

Université de Montréal

Visualisation de données dans le domaine de l'E-recrutement

par
Abdessamad Outerqiss

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté à la Faculté des arts et des sciences
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en informatique

Décembre, 2016

© Abdessamad Outerqiss, 2016.

Université de Montréal
Faculté des arts et des sciences

Ce mémoire intitulé:

Visualisation de données dans le domaine de l'E-recrutement

présenté par:

Abdessamad Outerqiss

a été évalué par un jury composé des personnes suivantes:

Miklós Csűrös,	président-rapporteur
Philippe Langlais,	directeur de recherche
Guy Lapalme,	membre du jury

Mémoire accepté le: 7 mars 2017

RÉSUMÉ

La récente profusion des données, communément appelée Big Data, nécessite une analyse pertinente de ces larges volumes de données afin d'en tirer l'information utile nécessaire à la prise de décision. La visualisation de données se révèle à cet égard un moyen efficace pour transmettre cette information de façon interactive et synthétique. Le travail décrit dans ce mémoire qui constitue un volet du projet BPP, collaboration entre le RALI et la société LittleBigJob (LBJ), vise à répondre à ce besoin par l'implémentation d'un tableau de bord permettant la visualisation des offres d'emploi sur le web. Ces offres étant composées de plusieurs sections : titre de l'offre, compagnie qui recrute, description de l'offre, etc. Certaines informations contenues dans la description de l'offre ne peuvent pas être extraites directement. Ainsi, pour l'extraction des compétences citées dans une offre, nous utilisons les techniques de l'apprentissage automatique et plus particulièrement les champs markoviens conditionnels (CRF) utilisés pour l'étiquetage des séquences. Les expériences menées visent également à tester la capacité de ces modèles à trouver ces compétences dans la description de l'offre avec un ensemble d'entraînement partiellement étiqueté, d'une part parce que nous ne disposons pas d'une liste complète de compétences nécessaire à l'étiquetage, et d'autre part, parce que de nouvelles compétences apparaissent continuellement.

Mots clés: E-recrutement, visualisation de données, extraction d'information, apprentissage automatique.

ABSTRACT

The large amount of data available nowadays, so-called Big Data, requires a relevant analysis to derive information and get insights for decision making. Data visualization is an effective way to convey this information interactively and synthetically. This work, which is part of BPP Project, a collaboration between the RALI and LBJ, aims to meet this need by implementing a dashboard for visualization of job offers on the web. These offers consist of several sections: title, company, description, etc. Some information contained in the description cannot be extracted directly. Thus, for the extraction of skills from the description of an offer, we use machine learning techniques, especially Conditional Random Fields (CRF) used for sequence labeling. We also tested the ability of those models to find skills in the description of the offer with partial labeled training dataset, as we do not have a complete list of skills required for labeling, and also because new skills appear constantly.

Keywords: E-recruitment, Data Visualization, Information Extraction, Machine Learning.

TABLE DES MATIÈRES

RÉSUMÉ	iii
ABSTRACT	iv
TABLE DES MATIÈRES	v
LISTE DES TABLEAUX	vii
LISTE DES FIGURES	viii
LISTE DES ANNEXES	x
LISTE DES SIGLES	xi
REMERCIEMENTS	xii
CHAPITRE 1 : INTRODUCTION	1
1.1 Préambule	1
1.2 Contexte du projet	2
1.2.1 L'appariement profils/offres	3
1.2.2 La génération automatique d'ontologie	4
1.2.3 La génération des lettres de motivation	4
1.2.4 Baromètre de l'emploi	5
1.3 Structure du mémoire	5
CHAPITRE 2 : ÉTAT DE L'ART	6
2.1 Évolution de la visualisation d'information	7
2.2 Types de données	10

2.3	Taxonomies de visualisation d'information	12
2.4	Visualisation de l'information dans le web	16
CHAPITRE 3 : BAROMÈTRE DE L'EMPLOI		20
3.1	Préparation des données	20
3.2	Choix du type des graphiques	26
CHAPITRE 4 : EXTRACTION DES COMPÉTENCES		31
4.1	Données	31
4.1.1	Description des données	31
4.1.2	Prétraitement des données	36
4.1.3	Identification de la section des compétences	37
4.2	Approche basée sur l'apprentissage automatique	41
4.2.1	Motivations du choix des CRF	41
4.2.2	Protocole expérimental	43
4.2.3	Résultats	50
CHAPITRE 5 : CONCLUSION		55
BIBLIOGRAPHIE		57

LISTE DES TABLEAUX

2.I	Problèmes à résoudre par type de données	14
4.I	Compétences annotées	35
4.II	Distribution des compétences dans une offre	39
4.III	Étapes de prétraitement des données	41
4.IV	Nombre de mots et de compétences	44
4.V	Matrice de confusion	45
4.VI	Performances par trait	50
4.VII	Performances des combinaisons des traits	51
4.VIII	Performances de l'annotation manuelle	52

LISTE DES FIGURES

2.1	Carte de la campagne de Napoléon en Russie (1812-1813) par C.J. Minard	8
2.2	Carte des décès pendant l'épidémie de Londres de 1854	9
2.3	Coordonnées parallèles appliquées au jeu de données Iris à 4 dimensions	12
2.4	Classification des techniques de visualisation selon Keim	15
3.1	Visualisation interactive des offres d'emploi	21
3.2	Exemple de coordonnées de la province de Saskatchewan au format GeoJSON	22
3.3	Exemples de formulations des salaires dans les offres d'emploi	23
3.4	Utilisation de Crossfilter et dc.js pour la conception du graphique <i>Company</i>	25
3.5	Contrôle du graphique <i>Company</i>	26
3.6	Matrice du choix du type du graphique (Zelazny)	28
3.7	Visualisation du nombre d'offres par province	29
3.8	Vue d'ensemble du baromètre avec sélection de mois et provinces	30
4.1	Exemple d'une offre d'emploi	32
4.2	Même offre au format JSON	33
4.3	Loi de Zipf appliquée aux compétences du corpus des offres	35
4.4	Extrait de la liste des 6000 compétences	37
4.5	Exemple de d'étiquetage de compétences	38
4.6	Exemple de structure d'une offre d'emploi	40
4.7	Modèle graphique d'un CRF linéaire	43

4.8	Annotation selon les ensembles de test	46
4.9	Format des données préparées pour l'entraînement du CRF	48

LISTE DES ANNEXES

Annexe I :	Liste des univers	xiii
Annexe II :	Titres utilisés pour la segmentation	xvi
Annexe III :	Mots fréquemment cités autour des compétences	xxi

LISTE DES SIGLES

BPP	Butterfly Predictive Project
CRF	Conditional Random Fields
CSS	Cascading Style Sheets
DOM	Document Object Model
D3	Data Driven Documents
HMM	Hidden Markov Model
JSON	JavaScript Object Notation
LBJ	LittleBigJob
POS	Part-Of-Speech
RALI	Recherche Appliquée en Linguistique Informatique Laboratoire du Traitement Automatique du Langage Naturel de l'Université de Montréal
GIS	Geographic Information System
SVG	Scalable Vector Graphics
SVM	Support Vector Machines

REMERCIEMENTS

Je tiens d'abord à remercier mon directeur de recherche Philippe Langlais qui m'a accueilli, guidé et soutenu tout au long de ce travail. Mes remerciements vont également à Guy Lapalme pour ses conseils, ainsi qu'à tous les membres du RALI.

