000 critiques positives et 1000 négatives ainsi qu'un ensemble de données de subjectivité composé de 5000 phrases subjectives et 5000 phrases objectives. Les deux ensembles de données ont été introduits dans Pang et Lee (2004) et ont été utilisés pour former des classificateurs de fouille d'opinion.

Compte tenu de la spécificité des domaines de ces collections, il s'est avéré qu'elles conduisaient à des classificateurs précis pour les données appartenant aux mêmes domaines ou à des domaines similaires. Plus récemment, un ensemble de données sur les critiques de films beaucoup plus important a été introduit (Maas et al., 2011), comprenant 50 000 critiques complètes recueillies sur le site Web de l'IMDB.

La recherche en analyse des sentiments a également bénéficié du nombre croissant de critiques de produits disponibles en ligne, sur des sites Web tels qu'Amazon, qui peuvent être utilisées pour créer de très grands ensembles de données annotées sur les sentiments (Hu et Liu, 2004). Ces revues sont généralement disponibles dans de nombreuses langues, permettant ainsi la construction d'outils d'analyse des sentiments dans d'autres langues que l'anglais (Nakagawa, Inui et Kurohashi, 2010).

7.3. Approches de la fouille d’opinion

Il existe un grand nombre d'approches développées à ce jour pour l'analyse des sentiments et de la subjectivité en anglais. Les méthodes peuvent être grossièrement classées en deux catégories :

- Les systèmes basés sur des règles, s'appuyant sur des lexiques construits manuellement ou semi automatiquement.

- Des classificateurs d'apprentissage automatique, formés sur des corpus annotés d'opinion.

Parmi les systèmes basés sur des règles, l'un des plus utilisés est OpinionFinder (Wiebe et al., 2005), qui annote automatiquement la subjectivité d'un nouveau texte en fonction de la présence (ou de l'absence) de mots ou de phrases dans un grand lexique. En bref, le classificateur de haute précision OpinionFinder s'appuie sur trois heuristiques principales pour étiqueter les phrases subjectives et objectives :

- Si deux ou plusieurs expressions subjectives fortes se produisent dans la même phrase, la phrase est étiquetée subjective.
- Si aucune expression subjective forte ne se produit dans une phrase, et deux expressions subjectives faibles au plus se produisent dans la phrase précédente, actuelle et suivante combinées, alors la phrase est étiquetée objectif.
- Si aucune des règles précédentes ne s'applique, la phrase est étiquetée inconnue.

Le classificateur utilise les indices d'un lexique de subjectivité et les règles mentionnées précédemment pour récolter des phrases subjectives et objectives à partir d'une grande quantité de texte non annoté ; les données sont ensuite utilisées pour identifier automatiquement un ensemble de modèles d'extraction, qui sont ensuite utilisés de manière itérative pour identifier un plus grand ensemble de phrases subjectives et objectives.

En plus du classificateur de haute précision, OpinionFinder comprend également un classificateur à couverture élevée. Ce classificateur de haute précision est utilisé pour produire automatiquement un ensemble de données étiqueté en anglais, qui peut ensuite être utilisé pour former un classificateur de subjectivité à couverture élevée. Lorsqu'il a été évalué sur le corpus MPQA, le classificateur de haute précision s'est révélé conduire à une précision de **86,7%** et un rappel de **32,6%**, tandis que le classificateur à couverture élevée a une précision de **79,4%** et un rappel de **70,6%.**

Un autre système non supervisé qui mérite d'être mentionné, basé cette fois sur des mots ou expressions étiquetés automatiquement, est celui proposé dans Turney (2002), qui s'appuie sur des travaux antérieurs de Hatzivassiloglou et McKeown (1997). En commençant par deux mots de référence – "excellent" et "pauvre" - Turney a classé la polarité d'un mot ou d'une phrase en mesurant la fraction entre son information mutuelle ponctuelle (PMI) avec la référence positive (excellent) et le PMI avec la référence négative (pauvre). Le PMI de deux mots w1 et w2 est défini comme la

probabilité de voir les deux mots ensemble divisée par la probabilité de voir chaque mot individuel : **PMI (w1, w2) = p (w1, w2) / p (w1) p (w2).**

Les scores de polarité attribués de cette manière sont utilisés pour annoter automatiquement la polarité des critiques de produits, de sociétés ou de films. Notez que ce système est totalement non supervisé et donc particulièrement attrayant pour une application dans d'autres langues. Lorsque des corpus annotés sont disponibles, les méthodes d'apprentissage automatique sont un choix naturel pour construire des classificateurs de subjectivité et de sentiment.

Par exemple, Wiebe, Bruce et O'Hara (1999) ont utilisé un ensemble de données annoté manuellement pour la subjectivité afin de former un classificateur d'apprentissage automatique, ce qui a conduit à des améliorations significatives par rapport à la ligne de base. De même, en commençant par des ensembles de données construits de façon semi-automatique, Pang et Lee (2004) ont construit des classificateurs pour l'annotation de subjectivité au niveau de la phrase, ainsi qu'un classificateur pour l'annotation des sentiments au niveau du document. Dans la mesure où des données annotées sont disponibles, ces classificateurs d'apprentissage automatique peuvent être utilisés aussi bien dans d'autres langues.

Récemment, un outil d'analyse des sentiments basé sur des techniques d'apprentissage approfondi a été introduit conjointement avec un Sentiment Treebank (Socher et al., 2013), où des étiquettes de sentiments à grain fin au niveau des mots et des phrases sont utilisées avec un arbre d'analyse pour composer le sentiment d'un texte. Contrairement à la plupart des méthodes précédentes qui supposaient qu'un texte avait un sentiment cohérent, cette méthode de composition permet des changements de sentiment à l'intérieur d'un texte, comme dans « J'aime généralement le téléphone, mais je n'aime pas beaucoup le petit clavier », où un sentiment positif et un sentiment négatif sont mélangés dans la même phrase.

8. Extraction de l'information

Dans cette section, nous passons en revue les principales directions de recherche axées sur le développement de méthodes et d'outils IE (Extraction d'Informations). Plus précisément, nous abordons en détail les sujets de l'extraction d'entités et de relations, nous présentons les travaux récents dans Web IE et discutons également de la tâche de remplissage de modèle qui combine plusieurs extracteurs d'informations ensemble.

Il existe plusieurs systèmes IE à l'échelle du Web qui ont été construits à ce jour. À des fins d'illustration, nous en décrivons brièvement deux.

- Le premier système est KnowItAll (Etzioni et al., 2004), qui extrait des faits et des relations du Web. Le système est enrichi d'informations obtenues à partir d'une ontologie, ainsi que de quelques modèles génériques, qu'il utilise ensuite pour créer des règles d'extraction de texte. Par exemple, le modèle général "NounPhrase1 tel que NounPhraseList" peut être appliqué sur du texte avec certaines graines comme Paris et Londres sont des villes pour déduire un modèle syntaxique générique de la forme "villes telles que <?>" qui peut être utilisé pour extraire des noms de villes supplémentaires. Notez que les règles de modèle sont génériques et indépendantes du domaine et peuvent donc être automatiquement instanciées pour l'extraction de diverses entités (par exemple, les villes, les pays, les couleurs). Les requêtes formées à partir de ses règles d'extraction de texte sont exécutées sur un moteur de recherche et les informations obtenues sont validées par un module statistique dans KnowItAll qui évalue la probabilité d'exactitude pour chaque élément d'IE. Les informations sont ensuite stockées dans une base de données pour une analyse plus approfondie. Alors que la version initiale de KnowItAll comprenait un peu plus de 50 000 faits, des suivis tels que ReVerb (Fader, Soderland et Etzioni, 2011) et TextRunner (Banko, Cafarella, Soderland, Broadhead et Etzioni, 2007) avaient une couverture beaucoup plus large, y compris plus de 3 millions d'entités et 600 000 extractions de relations, entre autres éléments de connaissance.
- Le deuxième système Web est “Never-Ending Language Learning“ (NELL; Carlson et al., 2010), qui est un autre système d'extraction d'informations (IE) Web qui crée des « croyances » candidates en traitant un très grand nombre de pages Web.

9. Remplissage du modèle

Il existe de nombreuses situations où les informations extraites sont liées les unes aux autres en ce qu'elles représentent différents aspects du même type de situation ou d'événement. Par exemple, si nous parlons d'une attaque terroriste, les informations pertinentes seraient le lieu de l'attaque, la date et l'heure de l'attaque, le groupe derrière l'attaque, le nombre de victimes, etc. Ces aspects d'un événement sont souvent appelés créneaux horaires et forment un modèle pour cet événement. Le processus de recherche des valeurs pour chacun des emplacements est appelé remplissage des emplacements, et

le processus global de remplissage des valeurs pour tous les aspects d'un modèle est appelé remplissage du modèle.

Dans la plupart des cas, des algorithmes IE distincts peuvent être formés en suivant le même ensemble d'étapes que celui décrit ci-dessus :

- On annote d'abord un ensemble de documents avec les informations d'intérêt (c'est-à-dire, chacun des emplacements marqués explicitement dans le texte) ou on identifie un ensemble de semences valeurs pour chaque emplacement.
- Un classificateur est ensuite formé pour reconnaître les modèles associés à ces occurrences d'emplacement.
- Un processus d'amorçage est ensuite mis en place afin d'agrandir automatiquement les listes de modèles et les valeurs possibles pour chaque emplacement.

Des travaux récents ont également envisagé l'application conjointe des algorithmes IE pour les créneaux individuels, avec l'idée qu'il pourrait y avoir des dépendances entre les créneaux. Par exemple, le lieu des attaques terroristes pourrait avoir une certaine association avec les groupes derrière les attaques et ainsi de suite. C'est ce qu'on appelle l'apprentissage des aspects conjoints ou le remplissage des fentes communes (Mukherjee et Liu, 2012).

10. Conclusion

IE est l'une des tâches principales de la fouille de texte, utilisée pour transformer du texte non structuré en données structurées. Alors que certains des outils les plus courants pour IE se concentrent sur quelques types d'entités pour lesquels de grandes quantités de données de formation sont disponibles (par exemple, les emplacements), les méthodes existantes pour IE peuvent facilement être personnalisées pour de nouveaux types. Il est important de noter que des approches proposées pour répondre à cette tâche peuvent être amorcées, ce qui ne nécessite que de petites quantités d'exemples annotés soient couplées à de grandes quantités de données non étiquetées.

Ce chapitre a présenté les principales approches pour IE, couvrant les méthodes d'extraction d'entités et d'extraction de relations, ainsi que les approches visant à traiter les informations à grande échelle, souvent appelées IE (open) Web.