Merci ma famille pour votre appui continu.

CHAPITRE 1

INTRODUCTION

1.1 Préambule

L'internet est devenu un élément incontournable de notre quotidien, tant au niveau personnel que professionnel. La montée en flèche de l'utilisation des applications mobiles et l'apparition de nouveaux concepts comme l'internet des objets ¹ ont permis de générer d'immenses quantités de données, à tel point que la création de 90% des données au monde jusqu'en 2013, avait eu lieu pendant les deux années précédentes ². Le passage au web 2.0 marqué par l'émergence des médias sociaux a refaçoné les formes de communication entre les humains. On se réfère couramment à cette explosion de données par la notion de big data définie comme étant des données de grands volumes, grande variété (issues de différentes sources) et grande vélocité (temps réel), qui nécessitent des formes innovantes et rentables de traitement de l'information afin d'améliorer l'aperçu, la prise de décision, et l'automatisation des processus ³. Ces données peuvent être structurées dans le cas des bases de données privées des entreprises ou bien non structurées à l'instar des pages web, courriels, photos, vidéos, etc. Cette révolution touche presque tous les domaines, du marketing, finances aux services de santé. Ces données provenant de diverses plates-formes d'information, une fois agrégées, contiennent de l'information à haute valeur commerciale qui contribue à la réussite de l'entreprise. Les compagnies donnent de plus en plus d'importance aux modèles d'entreprises basés sur les données ⁴, et les décisions ainsi prises sont axées sur les données ⁵, non seulement pour être

¹https://en.wikipedia.org/wiki/Internet_of_things

²<https://www.sciencedaily.com/releases/2013/05/130522085217.htm>

³<http://www.gartner.com/it-glossary/big-data/>

⁴Data-Driven Business Model

⁵Data-Driven Decisions

compétitives mais pour survivre dans le monde des affaires. Le monde des ressources humaines, et plus particulièrement celui du recrutement, ne fait pas exception. Le recrutement a connu plusieurs formes, de la publication des annonces dans les journaux, en passant par la publication sur les sites de la compagnie. Désormais les recruteurs se tournent vers les réseaux sociaux (LinkedIn à titre d'exemple avec 467 millions abonnés⁶). L'E-recrutement présente plusieurs avantages : promouvoir la marque de la société, garantir une meilleure rentabilité, atteindre plus de candidats potentiels, même ceux passifs pour les inciter à penser à de nouvelles opportunités, réduire la durée du processus de recrutement, assurer l'interactivité et la fluidité d'échanges entre le recruteur et le candidat et garantir un meilleur taux de succès de trouver le bon candidat. Ces technologies sont faciles d'utilisation tant pour les candidats que pour les recruteurs. Elles permettent également d'avoir une information plus précise sur le candidat par rapport aux outils traditionnels.

Explorer et analyser ces larges volumes de données et réussir à en tirer l'information utile nécessite l'adoption d'une approche narrative. Le *storytelling*⁷ s'avère un atout majeur à cet égard pour communiquer cette information à une audience qui en a besoin pour la prise de décision. Et l'une des techniques les plus efficaces pour accomplir cette tâche est la visualisation.

1.2 Contexte du projet

Le projet BPP s'inscrit dans le cadre du partenariat entre LittleBigJob et le RALI, qui vise à répondre aux problématiques liées au domaine de l'E-recrutement, par l'application des techniques du traitement automatique du langage naturel qui constitue le champ d'expertise du laboratoire RALI.

⁶<https://press.linkedin.com/about-linkedin> (Statistiques prises en date du 11 février 2017)

⁷[https://fr.wikipedia.org/wiki/Storytelling_\(technique\)](https://fr.wikipedia.org/wiki/Storytelling_(technique))

Du fait de la quantité des profils disponibles sur le web, il devient de plus en plus coûteux pour les recruteurs de cibler les candidats adéquats à une offre donnée. LittleBigJob dispose de données issues des sites d'emploi ainsi que des réseaux sociaux professionnels, composées de millions de profils et des milliers d'offres d'emploi collectées périodiquement. Sa collaboration avec le RALI vise l'automatisation des différentes étapes du recrutement, ce qui aura un impact tant sur la qualité que sur le coût de ce processus. En effet, après la réception de l'offre, le chasseur de tête entame la phase de recherche en passant en revue les profils des candidats à sa disponibilité pour en sélectionner ceux jugés pertinents, puis il contacte ceux retenus pour les inciter à répondre à l'offre. Chacun des quatre volets suivants du projet BPP vise à répondre aux besoins liés à chacune des étapes de ce processus.

1.2.1 L'appariement profils/offres

La première étape du processus de recrutement pour une agence de placement consiste à chercher les profils potentiels qui répondent aux exigences énumérées dans l'offre. Ce processus est manuel dans la plupart des cas où le *Career Manager* procède par des recherches par mots clés pour trouver les candidats correspondants. Il faut noter l'impact de cette approche sur la qualité des résultats obtenus. Développer un système d'appariement qui suggère une liste de candidats potentiels mêmes passifs pour les inciter à répondre l'offre en question fait objet du travail de Dieng [11] qu'il a réalisé dans le cadre de son mémoire. Le système développé consiste à faire l'extraction des principales sections d'une offre d'emploi, à savoir le titre, les compétences et les expériences, puis l'interrogation de la base de données des profils avec compétences et expériences correspondantes. Un score est affecté à chaque profil selon son degré de similarité avec l'offre ce qui permet de déterminer la liste finale des profils retenus. Ces extractions nécessitent la disponibilité de ressources linguistiques structurées, autrement dit des on-

tologies spécifiques au domaine des ressources humaines.

1.2.2 La génération automatique d'ontologie

Faire correspondre une offre d'emploi à un profil nécessite le parcours du texte pour l'extraction de différentes entités telles que le titre, les compétences, les expériences. Cette extraction nécessite la disponibilité de ressources linguistiques normalisées et structurées qui jouent le rôle de référence linguistique du projet. Ces ressources doivent être dynamiques, autrement dit, capables d'évoluer en réponse aux mutations que connaît le marché de l'emploi, ce qui implique l'apparition de nouveaux métiers, compétences, etc. C'est dans ce cadre que s'inscrit le travail de Kessler et al. [23] qui consiste à générer une ontologie dans le domaine des ressources humaines. La création de cette ontologie consiste à normaliser les offres d'emploi, puis à utiliser les titres de ces offres pour l'extraction des métiers et des univers⁸, ensuite l'utilisation du vocabulaire pour extraire les compétences, et finalement la transformation en une ontologie au format RDF enrichie d'une manière semi-automatique, avec une classification univers, métiers et compétences.