Chapitre 4 : RCrawler

1. Introduction

Un web crawler ou robot d'indexation est généralement connu efficace pour collecter des pages Web, mais lorsqu'un robot peut également effectuer une extraction de données pendant l'analyse, il peut être appelé aussi un grattoir Web ou Web scraper. Dans ce cadre, ce chapitre décrit l'architecture et l'implémentation de RCrawler, un web crawler et un web scraper basé sur R que nous avons développé durant nos études doctorales.

La motivation de cette étude provient de la nécessité de créer un robot d'indexation Web basé sur R qui peut analyser et extraire le contenu des pages Web (articles, titres et métadonnées) de manière automatisée pour produire un ensemble de données structuré. L'aspect difficile a été la mise en œuvre d'un robot parallèle dans un environnement principalement dédié aux calculs et au calcul statistique, plutôt qu'au traitement automatique des données. Ainsi, pour permettre la collecte de données dans l'environnement R, notre défi était de surmonter ses faiblesses et d'adapter l'environnement à nos besoins.

R est un environnement logiciel très efficace pour l'analyse statistique et le traitement des données, et fournit également un support puissant pour la fouille du Web (Zhao Y., 2012). En fait, R fournit un large ensemble de fonctions et de packages qui peuvent gérer les tâches de fouille de sites Web (Thomas, Scott, Patrick, Karthik et Christopher, 2016). Des packages R sont disponibles pour les processus de collecte de données, tels que Rvest, tm.plugin.webmining et scrapeR. Cependant, ces packages ne fournissent pas d'analyse de base, car ils peuvent uniquement analyser (Munzert, Rubba, Meißner et Nyhuis, 2014) et extraire le contenu des URL, que l'utilisateur doit collecter et fournir manuellement. Par conséquent, ils ne peuvent pas parcourir les pages Web et collecter automatiquement les liens et les données.

Par exemple, dans la plupart des cas, les utilisateurs s'appuient sur des outils externes pour effectuer cette tâche. Dans notre projet, nous visons à incorporer le processus de fouille dans l'environnement R, afin d'offrir une plate-forme de flux de données complète comprenant les étapes avant et après l'analyse réelle des données. En fait, à partir d'une URL donnée, RCrawler peut explorer et analyser automatiquement toutes les URL de ce domaine et extraire du contenu spécifique de ces URL qui

correspond le mieux aux critères de l'utilisateur. Le tableau 4.1 présente une comparaison de certains packages de collecte de données populaires et illustre l'utilité de notre nouveau package.

Tableau 4.1. Comparaison de certains packages R populaires pour la collecte de données

Nom du package	Indexation	Récupération	Analyse	Description
ScrapeR	Non	Oui	Oui	À partir d'un vecteur d'URL donné, ce package récupère les pages Web et les analyse pour extraire les informations d'intérêt à l'aide d'un modèle XPath.
tm.plugin.webmining	Non	Oui	Oui	Ce package suit les liens sur les formats de flux Web tels que XML et JSON, et extrait le contenu à l'aide de la méthode de chaudronnerie.
Rvest	Non	Oui	Oui	Enveloppe les packages xml2 et httr afin qu'ils puissent facilement télécharger puis manipuler HTML et XML.
RCrawler	Oui	Oui	Oui	Analyse les sites Web et extrait leur contenu à l'aide de diverses techniques.
Quelques outils de base				
XML, XML2	Non	Non	Oui	Analyseurs HTML / XML
jsonlite, RJSONIO	Non	Non	Oui	Analyseur JSON
RSelenium	Non	Non	Oui	Automatisation du navigateur
Selectr	Non	Non	Oui	Analyse les sélecteurs CSS3 et les traduit en XPath 1.0
Httr, RCurl	Non	Oui	Non	Gère les requêtes HTTP / HTTPS

Comme décrit dans le tableau 4.1, scrapeR et rvest nécessitent une liste d'URL à fournir à l'avance. Pendant ce temps, tm.plugin.webmining peut obtenir des améliorations, car il peut récupérer des URL à partir de certains formats de flux tels que

XML et JSON, mais son utilisation est toujours limitée par le fait que tous les sites Web n'ont pas de flux, et même si un flux existe, il peut ne pas contenir l'arborescence complète du site Web. Tous ces packages peuvent récupérer et analyser efficacement des pages Web spécifiques. Cependant, la faiblesse commune de ces outils est qu'ils sont limités dans le traitement de plusieurs demandes, et toute tentative de boucle sur la même fonction pour une liste d'URL peut entraîner des erreurs ou des plantages en raison des contraintes de politesse.

RCrawler peut facilement être appliqué à de nombreux problèmes de fouille de contenu Web, tels que l'exploration d'opinion (Pang et Lee, 2008), la détection d'événements ou de sujets et les systèmes de recommandation. De plus, sa fonctionnalité lui permet d'être étendu pour explorer des structures de sites Web spécifiques. Cependant, son utilisation est encore limitée aux petits projets, en raison de la nature de l'environnement R, qui n'est pas entièrement dédié à la gestion de l'exploration massive de données.

2. Objectifs et exigences

Dans cette partie, nous décrivons les exigences fonctionnelles et les objectifs système qui ont guidé notre implémentation du robot. Il y a cinq exigences principales :

1. **R-natif**: Habituellement, lorsque les utilisateurs R ont besoin d'explorer et de gratter une grande quantité de données automatiquement, ils se tournent vers des outils externes pour collecter des URL ou effectuer la tâche complète, puis importent les données collectées dans R. Ainsi, nos principaux raisons de l'écriture de RCrawler étaient de prendre en charge la fouille et le raclage Web dans l'environnement R. Ainsi, notre solution devrait être implémentée nativement dans R.
2. **Parallélisme** : RCrawler devrait tirer parti du parallélisme, afin d'obtenir des améliorations de performances significatives et d'utiliser efficacement les différentes ressources système, y compris les processeurs.
3. **Politesse**: le robot doit être capable d'analyser et d'obéir aux commandes robots.txt. De plus, le robot doit éviter de demander trop de pages dans un court intervalle de temps à un hôte donné.

4. **Efficacité**: Notre solution doit utiliser intelligemment les ressources et être résistante aux pièges à araignées. Le robot doit pouvoir détecter et éviter les URL ou les pages Web en double.
5. **Flexibilité et reconfiguration** : nous souhaitons concevoir un système flexible pouvant être appliqué dans différents scénarios. Voici une liste résumée des options de paramètres pour RCrawler:

- Nom et répertoire du projet.
- Agent utilisateur, délai d'expiration de la connexion et délai de demande.
- Filtres: inclure / exclure le type de contenu (MIME), les pages d'erreur, l'extension de fichier et les URL correspondant à un modèle d'expression régulière spécifique.
- Exploration parallèle: l'utilisateur spécifie le nombre de nœuds et le nombre de connexions (requêtes simultanées).
- Choix d'honorer le fichier Robots.txt ou de l'ignorer.
- Niveau de profondeur maximum pour contrôler la profondeur de fouille.
- Le robot doit permettre d'appliquer certaines fonctions spécifiques d'extraction de contenu, telles que les modèles XPath, à chaque page analysée pendant le processus d'exploration.

En ce qui concerne la visualisation des résultats, le robot doit renvoyer les données analysées dans des structures de données bien organisées et prêtes à l'emploi, telles que des vecteurs, des listes et des trames de données. En fait, les données renvoyées par d'autres packages sont toujours réparties, et des efforts supplémentaires sont nécessaires pour organiser cela en une seule structure de données. Nous décrivons les structures de données pour RCrawler comme suit :

- Un bloc de données représentant l'index URL générique, y compris la liste des URL récupérées et les détails de la page (type de contenu, état HTTP, nombre de liens sortants et entrants, type d'encodage et niveau).
- Un référentiel de fichiers qui contient toutes les pages téléchargées.
- Un vecteur pour le contenu exploré.
- Un message incluant les statistiques d'exploration.
- Pour l'analyse des liens, une structure de données est nécessaire pour représenter la connectivité du graphique Web (bords).
- Pendant le processus d'analyse, le robot doit afficher l'état d'analyse.

3. L’architecture RCrawler

Les robots d’indexation du Web doivent relever simultanément plusieurs défis, dont certains se contredisent (Olston et Najork, 2010). Inspirés par des travaux antérieurs, tels que Mercator (Heydon et Najork, 1999) et Ubicrawler (Boldi, Codenotti, Santini et Vigna, 2004), nous avons tenté de rendre l'architecture du robot aussi optimisée et simple que possible afin d'éviter de surcharger l'environnement hôte (voir figure 5.1).

Le robot commence à partir d'un URL de site Web donné, fourni par l'utilisateur, et le récupère progressivement afin d’extraire de nouveaux URL (liens sortants). Ces derniers sont à leur tour ajoutés à la liste des frontières à traiter. Le processus de fouille s'arrête lorsque tous les URL de la frontière sont traitées.

Premièrement, le robot lance l'environnement de travail, comprenant la structure d'index, le référentiel qui contiendra la collection de documents Web et les nœuds de cluster (travailleurs) pour le calcul parallèle. L'analyse est effectuée par plusieurs threads de travail et le composant gestionnaire de pool de travail prépare un pool d'URL à traiter en parallèle. Ensuite, chaque nœud exécute les fonctions suivantes pour l'URL donné :

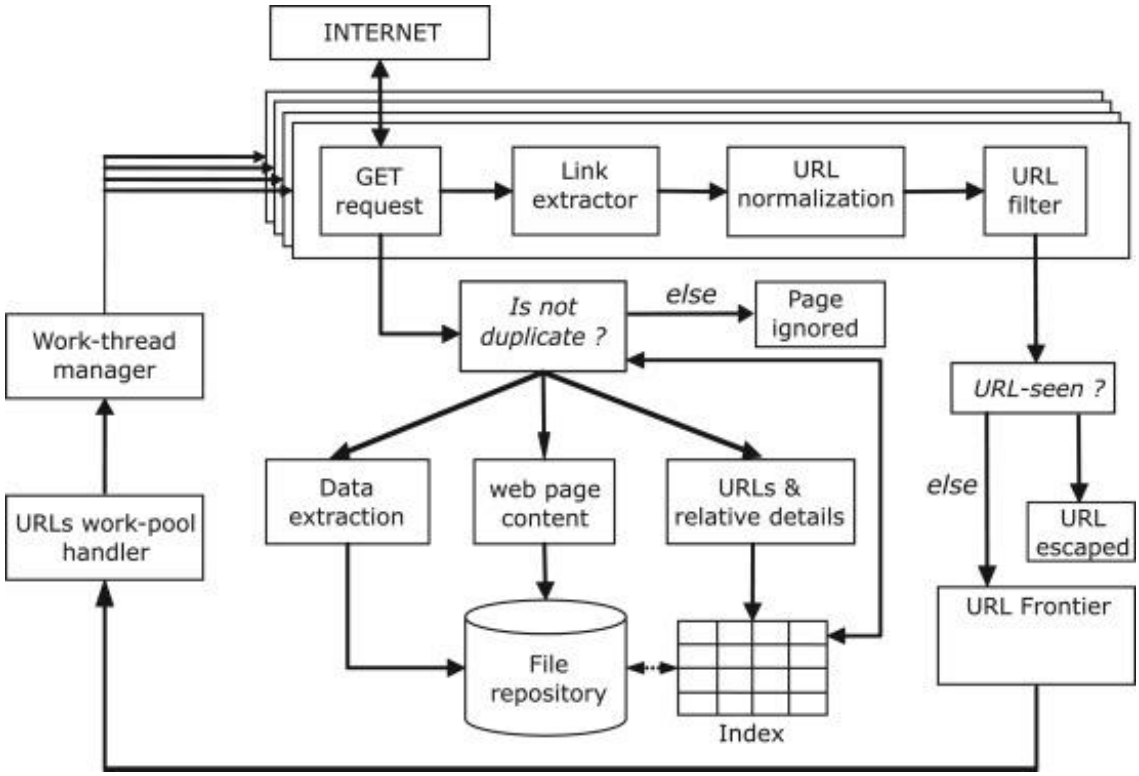
1. Télécharger le document correspondant et son en-tête HTTP à l'aide d'une requête GET.
2. Analyser et extraire tous les liens contenus dans le document.
3. Procéder à la canonisation et à la normalisation des URL.
4. Appliquer un filtre d'URL en ne conservant que les URL qui correspondent à la configuration fournie par l'utilisateur, au type de fichier et aux URL spécifiques au domaine.