1.2.3 La génération des lettres de motivation

Après avoir sélectionné les candidats, vient l'étape d'envoi des propositions pour les inciter à répondre à l'offre, via la génération automatique personnalisée des lettres de motivation. Au lieu d'envoyer des lettres basées sur un modèle figé qui ne prend pas en compte le profil du candidat, il s'avère bénéfique de personnaliser la lettre de motivation en fonction du profil ciblé, afin d'augmenter les chances d'attirer les candidats passifs. Le projet de mémoire de Grand'Maison [15] qui s'inscrit dans le cadre de l'un des principaux domaines du traitement automatique du langage naturel qui est la génération

⁸Un univers est un groupement de secteurs d'activité similaires

automatique du texte, vise à répondre à ce besoin et consiste à développer un système de génération de lettres destinées aux candidats sélectionnés par le système d'appariement. Une lettre est typiquement composée d'une section *Présentation* pour présenter le recruteur, *Qualifications* qui relate les compétences du candidat, *Formules d'appel*, *Contact* et *Salutations*.

1.2.4 Baromètre de l'emploi

Cette partie constitue le sujet de travail de ce mémoire, où il s'agit de développer une solution pour la visualisation de données. En effet, chaque jour, des milliers d'offres d'emploi sont diffusées sur Internet en même temps que des milliers de profils sont créés ou modifiés sur les sites d'emploi ou les réseaux sociaux. Plusieurs organismes privés ou publics publient périodiquement des statistiques permettant de mesurer les tendances. Il ne s'agit cependant que de « photos » reflétant l'état du marché à une période donnée. L'objectif de ce volet serait donc de concevoir et d'implémenter des outils intelligents identifiant sur le web en quasi temps réel l'ensemble des profils et des offres qui paraissent et permettant de visualiser la tendance du recrutement.

1.3 Structure du mémoire

Dans le deuxième chapitre de ce mémoire nous présentons l'état de l'art de la visualisation des données, son évolution, les techniques utilisées et les domaines d'application. Au chapitre 3 nous expliquons la solution mise en place pour la visualisation, sa structure et les différentes fonctionnalités offertes. Dans le chapitre 4 nous allons décrire l'approche suivie pour l'extraction des compétences, la méthodologie suivie pour l'implémentation, les défis rencontrés et l'analyse des résultats obtenus.

CHAPITRE 2

ÉTAT DE L'ART

Avant d'aborder la littérature de la visualisation, nous estimons nécessaire de faire la clarification quant à la terminologie, puisqu'on trouve dans les travaux du domaine de la visualisation les deux termes 'visualisation des données' et 'visualisation de l'information'.

Le théoricien des organisations américain Russell Ackoff dans sa définition des quatre concepts données, information, connaissances et sagesse, établit cette définition : les données sont des symboles qui représentent les propriétés des objets et des événements. L'information se compose de données traitées, ce traitement visant à accroître leur utilité. Par exemple, les recenseurs recueillent des données. Le Bureau du recensement traite ces données, les convertit en information qui est présentée dans de nombreux tableaux publiés sous forme de statistiques. Comme les données, l'information représente également les propriétés des objets et des événements, mais de manière plus compacte et utile que les données. La différence entre les données et l'information est fonctionnelle, et non structurelle. L'information est contenue dans les descriptions, les réponses aux questions qui commencent par des mots tels que, qui, quoi, quand, où, et combien [1].

Friendly et Denis [14] soulignent également la différence entre 'la visualisation de l'information' et 'la visualisation de données' : le terme visualisation de l'information est généralement appliqué à la représentation visuelle, à grande échelle, des collections d'information non numérique, telles que des fichiers et des lignes de code dans les systèmes logiciels, les bases de données des bibliothèques et bibliographiques, réseaux de relations sur l'Internet, etc. Par contre, la visualisation des données est la science de la représentation visuelle des **données**, définies comme étant une information abstraite sous

forme schématique, y compris les attributs ou variables pour les unités d'information.

La visualisation de l'information, comme outils et techniques utilisés aujourd'hui tant dans le domaine académique que dans l'industrie, est une discipline récente. On trouve plusieurs définitions de la visualisation d'information selon le contexte d'application. Card et al. [6] la définissent comme *'The use of computer-supported, interactive, visual representations of abstract data to amplify cognition'*. Cette définition met en évidence l'utilité de la visualisation comme moyen de communication qui réside dans sa capacité à optimiser la perception humaine qui traite l'image plus rapidement que le texte et au fait que l'image permet de véhiculer les concepts indépendamment du langage utilisé.

2.1 Évolution de la visualisation d'information

Les premières visualisations ont été qualitatives et utilisées principalement pour les voyages, le commerce, la communication et la religion [37]. Dans la carte de Peutinger¹ par exemple, figurent les routes et les villes principales de l'Empire romain. Ces représentations manquent toutefois de précision, les distances entre les villes par exemple ne sont pas exactes. Au 17^{ème} siècle, avec la naissance de la théorie des probabilités et l'évolution des statistiques, la visualisation a connu un important essor. La représentation visuelle des données quantitatives par un système de coordonnées à deux dimensions, la forme la plus commune de ce que nous appelons des graphiques, est apparue avec René Descartes, qui a inventé cette méthode à l'origine, non pas pour présenter des données, mais pour effectuer des opérations mathématiques basées sur un système de coordonnées (coordonnées cartésiennes). Plus tard, cette représentation a été reconnue comme un moyen efficace de visualisation de l'information [13].

Deux exemples classiques dans le domaine de la visualisation montrent l'importance

¹<http://peutinger.atlantides.org/map-a/>

des représentations graphiques. La figure 2.1² de l'ingénieur français Charles Joseph Minard (1781-1870) est un exemple de communication efficace qui montre la défaite de l'armée française en Russie. L'épaisseur des deux bandes représente l'effectif de l'armée, la couleur précise la direction : la bande en noir présente l'effectif de l'armée durant le retrait, l'autre représente l'invasion. En bas, une courbe de températures, avec les dates correspondantes, est liée aux différents points à travers le retrait. L'excellence de cette représentation réside dans la combinaison des données géographiques et des séries temporelles (time series) dans une représentation statique, et la représentation de données multivariées (six variables) : l'effectif de l'armée, sa position sur une surface de deux dimensions, sa direction, la température à différentes dates. Tuftte considère cette représentation comme étant le meilleur graphique statistique jamais dessiné [36].

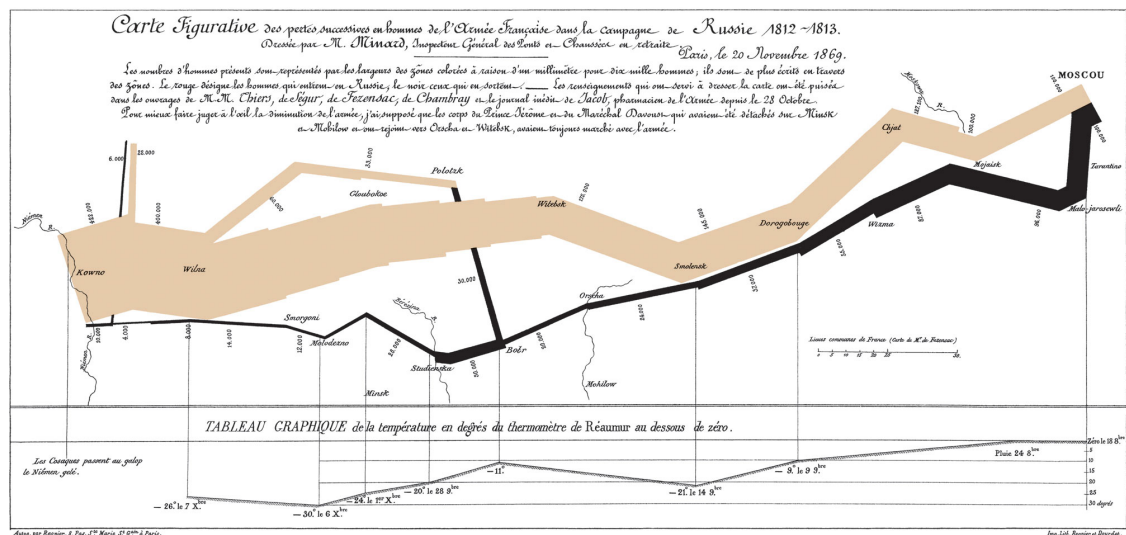


Figure 2.1 – Carte de la campagne de Napoléon en Russie (1812-1813) par C.J. Minard

L'autre exemple est la carte du physicien John Snow représentée en figure 2.2³. Cette visualisation illustre les décès suite à l'épidémie du Choléra survenue en 1854 à

²https://en.wikipedia.org/wiki/Charles_Joseph_Minard#/media/File:Minard.png

³https://en.wikipedia.org/wiki/John_Snow#/media/File:Snow-cholera-map-1.jpg

Londres. Les barres en noir représentent le nombre des cas de décès par habitation. Snow a constaté une forte concentration des décès autour du *Broad Street* et a recommandé la fermeture de la pompe de ce quartier ce qui a eu comme conséquence la fin de l'épidémie. Cet exemple montre l'importance du choix de la bonne présentation et la découverte de connaissances à travers une relation de cause à effet. L'hypothèse du Dr. Snow était que l'épidémie se propage à travers de l'eau. Il a eu recours à cette représentation visuelle qui illustre le nombre de morts pour confirmer son hypothèse.

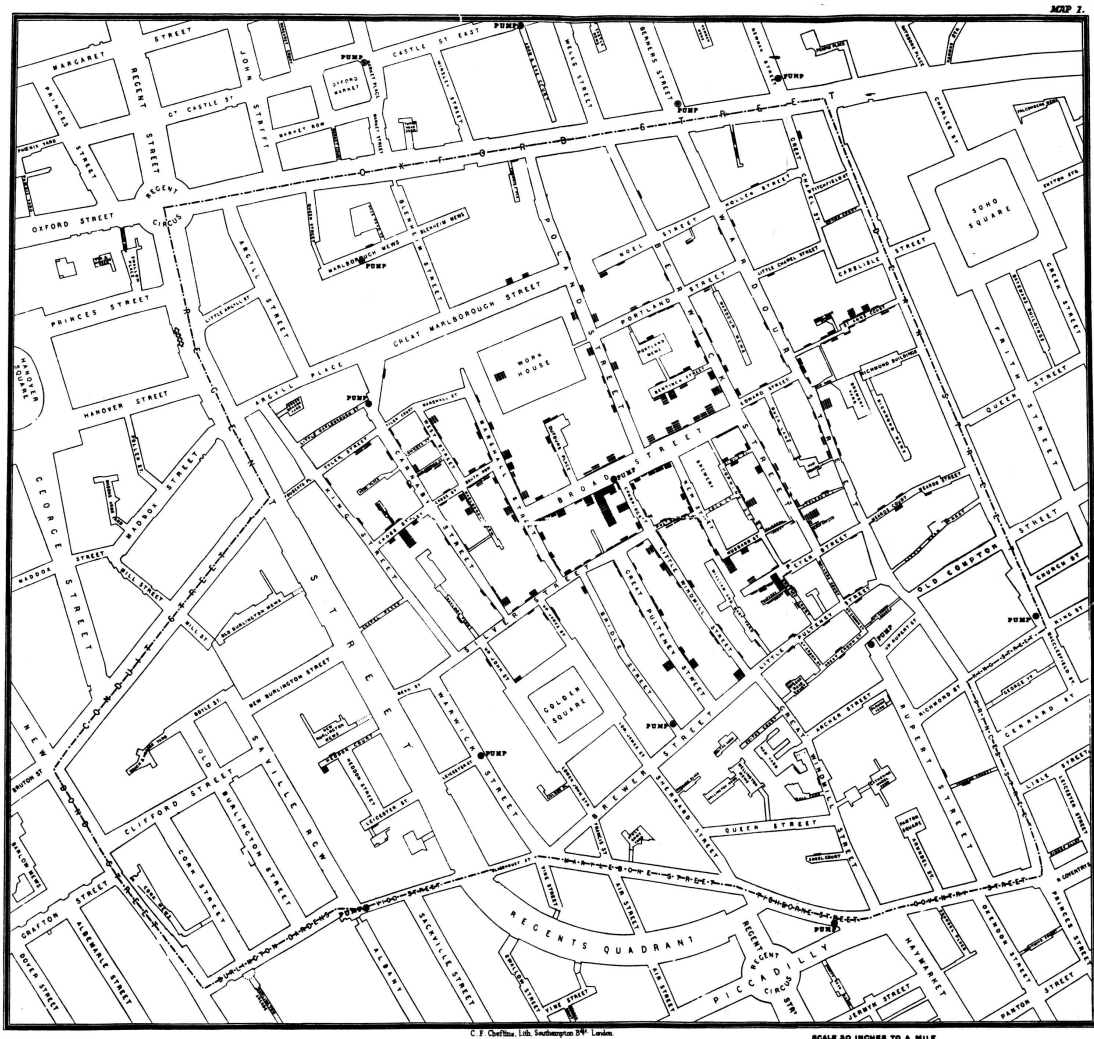


Figure 2.2 – Carte des décès pendant l'épidémie de Londres de 1854

A la fin du 18^{ème} et au début du 19^{ème} siècle William Playfair, ingénieur et économiste politique écossais, invente nombreux graphiques utilisés jusqu'à nos jours, y compris les graphiques à barres et les diagrammes circulaires (en secteur). En 1977, le professeur des statistiques John Tukey va développer une approche essentiellement visuelle pour explorer et analyser les données, appelée analyse exploratoire des données, qui recommande l'utilisation de méthodes visuelles pour donner un aperçu statistique rapide des ensembles de données. En 1983, un autre pionnier de la visualisation Edward Tufte publie son livre *The Visual Display of Quantitative Information* considéré comme l'une des principales références dans ce domaine. Il y a mis au point les règles essentielles pour une visualisation excellente et intégrée. Avec l'accès à des ordinateurs munis de capacités graphiques puissantes, la visualisation de l'information émerge comme une nouvelle branche de recherche dans le monde académique. En 1999, le livre *Readings in Information Visualization : Using Vision to Think* a rassemblé les travaux de ce domaine en un seul volume rendu accessible au-delà du milieu universitaire [13].

2.2 Types de données

Ward et al. [37] définissent deux catégories d'informations :

- Ordinale (valeurs numériques) avec trois sous-catégories : binaire qui admet deux valeurs 0 et 1 ; discrète qui définit des valeurs entières à partir d'un sous ensemble par ex. (2, 4, 6) ; et continue qui représente un intervalle de valeurs par ex. [0, 5].
- Nominale (valeurs non numériques) avec trois sous-catégories : catégorique où les valeurs sont sélectionnées à partir d'un ensemble fini de possibilités (rouge, bleu, vert) ; ordonnée qui est une variable catégorique impliquant un ordre (petit, moyen, grand) ; et aléatoire avec une plage infinie de valeurs sans aucun ordre donné (adresses). Ils définissent également une autre méthode de catégorisation

des variables par le biais de la notion d'échelle définie par trois attributs : la relation d'ordre, la métrique de distance et l'existence du zéro absolu.