Par conséquent, pour chaque URL, chaque travailleur renvoie son document HTML, les détails de l'en-tête HTTP et la liste des liens sortants découverts. La fonction URL vu vérifie si l'URL a déjà été traité ou mis en file d'attente avant d'être ajouté à la frontière. Avant de stocker le document dans le référentiel, le *Is-not-duplicate?* vérifie qu'il n'a pas été traitée avec un URL différent. Sinon, il est jeté. Nous préférons écrire la liste des frontières et l'index sur le disque à chaque itération du robot, pour les protéger contre la perte de données en cas de panne du robot. Voici les principales composantes de RCrawler :

- Gestionnaire de requêtes HTTP : gère les requêtes HTTP.
- Extracteur de liens : analyse et extrait les liens des documents analysés.

- Vérificateur de doublons de contenu : détecte les documents en double.
- Gestionnaire de threads de travail : gère le multithreading et le calcul parallèle.
- Extraction de données : un composant pour analyser et extraire le contenu d'une page Web.
- Index : une structure de données pour stocker des informations concernant les pages Web explorées.
- D'autres composants liés aux pièges d'araignée, à la canonisation d'URL, à l'analyseur robot.txt et à d'autres fonctionnalités sont abordés dans les sections suivantes.

Figure 4.1 Architecture et principales composantes de Rcrawler



4. Fonctionnalités principales et implémentation

Cette section présente les fonctionnalités de la version actuelle de RCrawler.

4.1. Indexation parallèle – multithreading

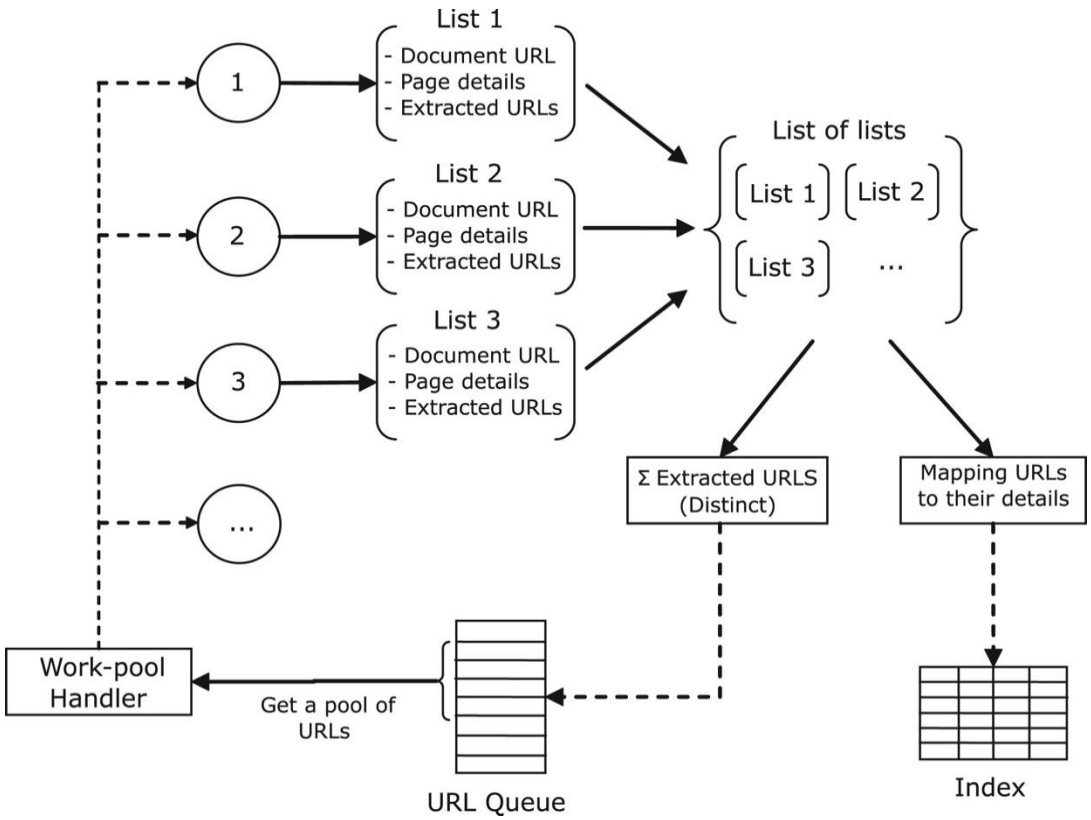
Le moyen le plus simple d'améliorer les performances et l'efficacité du robot est le calcul parallèle (Castillo, 2005). Cependant, la mise en œuvre d'un robot multi-threadé sous R implique de nombreuses limites et divers défis.

Le processus d'une opération d'analyse est effectué par plusieurs processus ou nœuds simultanés en parallèle, comme illustré sur la figure 5.2. Essentiellement, chaque nœud en cours d'exécution est affecté à un thread de contrôle physique fourni par le système d'exploitation et fonctionne indépendamment. Pour gérer le calcul parallèle dans un environnement R (Eddelbuettel, 2016), nous avons utilisé les packages *doParallel et parallèl* packages (Package 'parallel', 2015 ; Rich, Steve et Dan, 2015).

Tout d'abord, nous démarrons N processus de nœuds (avec N spécifié par l'utilisateur). Ensuite, nous initialisons tous les nœuds avec des bibliothèques, des fonctions et des structures de données partagées. Cependant, vu la difficulté de gestion d'accès des nœuds aux structures de données partagées dans R, nous avons limité cet accès en lecture seule. Par conséquent, toutes les opérations d'écriture sont effectuées par la fonction de fouille principale. Par la suite, le gestionnaire de pool de travail sélectionne un pool d'URL à partir de la frontière, à traiter en parallèle par les nœuds (la taille du pool est similaire au nombre de demandes simultanées). Notez que l'occurrence de trop de demandes simultanées à partir d'une seule adresse IP peut ressembler à une attaque DOS (demandes fréquentes qui empêchent le serveur de servir les demandes des clients authentiques).

Pour éviter de telles situations, il est recommandé de limiter le nombre de demandes actives à un hôte donné. Le package foreach (Rich et Steve, 2015) fournit la structure de boucle de base et utilise le *backend* parallèle afin d'exécuter les tâches et renvoyer une structure de données représentant la collection de résultats pour tous les nœuds. Notez que la boucle *foreach* renvoie une liste d'éléments dans le même ordre que le compteur, même lorsqu'elle s'exécute en parallèle. Ainsi, chaque nœud traite un URL donné et renvoi une structure de données de liste avec trois éléments (le document URL, les détails d'en-tête HTTP et la liste des URL extraits). Ces listes sont ensuite retournées ensemble sous la forme d'une collection unique (une collection de listes), à isoler et à mapper à leur URL d'origine par la fonction principale (voir figure 4.2). Le robot doit éviter de demander trop de pages à un hôte donné dans un court intervalle de temps. Par conséquent, nous avons introduit une fonction *RequestDelay* entre les requêtes adressées au même hôte. Enfin, les processus de noeud sont arrêtés et supprimés lorsque l'analyse est terminée.

Figure 4.2 : Conception de notre implémentation multithreading.



4.2. Gestionnaire de requêtes http

Le gestionnaire de requêtes HTTP est la composante en charge des connexions http. Elle fournit des fonctions pratiques pour composer des requêtes HTTP générales, récupérer des URL, obtenir et publier des formulaires, etc. Un robot efficace doit fournir un large degré de contrôle sur les connexions HTTP et les configurations des demandes. En fait, cela devrait permettre à l'utilisateur de définir certaines propriétés HTTP, telles que les suivantes :

- **User-agent :** lors de la fouille d'un site Web, RCrawler s'identifie par défaut comme RCrawler. Cependant, certaines sociétés d'hébergement Web peuvent bloquer un agent utilisateur s'il ne s'agit pas d'un navigateur ou s'il fait trop de demandes. Ainsi, il est important de changer le référent (agent utilisateur) pour simuler différents navigateurs et continuer à explorer sans être banni.
- **Time-out** : le temps de requête maximum, c'est-à-dire le nombre de secondes pour attendre une réponse jusqu'à l'abandon, afin d'éviter de perdre du temps à attendre les réponses des serveurs lents ou des pages volumineuses.

Certaines propriétés d'en-tête supplémentaires peuvent être ajoutées aux requêtes HTTP pour contrôler la manière dont le serveur traite les requêtes entrantes ou pour fournir des informations supplémentaires concernant le contenu demandé.

Pour améliorer les performances du robot, nous avons limité le nombre de demandes à une par URL. Par conséquent, la réponse doit inclure le contenu et toutes les informations requises pertinentes (code de réponse, type de contenu, codage, etc.). Le package httr (Hadley, 2016) a été importé pour gérer les connexions HTTP, car il fournit un ensemble riche de fonctions pour travailler avec HTTP et les URL. De plus, il offre une interface de niveau supérieur utilisant des connexions de socket R.

La gestion des exceptions et la vérification des erreurs sont des considérations importantes pendant le processus de récupération de page, afin d'éviter les plantages à chaque demande ayant échoué. En fait, le robot vérifie les propriétés de réponse avant d'analyser le contenu. S'il est nul, le nom de domaine n'existe pas ou n'a pas pu être résolu. Sinon, si le serveur hôte URL a émis un code d'erreur HTTP ou un type de contenu non pris en charge, les documents URL ne sont pas analysés et, selon la configuration utilisateur, cela peut ou non être indexé.

4.3. Analyse HTML et extraction de liens

Après avoir récupéré une page Web, chaque nœud d'exploration doit analyser son contenu pour extraire les hyperliens et les informations pertinentes nécessaires à l'indexation. Cependant, l'implémentation de l'analyse syntaxique pour l'extraction de contenu à l'intérieur des nœuds peut ralentir le processus d'analyse et entraîner une répartition de charge déséquilibrée.

L'analyse est implémentée à l'aide de la bibliothèque *xml2* (Hadley et Jeroen, 2015), qui est une interface simple et cohérente construite au-dessus de la bibliothèque C *libxml2*. Notez que nous analysons uniquement les hyperliens ou l'extraction de données spécifiques via le modèle XPath.

Premièrement, le contenu est récupéré de la requête HTTP et le document entier est analysé dans un modèle d'objet de document (DOM) (Le Hégaret, Whitmer et Wood, 2005) dans une structure de données arborescente en langage C. Dans cette structure de données, chaque élément qui se produit dans le HTML est maintenant représenté comme sa propre entité ou comme un nœud individuel. Tous les nœuds sont appelés ensemble de nœuds. Le processus d'analyse comprend également une étape de validation automatique des malformations. En fait, xml2 est capable de travailler sur des

documents HTML mal formés, car il reconnaît les erreurs et les corrige pour créer un DOM valide. Dans l'étape suivante, la structure de noeud de niveau C est convertie en un objet du langage R, via des fonctions dites de gestionnaire, qui gèrent les transformations entre C et R.

Ceci est nécessaire pour un traitement ultérieur du DOM. Enfin, l'extraction de liens est effectuée en extrayant tous les nœuds qui correspondent aux balises <href> dans le document, via la fonction *xml_find_all*. Par la suite, pour chaque nœud, nous saisissons les valeurs *href* à l'aide de la fonction *gsub*.

4.4. Normalisation et filtrage des liens

Le processus de normalisation des URL (Lee, Kim et Hong, 2005) transforme l'URL en sa forme canonique afin d'éviter les URL en double à la frontière. En fait, la normalisation élimine les URL qui sont syntaxiquement différents mais pointent vers la même page en les transformant en URL syntaxiquement identiques. Un URL canonique comprend les cinq règles suivantes (avec des exemples):

1. Chaine d'URL en minisucule :

   De : Http://www.Test.Com À : http://www.test.com.

2. Résoudre le chemin URL :

   De : http://www.test.com/.././f.htm À : http://www.test.com/f.htm

3. Non-www URL à www

   De : http://test.com/sub/sub/ À : http://www.test.com/sub/sub/.

4. URL relatives à l'URL de base absolue

   De : /sub/file.html À : http://www.test.com/sub/file.html.

5. Elimination des ancres

   De : http://www.test.com/f.html#tite À : http://www.test.com/f.htm.

Presque toutes ces fonctions sont implémentées à l'aide d'expressions régulières via la correspondance et le remplacement de modèles. Après la normalisation des liens, les liens extraits doivent être filtrés. A ce stade particulier, deux niveaux de filtrage sont distingués :

Le premier affecte le processus de fouille, il est appliqué aux liens avant qu'ils ne soient ajoutés à la frontière, pour ignorer certains types de fichiers ou chemins

d'accès. Une forme antérieure d'identification d'un type de fichier est basée sur des extensions de fichier, et l'identification ultérieure est confirmée par des en-têtes de réponse de type de contenu, ce qui est plus fiable. De plus, comme RCrawler est un robot spécifique au domaine, les liens extraits doivent appartenir à la source du domaine et tous les liens externes doivent être supprimés.

Une autre méthode de filtration implique des pièges à araignées. En fait, certains scripts côté serveur, tels que CGI (interface de passerelle commune), peuvent générer un nombre infini de chemins logiques (pages et répertoires), résultant en un site Web infiniment profond. De tels pièges d'araignée peuvent être évités simplement en vérifiant la longueur de l'URL (ou le nombre de barres obliques dans l'URL). Enfin, avant d'ajouter un nouveau URL au pool de travail, il faut vérifier si l'URL existe déjà dans la frontière ou s'il a été récupéré. Le deuxième niveau de filtrage est basé sur les paramètres utilisateur et les liens de contrôle à indexer par le robot.

4.5. Détection des doublons et quasi-doublons

Il existe de nombreux cas dans lesquels une page Web peut être disponible sous plusieurs URL. La probabilité que des pages en double se produisent sur des sites Web spécifiques est plus élevée pour les sites Web dynamiques, tels que les blogs, les forums ou le commerce électronique, en raison des paramètres facultatifs de l'URL. Cela entraînera le robot d'indexation Web à stocker le même document plusieurs fois, ce qui aurait un impact négatif sur les résultats de la fouille de contenu.

Par conséquent, pour réduire le stockage redondant, les coûts de traitement et les exigences de bande passante réseau, un robot d'indexation Web peut effectuer un test de détection des doublons pour déterminer si un document a déjà été stocké ou non (Manku, Jain, Das et Sarma, 2007). Une méthode rapide et efficace pour y parvenir consiste à mapper l'URL de chaque page à un nombre représentatif compact, en utilisant une fonction de hachage qui a une faible probabilité de collisions, comme MD5 ou SHA-1; ou en utilisant l'algorithme d'empreinte digitale de Rabin (Rabin et al, 1981), qui est beaucoup plus rapide et offre de fortes garanties probabilistes.

Il faut conserver une structure de données appelée empreintes digitales, qui stocke les sommes de contrôle du document téléchargé afin que seules les empreintes digitales soient comparées, au lieu de comparer l'ensemble du document. Cependant, il existe certaines situations où plusieurs URL peuvent présenter visuellement le même contenu,

mais ces contenus sont différents structurellement d'une manière globale ou dans une petite partie, comme les liens, les balises de script ou les publicités.

Le problème est que du point de vue d'un ordinateur, les pages ne sont identiques que si elles se correspondent exactement octet par octet. Ces pages sont considérées comme des quasi-doublons. L'empreinte digitale de Charikar (Charikar, 2002), également connue sous le nom de SimHash, est une technique qui aide à détecter les quasi-doublons. Cela constitue une méthode de réduction de dimensionnalité, qui mappe des vecteurs de grande dimension à des empreintes digitales de petite taille. Pour les pages Web, cela s'applique comme suit :

1. Définir une taille d'empreinte digitale, par exemple, 64 bits.
2. La page Web est convertie en un ensemble de jetons (fonctionnalités). Chaque jeton se voit attribuer un poids. Dans notre implémentation, nous ne faisons que tokeniser le document, et nous considérons que tous les jetons ont le même poids de 1.
3. Chaque jeton est représenté par sa valeur de hachage à l'aide d'une fonction de hachage traditionnelle.
4. Un vecteur V (de longueur 64) d'entiers est initialisé à 0. Ensuite, pour chaque valeur de hachage de jeton (h), le vecteur V est mis à jour : le ième élément de V est diminué du poids du jeton correspondant si le ième bit de la valeur de hachage est 0. Sinon, l'élément est augmenté du poids du jeton correspondant.

Pour mesurer la similitude de deux empreintes digitales A et B, nous comptons le nombre de bits qui diffèrent entre les deux comme mesure de non-similarité. Si le résultat est égal à 0, les deux documents sont alors considérés comme identiques. Nous avons décidé d'implémenter trois options possibles dans RCrawler:

1. Échappez au processus de détection des doublons.
2. Détection des doublons à l'aide du hachage MD5.
3. Détection presque en double à l'aide de SimHash.

Pour le hachage MD5, nous avons utilisé le package de résumé. Cependant, nous n'avons pas pu implémenter l'algorithme SimHash dans R, en raison de la taille entière limitée (2*10 ¨9), ce qui bloque le calcul. Par conséquent, nous avons décidé de l'implémenter en Java, puis de l'envelopper dans notre package RCrawler pour le rendre disponible dans R. Pour interfacer R avec Java, nous avons utilisé la bibliothèque rJava (Simon, 2016), qui fournit toutes les fonctions nécessaires pour appeler une classe Java de l'intérieur d'une fonction R. En conséquence, nous avons créé deux fonctions :

- **getSimhash** qui prend la page Web comme une chaîne et une taille de bit de hachage en entrée, et renvoie son empreinte digitale.
- **Getdistance** prend deux empreintes digitales en entrée et renvoie la valeur de similitude.

4.6. Gestion des structures de données

En mémoire ;

RCrawler est un robot spécifique au domaine qui ne traverse que les pages Web d'un nom de domaine particulier. Par conséquent, la taille de la frontière est limitée et peut tenir en mémoire. De plus, R peut allouer jusqu'à $2^34 - 1$ octets (8 Go) de mémoire et peut prendre en charge des vecteurs de $2^31 - 1$ éléments (2 milliards) sur des plates-formes de 64 bits, ce qui est suffisant dans notre cas. En conséquence, la structure de données de frontière est implémentée en tant que vecteur de caractères en mémoire. Au fur et mesure de l'exploration, les URL à explorer ensuite sont sélectionnées séquentiellement par la fonction de gestionnaire de pool de travail, et les URL nouvellement découvertes sont envoyés à la fin de la frontière. Après avoir récupéré une page Web, elle est indexée dans l'index du robot, qui est représenté en mémoire sous la forme d'un bloc de données.

Dans le référentiel ;

Pour chaque site Web, RCrawler lance un dossier local qui contiendra tous les fichiers de ce site Web, qui sont les suivants:

- **Extractedcontent.csv** :contient le contenu extrait.
- **Index.csv** le fichier d'index du robot.
- **pages**: collection de pages Web téléchargées.
- **backup**: l'état du robot pour restaurer / reprendre une session d'exploration.

4.7. Implémentation de fonctionnalités supplémentaires

Analyseur robot.txt

Tout ce qui peut être exploré ne doit pas forcément subir à un processus d'exploration. Il existe des moyens plus adéquats de le faire. En fait, les webmasters souhaitent parfois garder au moins une partie de leur contenu interdit d'exploration, ou dans certains cas, ils peuvent vouloir bloquer certains robots pour économiser la bande passante. Le fichier *robot.txt* est utilisé à cet effet. Ce protocole d'exclusion de robot

indique au robot d'exploration quelles informations sur le site peuvent être récoltées et lesquelles ne le peuvent pas. Dans ce but, nous avons implémenté une fonction qui récupère et analyse le fichier robots.txt au moyen d'expressions régulières et nous avons identifié les règles d'accès correspondantes. Au démarrage de RCrawler, si le paramètre Obeyrobots est défini sur TRUE, le robot analysera fichier robots.txt du site web et retourner ses règles (répertoires ou fichiers autorisés et interdits) à filtrer.