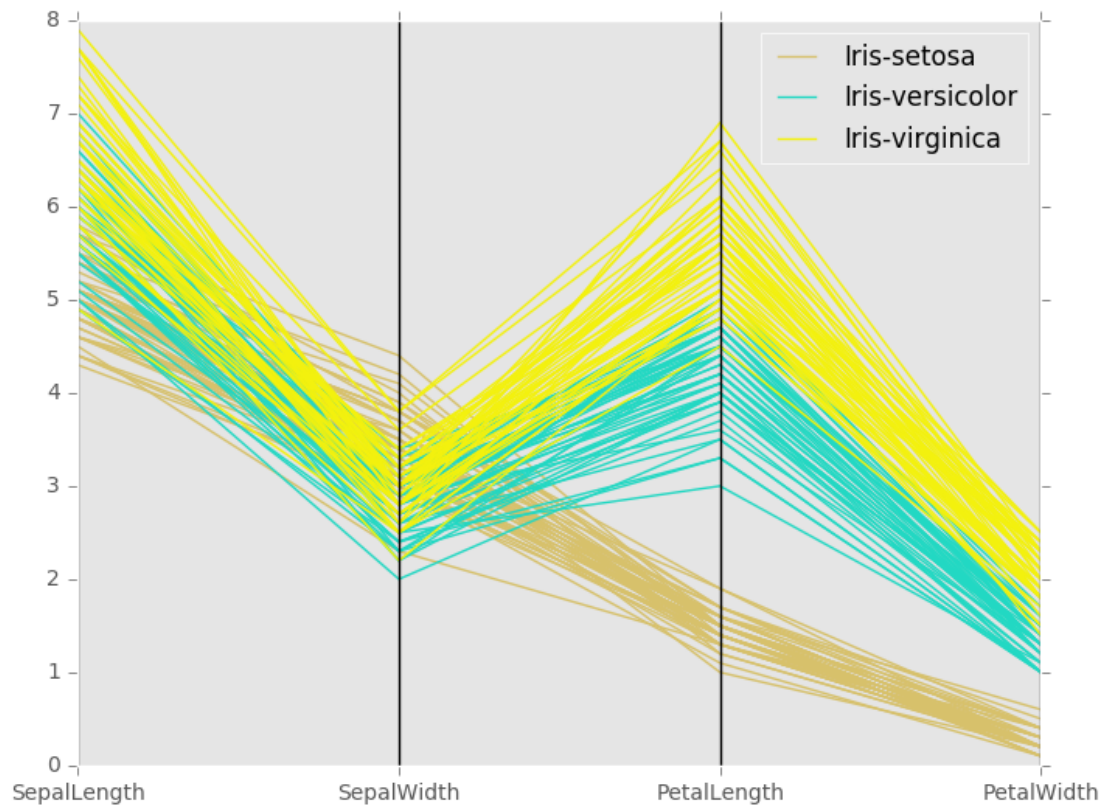
Shneiderman [33], adopte une autre classification des types de données. Il définit sept types de données :

- Les données à une dimension : qui sont des données linéaires comme les documents textuels ou du code source.
- Les données à deux dimensions : comprennent les données planaires comme le cas des cartes géographiques où chaque objet de la collection couvre une partie de l'espace global. Les systèmes d'information géographique en est un exemple.
- Les données à trois dimensions : Il s'agit d'objets du monde réel (par ex. molécules, corps humain) constitués d'objets avec un volume et une relation complexe avec d'autres objets.
- Les données temporelles : la caractéristique de ce type qui le distingue des données à une dimension est que les objets ont un temps de départ et d'arrivée et ces objets pourraient se chevaucher (par ex. les dossiers médicaux, la gestion de projet).
- Les données multidimensionnelles : Type de données courant dans les bases de données statistiques où les objets, définis avec n attributs, sont modélisés comme des points dans un espace à n dimensions. Parmi les techniques couramment utilisées pour représenter les données multidimensionnelles, les diagrammes de dispersion (scatter plots) et les coordonnées parallèles (parallel coordinates) illustrées en figure 2.3⁴
- Les données de type arbre (hiérarchiques) : arborescence d'objets, où chacun a un

⁴<http://pandas.pydata.org/pandas-docs/stable/visualization.html>

lien vers un seul objet parent (sauf la racine). Les objets et les liens entre fils et parent peuvent avoir plusieurs attributs.

- Les données de type réseau : Quand les objets sont liés arbitrairement à d'autres objets et que les relations entre ces objets ne peuvent pas être modélisées sous forme d'arbres. (par ex. fichier XML, jeu d'échecs sur ordinateur, réseaux sociaux).



2.3 Taxonomies de visualisation d'information

Les taxonomies s'avèrent un outil efficace pour représenter une classification afin de définir la hiérarchie et les relations entre les objets. Plusieurs taxonomies ont été élaborées.

rées pour classifier les techniques de visualisation, les critères de classification peuvent être les types de données, les techniques de visualisation ou les méthodes d'interaction. Dans ce qui suit, nous exposons certaines taxonomies liées à la littérature de la visualisation de l'information. Cette analyse comparative ne se veut pas être exhaustive, elle a comme objectif de présenter les approches courantes dans le domaine de la visualisation.

L'une des premières taxonomies est celle proposée par Shneiderman [33], qui résume les principes de base de la conception visuelle par ce que l'auteur appelle '*Visual Information Seeking Mantra : overview first, zoom and filter, then details-on-demand*'. Elle est définie comme une taxonomie de 'tâche par type de données' : en plus des sept types de données décrits dans la section 2.2, à savoir les données unidimensionnelles, bidimensionnelles, tridimensionnelles, multidimensionnelles, temporelles, de type arbre, et de type réseau, Shneiderman définit sept tâches : aperçu, zoom, filtre, détails à la demande, liens, historique, et extrait. Les données sont vues comme une collection d'objets et ces derniers ont plusieurs attributs. Sur la base de ces deux aspects à savoir les types de données et les tâches par domaine cette taxonomie est organisée comme indiqué dans le tableau 2.I.

Dos Santos [12] souligne un problème lié à l'application de cette taxonomie, qui ne fait pas la distinction entre les méthodes et les systèmes logiciels. Il considère que la compréhension de la méthodologie est plus important que la compréhension de son implémentation technique, étant donné que les systèmes logiciels sont utilisés pendant une durée de vie assez limitée par rapport aux méthodes qu'ils implémentent et que ces systèmes ne sont pas tous libres de droits, alors qu'une technique devient accessible une fois qu'elle a été publiée.