Liens sortants/ Compte des liens internes

Le nombre de liens sortants d'une page Web représente le nombre de liens extraits et filtrés à partir de cette page. Cependant, le nombre de liens entrants (liens à partir d'autres pages qui pointent vers cette page) est calculé pendant le processus d'exploration. Pendant le processus d'exploration, le programme recherche chaque URL extraite dans l'index et incrémente sa valeur de lien correspondante.

Contrôle du niveau de profondeur

Il ne s'agit pas de la profondeur de fichier dans une structure de répertoires, mais plutôt de la distance entre le document racine et tous les liens extraits. Par exemple, le niveau 0 représente le document racine. Ainsi, après avoir récupéré tous les liens de ce document, le niveau est incrémenté de 1, etc. Nous avons implémenté cette fonctionnalité pour contrôler le niveau de profondeur de l'exploration.

Sauvegarde et restitution d'une session

En tant que fonction régulière sur R, le robot renvoie ses résultats à la fin du processus d'exploration (lorsque toutes les URL de la frontière ont été traitées). Ainsi, pour éviter la perte de données dans le cas où une fonction est interrompue au milieu de l'action, nous devons enregistrer l'état du robot et les structures de données sur le disque. Cependant, l'écriture de toutes les données sur le disque à chaque itération prend du temps, en raison de la recherche de disque et de la latence. Par conséquent, une solution pratique consiste à initier une connexion de fichier et à ajouter uniquement de nouvelles entrées au fichier existant, au lieu de remplacer le fichier entier à chaque itération. Une autre solution consiste à utiliser des variables R globales. En fait, au sein du mécanisme de calcul R, les environnements jouent un rôle crucial, car R les utilise systématiquement en arrière-plan pour un calcul interactif. La tâche d'un environnement est d'associer ou de lier un ensemble de noms à un ensemble de valeurs. Par défaut, lorsqu'une fonction est appelée, un environnement local est créé. Ainsi, toutes ses variables sont associées à cet environnement et ne peuvent être utilisées que pour cette fonction. Dans le cas où il

se bloque, toutes ses variables dépendantes sont perdues. Cependant, en utilisant l'environnement global (l'environnement de niveau supérieur disponible) pour gérer certaines structures de données critiques, telles que l'index du robot, la frontière les rendra plus fiables contre les pannes de fonctions. En outre, cela est encore loin de la sécurité offerte par le stockage persistant.

Extraction de contenu

L'analyse Web et l'extraction de données peuvent être implémentées en tant que deux tâches consécutives distinctes (le robot récupère toutes les pages Web dans un référentiel local, puis le processus d'extraction est appliqué à l'ensemble de la collection), ou en tant que tâches simultanées (lors de la récupération des URL, le raclage du contenu est appliqué à chaque page individuellement). Pour étendre les capacités de RCrawler en tant que grattoir Web, nous avons introduit l'extraction manuelle à l'aide de modèles XPath. Les utilisateurs peuvent spécifier un nombre illimité de modèles nommés à extraire pendant le processus d'analyse. De plus, les utilisateurs peuvent spécifier les pages à extraire (pages de détail), en filtrant les URL correspondant à un modèle d'expression régulier spécifique à l'aide du paramètre *urlregexfilter*.

L'annexe L, présente la documentation complète du Rcrawler, avec des démonstrations et des exemples de manipulation de ce dernier.

5. Conclusion

Dans cet ce chapitre, nous avons présenté RCrawler, un robot d'indexation et exploration Web basé sur R, multithread, flexible et puissant qui fournit une suite de fonctions utiles pour l'exploration Web, le grattage Web et également l'analyse de liens potentiels.

L'implémentation de RCrawler met également en évidence certaines faiblesses de l'environnement informatique parallèle R, que nous avons pu surmonter en adaptant notre architecture aux normes R et en utilisant des algorithmes supérieurs lorsque cela était nécessaire.

RCrawler est un projet en cours. Nous avons l'intention d'améliorer ses performances en utilisant des fonctions de bas niveau en C ++ ou Rcpp. Nous prévoyons également d'étendre ses caractéristiques et fonctionnalités en implémentant l'exploration ciblée, la gestion des mises à jour et la détection et l'extraction automatiques de contenu.

Chapitre 5 : Résultats et discussions

Nous réalisons nos expériences sur un PC avec 6 Go de RAM et un disque dur SSD de 100 Go, où le CPU est un Core i7 et le système d'exploitation est Microsoft Windows 8.1. Notre package est installé sur une plateforme R version 3.2.2. Les pages sont collectées sur le site Web du New York Times.

1. Tests préliminaires

Lors des tests préliminaires, nous avons remarqué que la vitesse du robot diminue considérablement après avoir exploré 1000 pages. Pour identifier la composante qui ralentit le robot, nous avons exécuté des tests sur chaque composant en récupérant, analysant, vérifiant et ajoutant individuellement des liens à la structure de données.

La figure 5.1 illustre le temps requis pour chaque composante RCrawler. On peut voir que les processus d'appariement et d'ajout prennent plus de temps à mesure que l'exploration se poursuit. En fait, vérifier si l'URL existe dans le tableau prend plus de temps à mesure que la taille de la frontière augmente. Pour vérifier si un URL existe dans la frontière, nous utilisons l'opérateur% in%. Malheureusement, cet opérateur n'est pas conçu pour gérer de grands ensembles de données. Par conséquent, afin de maintenir la vitesse du robot, nous avons remplacé l'opérateur% in% par l'opérateur %chin% du package data.table (Dowle, Srinivasan, 2015), qui est optimisé pour les grands ensembles de données et les vecteurs de caractères. De plus, nous avons optimisé les opérations de filtrage et d'ajout. La figure 5.2 montre la vitesse du robot avant et après l'optimisation. On peut voir que le nombre de pages Web téléchargées devient plus stable et linéaire. Il est à noter que les petites variations de vitesse restantes sont dues aux serveurs et aux connexions Internet qui peuvent devenir surchargés et entraîner ainsi des retards par rapport au temps de réponse.

2. Évaluation du multithreading

Après avoir téléchargé le package RCrawler, nous avons testé l'exécution du robot sous différentes configurations, en faisant varier le nombre de cœurs entre un et huit et le nombre de connexions (demandes simultanées) entre un et 16. Nous avons désactivé des fonctionnalités supplémentaires, telles que l'extraction et filtration. La figure 5.3 présente le nombre moyen de pages analysées par minute pour chaque

scénario. En observant les résultats, il est évident que la vitesse du robot augmente en augmentant le nombre de cœurs et de connexions.

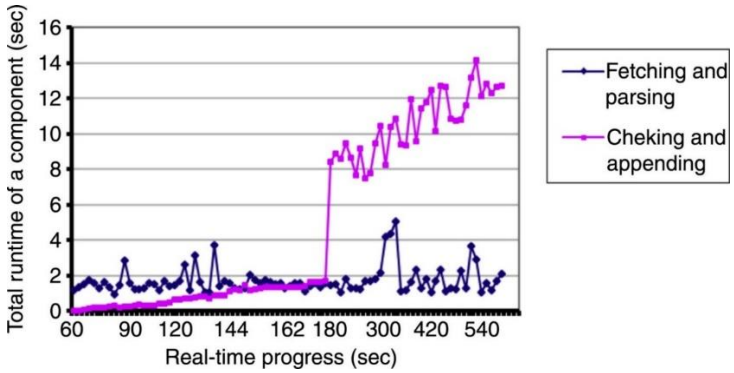


Figure 5.1 : Au fur et à mesure de l'exploration, les composants correspondants et ajoutés nécessitent plus de temps que les autres composants existants

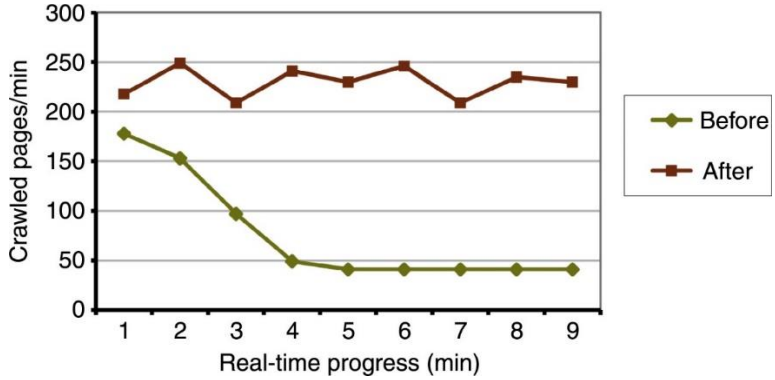


Figure 5.2 : Après avoir optimisé le composant correspondant, la vitesse du robot devient stable

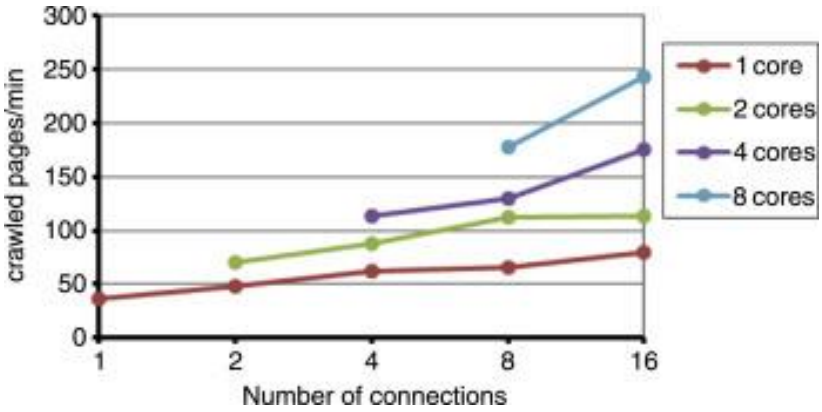


Figure 5.3: Le nombre de pages analysées / min augmente avec l'augmentation de processus et de connexions.

3. Évaluation de la performance en termes de vitesse

Pour faire une comparaison équitable et significative, nous avons sélectionné trois projets open source similaires pour différentes plateformes. Le tableau 7.1 présente une comparaison des robots d'exploration open source concernant le langage utilisé, le rôle essentiel et la prise en charge du parallèlisme.

Scrapy est un framework d'application Python open source pour l'écriture d'araignées Web qui explorent des sites Web. Ceci est généralement considéré comme le grattoir Web open source le plus rapide. Cependant, aucune recherche convaincante ne le prouve. Contrairement à d'autres robots d'exploration, *Scrapy* permet également l'extraction de données à partir de pages Web pendant le processus d'exploration. Il s'agit d'une caractéristique cruciale pour fournir des résultats significatifs dans notre évaluation.

Pour comparer les vitesses atteintes par les robots, nous tenterons de gratter les pages Web du site Web du New York Times et d'extraire les titres et les articles. Ensuite, nous évaluons et surveillons chaque robot individuellement. Il faut noter que RCrawler peut être directement comparé à Scrapy, car les deux sont de nature similaire et ont les mêmes caractéristiques. Par conséquent, les deux obtiennent la page de destination du site Web en entrée et ont les mêmes règles d'extraction de données et paramètres de configuration.

Cependant, *Rvest* ne prend pas en charge l'analyse ou le multithread. Pour faire une comparaison correcte, nous avons implémenté un script fait à la main de *Rvest*. Ici, nous fournissons une liste d'URL à l'avance, puis utilisons une fonction de boucle pour les parcourir. Nous appliquons des fonctions Rvest pour récupérer et analyser des pages Web et extraire les données souhaitées. Nous avons décidé de ne pas prendre en compte les autres packages R mentionnés dans la partie évaluation de ce chapitre, car ils étaient tous construits autour des mêmes boîtes à outils de base que *Rvest* et atteignent approximativement les mêmes performances.

Dans ce contexte, la figure 5.4 montre que RCrawler fonctionne mieux que les autres packages R en termes de vitesse. En fait, le multithreading permet une utilisation efficace des ressources disponibles. Cependant, notre robot d'exploration n'est toujours pas efficace en tant que grattoir Python, en raison des lacunes de R dans le traitement des données. Bien que la vitesse du robot d'indexation soit considérée comme un facteur

de performance, un robot d'indexation doit également être poli avec les serveurs Web, de sorte qu'il ne surcharge pas le serveur avec des demandes fréquentes dans un court laps de temps. Cela est particulièrement vrai pour les robots d'exploration de domaine (comme le nôtre). Par conséquent, il doit y avoir des délais suffisants entre les requêtes consécutives adressées au même serveur.

Tableau 5.1: Évaluation des librairies et frameworks open source pour le grattage du web

Nom du projet	Type	Role principal	Language	Parallélisme
RCrawler	Librairie	Indexation et grattage	R	Oui
Rvest	Librairie	Récupération et analyse	R	Non
Scrapy	Framework	Indexation et grattage	Python	Oui

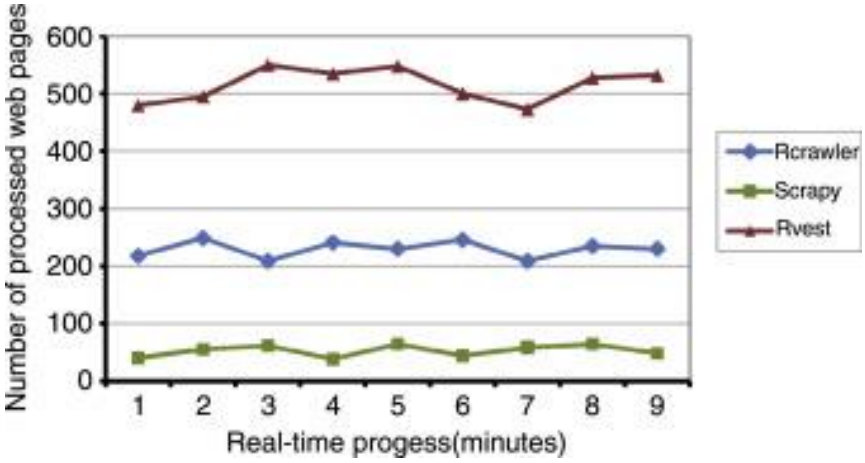


Figure 5.4: RCrawler atteint des performances élevées par rapport à Rvest, mais pas aussi rapidement que Scrapy.

4. Évaluation du grattage web

Nous avons collecté des données HTML afin d'analyser l'intégralité. Nous avons utilisé la technique d'extraction XPath, car c'est la seule technique prise en charge dans RCrawler jusqu'à présent. De plus, nous avons utilisé huit cœurs et huit connexions. Comme notre implémentation de scraper nécessite un URL pour fonctionner, nous avons utilisé un serveur Web de domaine local avec un nombre prédéfini de pages Web (1000 pages). Parmi ceux-ci, il y a 800 pages de contenu et 200 pages de menu. Chaque page Web de contenu contient deux éléments à extraire : le titre et la publication. Par

conséquent, nous savons à l'avance que la quantité de données à extraire s'élève à 1600 enregistrements.

Les statistiques résultantes pour l'expérience sont présentées dans le tableau 7.2. On peut voir que la quantité totale de données extraites obtenues par les grattoirs comparés se situe entre 99% et 100%. En fait, XPath est considéré comme l'une des techniques d'extraction les plus fiables, car il est basé directement sur une arborescence XML pour localiser les nœuds d'élément, les nœuds d'attribut, les nœuds de texte et tout ce qui peut se produire dans un document XML sans apprentissage automatique ou intelligence artificielle.

Ainsi, dans notre cas, la fonction de grattage n'a pas d'effet direct sur les performances du robot. Cependant, la qualité de couverture d'un robot joue un rôle très important dans ses performances. Cela décrit la possibilité d'extraire tous les liens accessibles d'une page Web analysée. À ce stade, Rvest ne peut pas être comparé, car il n'a pas la possibilité de suivre les URL découverts. En effet, nous pensons que RCrawler est la meilleure solution R pour l'exploration et le grattage de données dans un environnement R. Il peut explorer, télécharger des pages et même gratter du contenu avec une vitesse et une efficacité élevée. Outre sa facilité d'utilisation, RCrawler fournit également une solution complète pour gérer les tâches de collecte de données dans R, avec de nombreux paramètres et options personnalisables, et sans avoir besoin d'utiliser un robot / scraper Web externe.

Tableau 5.2 : Performance des projets en termes de grattage

Nom du projet	Pourcentage de la performance de grattage web
RCrawler	99.8%
RVest	99.8%
Scrapy	99.8%

Conclusion générale

Tout le monde s'accorde à dire qu'aucune innovation n'a pu changer la société aussi radicalement que le World Wide Web. En effet, de l'information et de la communication au travail et à l'éducation en passant par les comportements d'achat et de rencontres, le World Wild Web n'a épargné aucun domaine de la vie. Cela fournit un environnement riche pour la fouille de données et la découverte des connaissances. Assurément, la fouille du Web, qui se veut une branche scientifique ayant pour but de découvrir des connaissances par le biais d'informations disponibles sur le Web.

Il convient de rappeler que le processus de fouille de texte inclut communément la récupération d'informations et les applications de méthodes statistiques avancées et de traitement du langage naturel (NLP), en s'appuyant sur une panoplie de techniques d'acquisition et de traitement des données. Ces techniques reposent essentiellement sur des outils pour le web scraping et le web crawling. La fouille représente un secteur intéressant, regroupant les dernières avancées dans les méthodes de recherche, et des outils logiciels. Son utilisation ne se limite pas au monde universitaire, professionnel, mais s'étend aux organismes gouvernementaux., aux marchés boursiers, et aux protections politiques. Mais ce n'est pas tout : la fouille de texte est sollicitée dans un tas d'autres domaines, en particulier dans le marketing et le commerce.

Le design de recherche constitue l'un des piliers les plus importants dans les sciences sociales. En effet, il représente l'architecture de base permettant de maximiser les chances d'atteindre les objectifs de recherche. Pour ce faire, les décisions au début tout projet de recherche doivent êtres mûrement réfléchis, pour éviter toute erreur susceptible de conduire à des résultats non significatifs et non fiables. Les designs de recherche varient en fonction de la méthode utilisée (qualitative ou mixte), de la collection des documents (unique ou multiple), ainsi que des méthodes d'échantillonnage de données (inductives, déductives ou abductives).

Avec l'augmentation des performances des médias électroniques, la mise en réseau croissante et la croissance explosive des possibilités de stockage des données électroniques, il y a eu une augmentation des informations disponibles. Cependant, ces données disponibles électroniquement et la quantité de données ne sont plus comparables à la collecte de données sur papier. Dans ce contexte, on parle de mégadonnées : le nombre de données (principalement des millions d'enregistrements de données), la vitesse de l'enquête (en temps réel) et la gamme d'instruments d'enquête

(caméras, satellites, Internet, registres de scanner, ...) sont importants à tous égards. Ces énormes quantités de données imposent des exigences particulières à l'évaluation.

Le terme exploration de données se trouve dans le domaine des mégadonnées. Les méthodes exploratoires peuvent être incluses dans l'exploration de données, dans laquelle - en partie entièrement automatisées et en partie uniquement semi-automatisées - des informations sont obtenues à partir de grandes quantités de données. L'objectif de l'exploration de données est de promouvoir les dépendances, les régularités et les modèles de données brutes autrement non liées ou non structurées.

Lorsqu'elles utilisent les diverses techniques d'analyse et d'évaluation de l'exploration de données de manière réfléchie, ces méthodes offrent des informations précieuses et des avantages concurrentiels. Tous ces processus impliquent des défis particuliers. L'un des problèmes les plus importants de l'exploration de données est que chaque méthodologie doit d'abord être définie manuellement. Les humains sont responsables de la détermination des variables dépendantes et indépendantes, des classes et des techniques analytiques utilisées. Les résultats de l'exploration de données sont fondamentalement falsifiés par certaines hypothèses, idées et objectifs.

- Les changements incessants dans le domaine du web ont donné naissance à une panoplie de contraintes, nécessitant une maintenance régulière. Nous pouvons en citer comme exemples :
- Les réglementations légales qui régissent les relations entre les fournisseurs d'informations et les utilisateurs ;
- Les réglementations de sécurité, comprenant généralement (les précautions d'usage, les clauses de confidentialité, les mentions légales…);
- Les mises à jours continues des technologies web qui concernent plus particulièrement les protocoles de transmission des données et les langages de représentation des données, tel que le HTTP, le W3C , l'HTML , CSS, et le JAVASCRIPT ..etc.

Il s'ensuit donc que la recherche constante est la clé de voûte pour le bon fonctionnement des systèmes de collecte et d'extraction des données. Cependant, pour accompagner les changements immuables consubstantiels à l'environnement du web, l'idéal serait d'établir des normes et des standards permettant la création des canaux dédiés à l'accès aux données, tel que les web services ou les API REST. Ainsi, les chercheurs pourront récupérer les données de manière structurée, efficace et légitime.

Références

Andersen, D. (2015). Stories of change in drug treatment: A narrative analysis of "whats" and "hows" in institutional storytelling. Sociology of Health & Illness, 37(5), 668–682.

Asher, K., & Ojeda, D. (2009). Producing nature and making the state: Ordenamiento territorial in the Pacific lowlands of Colombia. Geoforum, 40(3), 292–302.