Une autre taxonomie est celle proposée par Keim [22]. Il s'agit d'une classification à trois dimensions : le type de données à visualiser (données unidimensionnelles, bidimensionnelles, multidimensionnelles, texte et hypertexte, hiérarchies et graphes, et

Type de données	Tâche ou problème à résoudre
Unidimensionnel	- Trouver le nombre d'objets ; - Trouver des objets avec des attributs.
Bidimensionnel	- Trouver des objets adjacents ; - Trouver les chemins entre les objets ; - Compter, filtrer, détails à la demande.
Tridimensionnel	- Adjacence, au-dessus/au-dessous et les relations à l'intérieur/à l'extérieur.
Multidimensionnel	- Trouver des patterns, des clusters des corrélations entre les paires de variables les gaps et les valeurs aberrantes (outliers).
Temporel	- Trouver tous les événements avant, après un moment ou pendant une période donnée.
Arbre	- Identifier des propriétés structurelles : par exemple niveaux de l'arbre, le nombre de fils d'un objet.
Réseau	- Chemin le plus court ou le moins coûteux entre deux objets ; - Traverser tout le réseau.

Tableau 2.I – Problèmes à résoudre par type de données

algorithme et logiciel), la technique de visualisation (2D et 3D, géométrique, iconique, pixel dense et empilée), et le troisième axe qui représente la technique d'interaction et de distorsion utilisée (projection dynamique, filtrage interactif, zoom interactif, distorsion interactive et raccordement et brossage interactifs), les trois axes sont considérés orthogonaux. Selon Keim, il faut combiner les qualités humaines telles que la créativité, la flexibilité et la connaissance avec les capacités de stockage et de calcul des machines pour une exploration visuelle des données volumineuses. En effet, il considère la visualisation des données comme un processus de génération d'hypothèses dont la validation peut se faire soit par la visualisation elle-même, où par d'autres techniques statistiques.

Selon ce schéma de classification, une technique de visualisation est associée avec un type de données. Ce dernier à son tour est associé à une technique d'interaction et de distorsion.

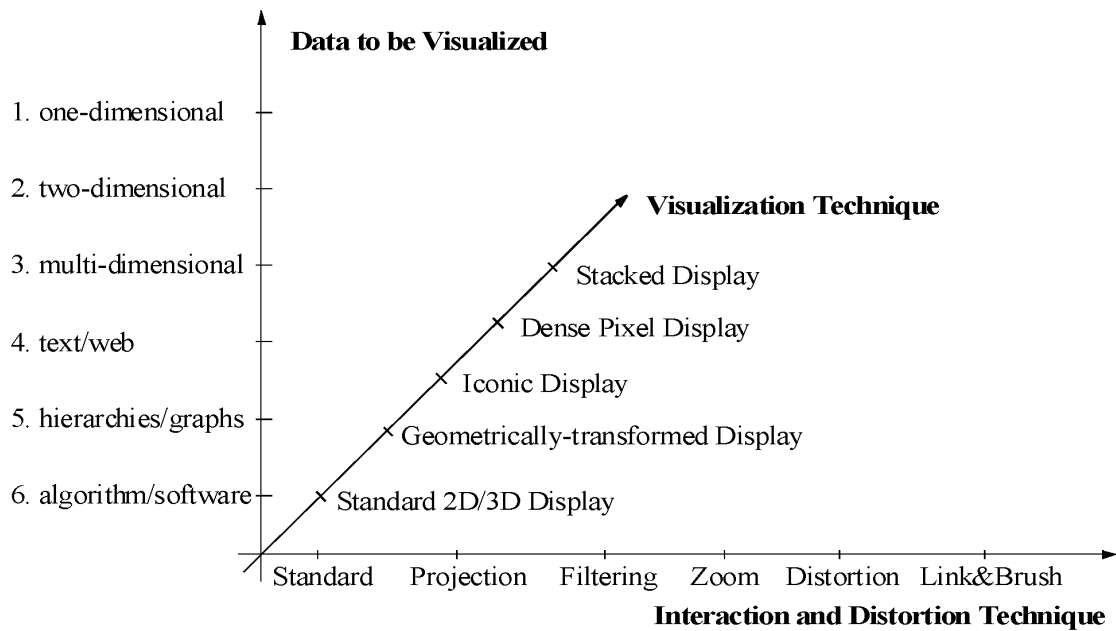


Figure 2.4 – Classification des techniques de visualisation selon Keim

Dans un travail antérieur du même auteur [21], il a présenté une classification basée sur trois concepts : la technique de visualisation, la technique d’interaction et la technique de distorsion, le type de données n’étant pas pris en compte comme dimension. Dos Santos [12] considère cette conception inadéquate puisque les techniques de distorsion peuvent être considérées comme un aspect interactif d’une technique de visualisation, qui modifie la façon dont les données sont présentées. Cette lacune a été comblée dans la deuxième version de la classification [22] où les techniques d’interaction et de distorsion sont fusionnées en un seul axe.

Une autre classification intéressante est présentée par Chi [8], qui adopte un modèle d’état de données *data state model*, plus détaillé, qui consiste à analyser les différentes techniques de visualisation par différents opérateurs qui peuvent servir soit à la transformation des données, la transformation de la visualisation ou la transformation du ‘mapping’ visuel.

La taxonomie définie par Tory et Moller [35] classifie les techniques selon les ca-

ractéristiques du modèle de données plutôt que les données elles-mêmes. La structure de la taxonomie est basée sur deux critères : la nature des modèles de conception qui sont soit continus, soit discrets et le degré de l’affichage des attributs par le concepteur. CARR [7] décrivent une classification cognitive pour les visualisations 3D qui est basée sur trois aspects : l’attention, l’abstraction et l’interaction, contrairement à un système de visualisation scientifique. Rodrigues et al. [30] considèrent cette taxonomie comme un guide pour comprendre la nature subjective des techniques de visualisation.

2.4 Visualisation de l’information dans le web

Afin de permettre la comparaison des outils de visualisations, Bostock et Heer [4], définissent deux catégories d’outils de visualisations : les systèmes graphiques qui fonctionnent sur des primitives graphiques de bas niveau. Ces systèmes comprennent des programmes de dessin vectoriel (par ex. Adobe Illustrator) et des APIs de faible rendu (par ex. OpenGL). Puis, les systèmes de visualisation qui se basent sur des abstractions et des modèles mathématiques et qui supportent la gestion des données, les algorithmes de mise en page, l’interaction et l’animation. Bostock et Heer définissent trois catégories de systèmes de visualisation :

- Les logiciels grand public : Ces outils de visualisation sont largement utilisés en raison de la facilité de leur utilisation, la création d’un diagramme ne nécessite pas des actions compliquées. Des exemples de ces outils sont les tableurs tels que Microsoft Excel et Google Spreadsheets.
- Les outils analytiques et exploratoires : Contrairement à la catégorie précédente, un bon nombre de ces outils ont vu le jour au sein de la communauté de la recherche et offrent plus d’options pour l’exploration visuelle des données. Malgré leur puissance, ces outils n’offrent pas une flexibilité suffisante à l’utilisateur pour

la personnalisation des visualisations. Le produit Tableau ⁵ est un exemple de ces systèmes.

- Les outils (toolkits) de programmation : certains de ces outils offrent un choix restreint de graphiques comme GoogleChart, pour permettre une manipulation facile à l’instar des logiciels grand public. En revanche, d’autres solutions plus riches, comme InfoVis et Prefuse, offrent un cadre intégré de gestion de données couplé à des composants de visualisation et d’interaction. Elles peuvent être étendues par la création de nouveaux composants, ce qui nécessite toutefois une maîtrise significative du développement logiciel.