Bamberg, M. (2004). Form and functions of "slut bashing" in male identity constructions in 15-year-olds. Human Development, 47(6), 331–353.

Banerjee, S., & Pedersen, T. (2002). An adapted Lesk algorithm for word sense disambiguation using WordNet. Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, Mexico.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O. (2007, January). Open information extraction from the web. Communications of the ACM—Surviving the Data Deluge, 51(12), 68–74.

Bauer M. W., Gaskell, G., & Allum, N. (2000). Quantity, quality and knowledge interests: Avoiding confusions. In M. W. Bauer & G. Gaskell (Eds.), Qualitative researching with text, image and sound (pp. 3–17). Thousand Oaks, CA: Sage.

Bauer, M. W., Bicquelet, A., & Suerdem, A. K. (Eds.). (2014). Text analysis: An introductory manifesto. In M. W. Bauer, A. Bicquelet, & A. K. Suerdem (Eds.), Textual analysis (SAGE benchmarks in social research methods) (Vol. 1). Thousand Oaks, CA: Sage.

Berelson, B. (1952). Content analysis in communication research. Glencoe, IL: Free Press.

Berglund, E. (2001). Facts, beliefs and biases: Perspectives on forest conservation in Finland. Journal of Environmental Planning and Management, 44, 833–849.

Bickes, H., Otten, T., & Weymann, L. C. (2014). The financial crisis in the German and English press: Metaphorical structures in the media coverage on Greece, Spain and Italy. Discourse & Society, 25(4), 424-445.

Boldi P, Codenotti B, Santini M, Vigna S. Ubicrawler: A scalable fully distributed web crawler. Softw. - Pract. Exp. 2004;34(8):711–26.

Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. Cognition, 75(1), 1–28.

Bourdieu, P., & Thompson, J. B. (1991). Language and symbolic power. Cambridge, MA: Harvard University Press.

Boyatzis, R. E. (1998). Transforming qualitative information: Thematic analysis and code development. Thousand Oaks, CA: Sage.

Bradley, J. (1989). TACT user manual. Toronto, Canada: University of Toronto Press.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101.

Buchholz, M. B., & von Kleist, C. (1995). Psychotherapeutische Interaktion—Qualitative Studien zu Konversation und Metapher, Geste und Plan. Opladen:Westdeutscher Verlag.

Busanich, R., McGannon, K., & Schinke, R. (2014). Comparing elite male and female distance runners' experiences of disordered eating through narrative analysis. Psychology of Sport and Exercise, 15(6), 705–712.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E. R., Jr., & Mitchell, T. M. (2010, July). Toward an architecture for never-ending language learning. Proceedings of the Twenty-Fourth American Association for Artificial Intelligence Conference on Artificial Intelligence

Charteris-Black, J. (2009). Metaphor and political communication. In A. Musolff & J. Zinken (Eds.), Metaphor and discourse (pp. 97–115). Basingstoke, England: Palgrave Macmillan.

Charteris-Black, J. (2012). Comparative keyword analysis and leadership communication: Tony Blair—A study of rhetorical style. In L. Helms (Ed.), Comparative political leadership (pp. 142–164). Basingstoke, England: Palgrave Macmillan.

Charteris-Black, J. (2013). Analysing political speeches: Rhetoric, discourse and metaphor. Basingstoke, England: Palgrave Macmillan.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. Computational Linguistics, 16(1), 22–29.

Collins, M. (2002). Ranking algorithms for named-entity extraction: Boosting and the voted perceptron. Proceedings of the 40th Annual Meeting on Association for

Computational Linguistics (pp. 489–496). Stroudsburg, PA: Association for Computational Linguistics.

Couper, M. P. (2000). Web surveys: A review of issues and approaches. Public Opinion Quarterly, 64(4), 464–494.

Creswell, J. D. (2014). Research design: Qualitative, quantitative, and mixed methods approaches. Thousand Oaks, CA: Sage.

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., & Kleinberg, J. (2012, April 16–20). Echoes of power: Language effects and power differences in social interaction. WWW 2012. Retrieved March 29, 2016, from http://www.cs.cornell.edu/~cristian/Echoes_of_power_files/echoes_of_power.

Denzin, N. K, & Lincoln, Y. S. (2011). Epilogue: Toward a "refunctioned ethnography." The SAGE Handbook of Qualitative Research (pp. 715–718). Thousand Oaks, CA: Sage.

Eddelbuettel D. Cran task view: High-performance and parallel computing with r; 2016.

Esuli, A., & Sebastiani, F. (2006a). Determining term subjectivity and term orientation for opinion mining. Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy.

Esuli, A., & Sebastiani, F. (2006b). SentiWordNet: A publicly available lexical resource for opinion mining. Proceedings of the Fifth Conference on Language Resources and Evaluation, Genova, Italy.

Etzioni, O., Cafaraella, M., Downey, D., Kok, S., Popescu, A. M., Shaked, T., . .. Yates, A. (2004). Web-scale information extraction in KnowItAll: (Preliminary results). Proceedings of the Thirteenth International Conference on World Wide Web (pp. 100–110). New York, NY: Association for Computing Machinery.

Fader, A., Soderland, S., & Etzioni, O. (2011). Identifying relations for open information extraction. Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1535–1545). Stroudsburg, PA: Association for Computational Linguistics.

Fairclough, N. (1995). Critical discourse analysis: The critical study of language. London, England: Longman.

Fairclough, N. (1995). Critical discourse analysis: The critical study of language.London, England: Longman.

Feldman, J. (2006). From molecule to metaphor. Cambridge, MA: MIT Press.

Fellbaum, C. (Ed.). (1998). WordNet: An electronic lexical database. Cambridge, MA: MIT Press.

Fors, A., Dudas, K., & Ekman, I. (2014). Life is lived forwards and understood backwards—Experiences of being affected by acute coronary syndrome: A narrative analysis. International Journal of Nursing Studies, 51(3), 430–437.

Franzosi, R. (1987). The press as a source of socio-historical data: Issues in the methodology of data collection from newspapers. Historical Methods, 20(1), 5–16.

Franzosi, R., De Fazio, G., & Vicari, S. (2012). Ways of measuring agency: An application of quantitative narrative analysis to lynchings in Georgia (1875–1930). Sociological Methodology, 42(1), 1-42.

Gabrilovich, E., & Markovitch, S. (2007, January). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In IJcAI (Vol. 7, pp. 1606-1611).

Garton, L., Haythornthwaite, C., & Wellman, B. (1997). Studying online social networks. Journal of Computer Mediated Communication, 3(1) http://onlinelibrary.wiley.com/doi/10.1111/j.1083- 6101.1997.tb00062.x/abstract.

Gatti, L., & Catalano, T. (2015). The business of learning to teach: A critical metaphor analysis of one teacher's journey. Teaching and Teacher Education,45, 149–160.

Gee, J. P. (1991). A linguistic approach to narrative. Journal of Narrative and Life History, 1(1), 15–39.

Gibson, C. B., & Zellmer-Bruhn, M. E. (2001). Metaphors and meaning: An intercultural analysis of the concept of teamwork. Administrative Science Quarterly, 46(2), 274–303.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. Cambridge, MA: MIT Press.

Gorard, S. (2013). Research design: Creating robust approaches for the social sciences. Thousand Oaks, CA: Sage.

Grossetti, M. (2006). Trois échelles d'action et d'analyse : L'abstraction comme opérateur d'échelle. L'Année sociologique, vol. 56(2), 285-307. doi:10.3917/anso.062.0285.

Hardie, A., Koller, V., Rayson, P., & Semino, E. (2007). Exploiting a semantic annotation tool for metaphor analysis. In Corpus Linguistics Conference (CL2007).

Hardy, C. (2001). Researching organizational discourse. International Studies of Management & Organization, 31(3), 25–47.

Heritage, J., & Raymond, G. (2005). The terms of agreement: Indexing epistemic authority and subordination in talk-in-interaction. Social Psychology Quarterly, 68(1), 15–38.

Herrera, Y. M., & Braumoeller, B. F. (2004, Spring). Symposium: Discourse and content analysis. Qualitative Methods Newsletter, 15–19. Retrieved from http://www.braumoeller.info/wp-content/uploads/2012/12/Discourse-Content-Analysis.pdf

Hewson, C., & Laurent, D. (2012). Research design and tools for Internet research. In J. Hughes (Ed.), SAGE Internet research methods: Volume 1. Thousand Oaks, CA: Sage.

Heydon, A., & Najork, M. (1999). Mercator: A scalable, extensible web crawler. World Wide Web, 2(4), 219-229.

Hine, C. (2000). Virtual ethnography. Thousand Oaks, CA: Sage.

Hofstede, G. (1980). Culture's consequences: International differences in workrelated values. Beverly Hills, CA: Sage.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 168–177). New York, NY: Association for Computing Machinery.

Ignatow, G. (2003). "Idea hjamsters" on the "bleeding edge": Profane metaphors in high technology argon. Poetics, 31(1), 1–22.

Ignatow, G. (2009). Culture and embodied cognition: Moral discourses in Internet support groups for overeaters. Social Forces, 88(2), 643–669.

Johnson-Laird, P. N. (1983). Mental models: Toward a cognitive science of language, inference, and consciousness. Cambridge, MA: Harvard University Press.

Kassarjian, H. (1977). Content analysis in consumer research. Journal of Consumer Research, 4(1), 8–18.

Kim, S.-M., & Hovy, E. (2006). Identifying and analyzing judgment opinions. Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics.

Koppel, M., Argamon, S., & Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. Literary and Linguistic Computing, 17(4), 401–412.

Kozinets, R. V. (2002). The field behind the screen: Using netnography for marketing research in online communities. Journal of Marketing Research, 39(1), 61–72.

Kozinets, R. V. (2009). Netnography: Doing ethnographic research online.

Kozinets, R. V. (2009). Netnography: Doing ethnographic research online. Thousand Oaks, CA: Sage.

Krippendorff, K. (2013). Content analysis: An introduction to its methodology. Thousand Oaks, CA: Sage.

Kuckartz, U. (2014). Qualitative text analysis: A guide to methods, practice, and using software. Thousand Oaks, CA: Sage.

Labov, W. (1972). Sociolinguistic patterns. Philadelphia: University of Pennsylvania Press.

Laird, E. A., McCance, T., McCormack, B., & Gribben, B. (2015). Patients' experiences of in-hospital care when nursing staff were engaged in a practice development programme to promote person-centredness: A narrative analysis study. International Journal of Nursing Studies, 52(9), 1454–1462.

Lakoff, G. (1987). Women, fire, and dangerous things. What categories reveal about the mind. Chicago, IL: University of Chicago Press.