Évolution de la visualisation sur le web

Une des applications les plus courantes de visualisation de l’information peut être considérée comme la visualisation sur le Web, en raison de l’utilisation massive des navigateurs. Les outils de visualisation sur web se basent principalement sur trois technologies : SVG, canvas HTML5 et JavaScript. Afin de relater l’évolution de cette catégorie de visualisations, nous avons choisi trois bibliothèques, que nous décrivons en termes de leurs caractéristiques et des fonctionnalités qu’elles offrent :

- Prefuse [18] : développée en Java en utilisant la bibliothèque graphique Java2D, sa version Prefuse Flare offre une bibliothèque ActionScript pour la création de visualisations qui s’exécutent sur Adobe Flash Player. La base théorique de Prefuse est inspirée du modèle de référence de visualisation d’information [9] (aussi appelé data state model). Le processus de visualisation commence par l’étape du ‘filtering’ qui consiste en une série d’actions qui transforment les données abstraites en contenu visualisable avec des propriétés visuelles (position, couleur, taille, police, etc.). Puis les modules de rendu (Render modules), permettent de dessiner des

⁵<https://www.tableau.com/>

VisualItems pour construire des affichages interactifs. Toutefois, l'utilisation des applets Java limite la flexibilité et les performances de tels outils sur le web

- Protovis [4] : a été développé pour offrir plus de contrôle bas niveau sur le design, contrairement aux outils analytiques dont l'expressivité est limitée. Dans Protovis, les concepteurs spécifient les visualisations comme une hiérarchie de marks (primitives graphiques comme les barres, les lignes et les labels) avec des propriétés visuelles définies comme des fonctions de données. L'héritage des propriétés à partir des marks composées, semblable au CSS, permet des définitions de visualisation concises avec une grande expressivité et un minimum d'abstractions intermédiaires. Protovis supporte JavaScript, HTML 5 canvas, SVG et Flash.
- D3.js [5] : successeur de Protovis, D3 permet la manipulation directe du DOM. Autrement dit, les données peuvent être liées directement à des éléments du document. A l'inverse de certaines bibliothèques qui restreignent le choix en offrant des graphiques prédéfinis à utiliser, D3 qui combine SVG, CSS, JavaScript et HTML donne plus de contrôle au concepteur sur la page web, et permet la création de graphiques personnalisés.

Dans le cadre du projet BPP, nous avons fait le choix de la bibliothèque D3.js ⁶ pour la conception du baromètre de l'emploi. Bostock et Heer [4] définissent trois critères à prendre en considération lors du choix de l'outil de visualisation : l'expressivité qui reflète la diversité des visualisations offertes par l'outil ; l'accessibilité qui exprime le degré de difficulté à apprendre la représentation et l'efficacité qui correspond à la réduction de l'effort nécessaire pour spécifier une visualisation. D3, successeur de Protovis, grâce à la manipulation directe et transparente du DOM, apporte une amélioration à l'expressivité et l'accessibilité en offrant une efficacité comparable à celle de Protovis.

⁶<https://d3js.org/>

Nous avons donc fait dans ce chapitre le survol de la visualisation, son évolution à travers l'histoire, les principales taxonomies de la littérature, et l'impact de l'avènement de l'ère de l'informatique sur les applications de la visualisation, et notamment les atouts du web. Les efforts continus d'amélioration menés dans ce sens se sont concrétisés par la réalisation de plusieurs applications, dont la librairie D3 que nous avons retenue pour implémenter le baromètre de l'emploi, dont les motivations, la conception et le fonctionnement sont relatés au chapitre suivant.

CHAPITRE 3

BAROMÈTRE DE L'EMPLOI

La grande quantité des données apparaissant chaque jour sur le web tant en termes d'offres d'emploi que de profils présente un vrai défi pour les recruteurs et les candidats. Avoir les outils nécessaires pour explorer ces données se révèle un grand avantage, ce qui permet, à titre d'exemple, de visualiser la distribution géographiques des offres d'emploi, la disponibilité des profils, les tendances en termes des compétences recherchées, les employeurs potentiels, les secteurs avec plus d'opportunités, etc.

Le baromètre de l'emploi tente de répondre à ces besoins afin de donner à l'utilisateur final, qu'il soit candidat en quête de nouvelles opportunités ou recruteur qui vise à atteindre des profils potentiels, une vue globale sur le marché d'emploi en termes d'offres, sous forme graphique. Il s'agit d'un tableau de bord permettant la visualisation des différentes composantes relatives à une offre, de relater l'historique et de d'explorer les tendances du recrutement.

3.1 Préparation des données

Les données qui alimentent le baromètre sont issues de deux fichiers JSON : le premier fichier contient les coordonnées des provinces canadiennes qui servent au dessin de la carte géographique, il est au format GeoJSON. Il s'agit d'un format de codage destiné aux structures de données géographiques. Un objet GeoJSON peut représenter une géométrie, une caractéristique (feature), ou une collection de caractéristiques. GeoJSON supporte les types de géométrie suivants : le point, la ligne polygonale (LineString), le polygone, les multi-points, les multi-lignes polygonales (MultiLineString), et les collections géométriques (GeometryCollection). Les features en GeoJSON contiennent un

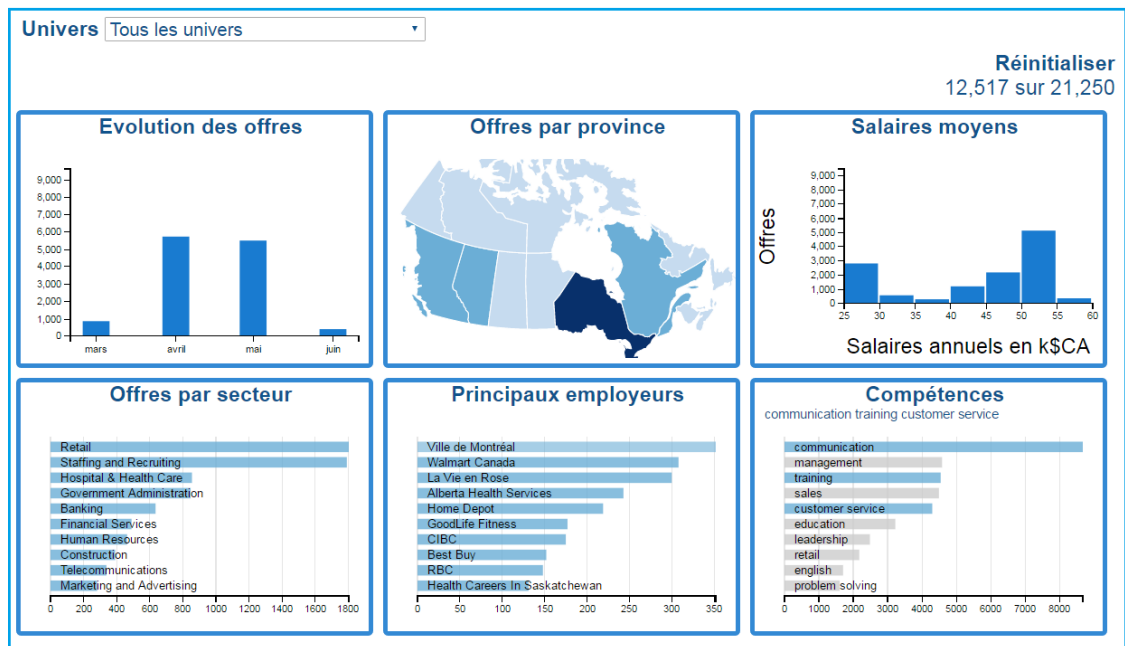


Figure 3.1 – Visualisation interactive des offres d’emploi

objet de géométrie et des propriétés supplémentaires, et une collection de features qui représente une liste de features.¹

Le deuxième fichier contient les données qui alimentent le baromètre. Les offres d’emploi collectées par LBJ sont relatives à la France et au Canada. Nous nous limitons aux offres du Canada pour la visualisation. Le fichier ainsi créé pour alimenter le baromètre contient les champs suivants, dont certains sont extraits directement du fichier des offres d’emploi, ce qui est le cas pour :

- **_id** : l’identifiant de l’offre.
- **city** : la ville où est situé le poste.
- **province** : code de la province. Les champs **city** et **province** sont le résultat de la décomposition du champ **place** du fichier des offres d’emploi où sont renseignés à la fois la ville et le code de la province.