Landauer, T. K. (2002). On the computational basis of learning and cognition: Arguments from LSA. Psychology of Learning and Motivation, 41, 43–84.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse Processes, 25(2–3), 259–284.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. Proceedings of the SIGDOC Conference 1986 (pp. 24–26). New York, NY: ACM.

Levina, N., & Arriaga, M. (2012). Distinction and status production on user- generated content platforms: Using Bourdieu's theory of cultural production to understand social dynamics in online fields. Information Systems Research, 25(3), 468–488.

Liu, B., & Mihalcea, R. (2007). Of men, women, and computers: Data-driven gender modeling for improved user interfaces. Paper presented at the Proceedings of the International Conference on Weblogs and Social Media, Boulder, CO.

Magnini, B., & Cavaglia, G. (2000). Integrating subject field codes into WordNet. Proceedings of the Conference on Language Resources and Evaluations (LREC-2000) (pp. 1413–1418). Athens, Greece.

Maguire, S., Hardy, C., & Lawrence, T. (2004). Institutional entrepreneurship in emerging fields: HIV/AIDS treatment advocacy in Canada. Academy of Management Journal, 47(5), 657–679.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics, 19(2), 313–330.

McCallum, A., & Nigam, K. (1998). A comparison of event models for Naive Bayes text classification. Paper presented at the AAAI-98 Workshop on Learning for Text Categorization.

Merkl-Davies, D. M., & Koller, V. (2012). "Metaphoring" people out of this world: A critical discourse analysis of a chairman's statement of a UK defence firm. Accounting Forum, 36(3), 178–193.

Mihalcea, R., Banea, C., & Wiebe, J. (2007). Learning multilingual subjective language via cross-lingual projections. Paper presented at the Proceedings of the Association for Computational Linguistics, Prague, Czech Republic.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems (pp. 3111–3119).

Mische, A. (2014). Measuring futures in action: Projective grammars in the Rio+20 debates. Theory & Society, 43(3–4), 437–464. New York, NY: Vintage Books. Political Communication, 27(2), 121–140.

Mukherjee, A., & Liu, B. (2012). Aspect extraction through semi-supervised modeling. Proceedings of the 50th Annual Meeting of the Association for Computational

Linguistics. (pp. 339–348).Stroudsburg, PA: Association for Computational Linguistics.

Munzert, S., Rubba, C., Meißner, P., & Nyhuis, D. (2014). Automated data collection with R: A practical guide to web scraping and text mining. John Wiley & Sons.

Nakagawa, T., Inui, K., & Kurohashi, S. (2010). Dependency tree-based sentiment classification using CRFs with hidden variables. In Human Language Technologies: The 2010 Stroudsburg, PA: Association for Computational Linguistics. Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 786–794).

Navigli, R., & Ponzetto, S. (2012). Artificial Intelligence, 193, 217–250.

O'Keefe, A., & Walsh, S. (2012). Applying corpus linguistics and conversation analysis in the investigation of small group teaching in higher education. Corpus Linguistics and Linguistic Theory, 8(1), 159–181.

O'Mara-Shimek, M., Guillén-Parra, M., & Ortega-Larrea, A. (2015). Stop the bleeding or weather the storm? Crisis solution marketing and the ideological use of metaphor in online financial reporting of the stock market crash of 2008 at the New York Stock Exchange. Discourse & Communication, 9(1), 103-123.

Olston, C., & Najork, M. (2010). Web crawling. Now Publishers Inc.

Ortony, A., Clore, G. L., & Collins, A. (1990). The cognitive structure of emotions. New York, NY: Cambridge University Press.

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. Proceedings of the Forty-Second Annual 271–278).Stroudsburg, PA: Association for Computational Linguistics. Meeting on Association for Computational Linguistics.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1–35.

Patton, M. Q. (2014). Qualitative research & evaluation methods: Integrating theory and practice (4th ed.). Thousand Oaks, CA: Sage.

Pennebaker, J. W., & King, L. (1999). Linguistic styles: Language use as an individual difference. Journal of Personality and Social Psychology, 77, 1296–1312.

Phillips, N., & Hardy, C. (2002). Discourse analysis: Investigating processes of social construction. Thousand Oaks, CA: Sage.

Propp, V. (1968). Morphology of the folktale. Austin: University of Texas Press.

Ravitch, S. M., & Riggan, J. M. (2016). Reason & rigor: How conceptual frameworks guide research. Thousand Oaks, CA: Sage.

Resnik, P., Garron, A., & Resnik, R. (2013). Using topic modeling to improve prediction of neuroticism and depression in college students. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (pp. 1348–1353). Stroudsburg, PA: Association for Computational Linguistics.

Rich C, Steve W, Dan T. do Parallel Foreach Parallel Adaptor for the 'parallel' Package; 2015. http://CRAN.R-project.org/package=doParallel. [Accessed 30 February 2020].

Ricoeur, P. (1991). Narrative identity. Philosophy Today, 35(1),73–81.

Roberts, C. W. (1997). Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts. Mahwah, NJ: Lawrence Erlbaum.

Roberts, C. W., Popping, R., & Pan, Y. (2009). Modalities of democratic transformation forms of public discourse in Hungary's latest newspaper, 1990–1997. International Sociology, 24(4), 498–525.

Roberts, C. W., Zuell, C., Landmann, J., & Wang, Y. (2010). Modality analysis: A semantic grammar for imputations of intentionality in texts. Quality & Quanitity, 44(2), 239–257.

Roget, P. (1987). Roget's thesaurus of English words and phrases. New York, NY: Longman. (Original work published 1911)

Rosenwald, G. C., & Ochberg, R. L. (1992). Storied lives: The cultural politics of self-understanding. New Haven, CT: Yale University Press.

Ruiz Ruiz, J. (2009). Sociological discourse analysis: Methods and logic. Forum: Qualitative Social Research, 10(2). Retrieved June 27, 2015, from qualitative-research.net/index.php/fqs/article/view/1298/2882

Salmons, J. (2014). Qualitative online interviews. Thousand Oaks, CA: Sage.

Santa Ana, O. (2002). Brown tide rising metaphors of Latinos in contemporary American public discourse. Austin: University of Texas Press.

Sapir, J., & Crocker, J. (Eds.). (1977). The social use of metaphor: Essays on the anthropology of rhetoric. Philadelphia: University of Pennsylvania Press.

Schmitt, R. (2005). Systematic metaphor analysis as a method of qualitative research. The Qualitative Report, 10(2), 358–394.

Schwandt, T. A. (2001). Dictionary of qualitative research. Thousand Oaks, CA: Sage.

Silverman, D. (1993). Interpreting qualitative data: Methods for analyzing talk, text and interaction. Newbury Park, CA: Sage.

Silverman, D. (Ed.). (2016). Qualitative research. Thousand Oaks, CA: Sage.

Silverman, D. (Ed.). (2016). Qualitative research. Thousand Oaks, CA: Sage. Snow C. P. (2013). The two cultures and the scientific revolution. London, England: Martino Fine Books. (Original work published 1959).

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Stone, P. J., & Hunt, E. B. (1963). A computer approach to content analysis: Studies using the General Inquirer system. AFIPS ’63 (Spring) Proceedings of the May 21–23, 1963, Spring Joint Computer Conference (pp. 241–256). doi:10.1145/1461551.1461583

Stone, P. J., Dunphry, D., Smith, M. S., & Ogilvie, D. M. (1966). The General Inquirer: A computer approach to content analysis. Cambridge, MA: MIT Press.

Strapparava, C., & Mihalcea, R. (2007). SemEval-2007 task 14: Affective text. Proceedings of the Fourth International Workshop on the Semantic Evaluations, Prague, Czech Republic (pp. 70–74). Stroudsburg, PA: Association for Computational Linguistics.

Strapparava, C., & Valitutti, A. (2004). WordNet-Affect: An affective extension of WordNet. Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal.

Stroet, K., Opdenakker, M.-C., & Minnaert, A. (2015). Need supportive teaching in practice: A narrative analysis in schools with contrasting educational approaches. Social Psychology of Education, 18(3), 585–613.

Tashakkori, A. M., & Teddlie, C. B. (2010). SAGE handbook of mixed methods in social & behavioral research (2nd ed.). Thousand Oaks, CA: Sage.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology, 29(1), 24–54.

Teddlie, C. B., & Tashakkori, A. M. (Eds.). (2008). Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences. Thousand Oaks, CA: Sage.

Thomas L, Scott C, Patrick M, Karthik R, Christopher G, CRAN Task View: Web Technologies and Services; 2016. https://cran.r-project.org/web/views/WebTechnologies.html. [Accessed 29 February 2020].

Trappey, C., Wu, H., Liu, K., & Lin, F. (2013, September 11–13). Knowledge discovery of service satisfaction based on text analysis of critical incident dialogues and clustering methods. 2013 IEEE 10th International Conference on e-Business Engineering (pp. 265–270). Coventry, United Kingdom: ICEBE 2013.

Turney, P. D. (2002). Th/umbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of the Fortieth Annual Meeting on Association for Computational Linguistics (pp. 417–424). Stroudsburg, PA: Association for Computational Linguistics.

Turney, P., Neuman, Y., Assaf, D., & Cohen, Y. (2011, July). Literal and metaphorical sense identification through concrete and abstract context. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (pp. 680-690).

Walejko, G. (2009). Online survey: Instant publication, instant mistake, all of the above. In E. Hargittai (Ed.), Research confidential: Solutions to problems most social scientists pretend they never have (pp. 101–115). Ann Arbor: University of Michigan Press.

Wertsch, J. V. (1985). Vygotsky and the social formation of mind. Cambridge, MA: Harvard University Press.

White, H. (1978). Tropics of discourse: Essays in cultural criticism. Baltimore, MD: Johns Hopkins University Press.

Wiebe, J., Bruce, R., & O'Hara, T. (1999). Development and use of a gold- standard data set for subjectivity classifications. Proceedings of the Thirty- Seventh Annual Meeting of the Association for Computational Linguistics (pp. 246–253). Stroudsburg, PA: Association for Computational Linguistics.

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, 39(2–3), 165–210.

Wilson, T. (2008). Fine-grained subjectivity and sentiment analysis: Recognizing the intensity, polarity, and attitudes of private states (PhD thesis, University of Pittsburgh).

Windelband, W. (1998). On history and natural science. History and Theory, 19, 165–185. (Original work published 1894)

Windelband, W. (2001). A history of philosophy. Cresskill, NJ: The Paper Tiger. (Original work published 1901)

Winkel, G. (2012). Foucault in the forests—A review of the use of "Foucauldian" concepts in forest policy analysis. Forest Policy and Economics, 16, 81–92.

Yarowsky, D. (2000). Hierarchical decision lists for word sense disambiguation. Computers and the Humanities, 34(1), 179–186.

Zhao Y. R and Data Mining: Examples and Case Studies. Academic Press; 2012.