¹<http://geojson.org/>

```

{"type": "FeatureCollection",
  "features": [{"type": "Feature",
    "properties": {"name": "SK"},
    "geometry": {"type": "Polygon",
      "coordinates": [[[-110,50],[-110,51],[-110,53], [-110,55],[-110,56],
        [-110,58],[-110,60],[-108,60],[-107,60],[-105,60],[-103,60],
        [-102,60],[-102,57.3531608581543],[-102,55.804649353],
        [-101.84937286376953,54.41109085083008],
        [-101.7439193725586,53.36949157714844],
        [-101.66483306884764,53.018],
        [-101.64192199707,52.318244934082024],
        [-101.5716400,51.976516],
        [-101.5643386840,51.3036766052],
        [-101.4711608886,50.593856811],
        [-101.34937286376952,49.00117492675781],[-103,49],
        [-105,49],[-108,49],[-109.98324268599357,49],[-110,50]]]]
    }
  ]
}

```

Figure 3.2 – Exemple de coordonnées de la province de Saskatchewan au format GeoJSON

- **date_update** : date de publication de l’offre.
- **company_name** : le nom de la compagnie. Les champs suivants sont déduits à partir des champs du fichier des offres d’emploi.
- **sector** : le secteur d’activité de l’entreprise qui a publié l’offre.
- **universe** : il s’agit d’un concept élaboré par LBJ, qui permet de regrouper plusieurs secteurs d’activité similaires en une seule entité. Par exemple, les secteurs *Banking/Credit*, *Capital Markets*, *Financial Services*, *Investment Banking*, *Investment Management* et *Venture Capital and Private Equity* appartiennent à l’univers *banking, finance, capital risk, private funds*. Cette catégorisation a donné lieu à 48 univers. L’annexe I relate la liste des univers définis chacun avec son identifiant

et son libellé. Pour déterminer le secteur et l'univers associé à chaque offre, nous nous basons sur une base de données composée de 333 422 lignes, avec quatre colonnes : le nom de la compagnie, le nom normalisé, le secteur d'activité et l'univers. La normalisation du nom de la société sert à unifier les noms qui figurent avec différentes variantes.

- **salary** : l'extraction des salaires montre qu'ils sont renseignés dans seulement 14% d'offres. En plus, ils sont formulés de différentes façons, illustrées en figure 3.3 : taux horaire, salaire annuel, salaire de base plus bonus, etc. Ainsi, pour l'intégration des salaires dans le baromètre, nous nous sommes basés sur les données issues du site *Emploi Québec* ². Le salaire affiché dans le baromètre est le salaire annuel moyen par univers.

- Remuneration \$200,000 estimated
- These positions pay a base with a bonus for sales, around \$65K/yr
- The starting salary is \$12.00/hr
- \$12.00 per hr
- Rate of Pay : Pay Band 11 \$20.370 to \$21.810 (3 step range)
- Up to \$70,000
- \$14.00 - \$15.00/hr
- Wage : \$60,000 - \$65,000K
- CDL truck drivers are eligible for : Up to \$65,000 per year* (up to \$0.40 per mile*)
\$3,000 sign-on bonus for experienced drivers Up to \$6,000 tuition reimbursement for qualified drivers Accessorial pay plus the potential for \$0.02/mile performance bonuses
- Starting wage is \$11.20/hour with a \$1.00 increase after 12 AM (MST)
- Starting wage range : \$15.00 to \$18.50 per hour

Figure 3.3 – Exemples de formulations des salaires dans les offres d'emploi

- **skills** : les compétences sont énumérées dans le champ description de l'offre, leur extraction nécessite le recours à l'apprentissage automatique, ceci fait l'objet du chapitre suivant.

²<http://imt.emploi.quebec.gouv.qc.ca/>

Mettre en place un tableau de bord en se basant uniquement sur D3 est complexe. Etant donné que l'objectif du tableau de bord est d'intégrer différents composants graphiques, les mettre en interaction et faire des sélections et des filtrages en tenant compte des différentes dimensions, nous utilisons, outre D3, deux bibliothèques basées également sur JavaScript conçues pour faciliter cette tâche : Crossfilter et dc.js. Crossfilter, une bibliothèque développée à l'origine par Mike Bostock pour Square ³ qui permet l'exploration de larges ensembles de données via la génération de dimensions sur ces ensembles. Une dimension de données peut être considérée comme un type de groupement de données, où chaque élément de ces données dimensionnelles est une variable catégorique [40]. Crossfilter fournit la fonction `Group` basée sur le concept `Mapreduce` [10] pour faire des groupements selon sur une dimension donnée. Comme Crossfilter est conçue pour manipuler les données, il devient difficile de tirer parti de la puissance de cette bibliothèque lors de son utilisation avec D3. C'est pour pallier à cette contrainte que dc.js a été introduite. Cette dernière, développée par Nick Qi Zhu, nécessite D3 et Crossfilter comme prérequis, dans le but de permettre de concevoir des graphiques plus facilement tout en exploitant la puissance de D3 comme bibliothèque de manipulation des graphiques et de Crossfilter comme bibliothèque de manipulation de données.

La figure 3.4 montre l'exemple de conception du graphique *Company*, qui illustre les principaux employeurs. Les trois premières commandes permettent de créer une instance Crossfilter définie par rapport à la colonne *company_name* du fichier des données JSON des données du baromètre décrit à la section 3.1. La création d'une dimension et d'un groupement par rapport à ce champ, permet d'agrèger les mêmes instances de la colonne en question et d'interagir avec les sélections effectuées sur les autres graphiques du baromètre. La dernière commande illustre l'utilisation de dc.js pour la création du graphe par la définition de ses différents attributs tels que la longueur, la largeur, la cou-

³<https://github.com/square/crossfilter>

leur, le ‘mapping’ avec la dimension Crossfilter définie (en l’occurrence *Company*), la définition de l’ordre de classement des barres de graphes, le nombre de barres à afficher, etc. La figure 3.5 montre le résultat de cette implémentation, où le graphe des principaux employeurs illustre les différentes compagnies correspondantes à la sélection de trois provinces, deux mois, et deux secteurs.

```
// Créer une instance Crossfilter  
var ndx = crossfilter(dataSet);  
// Définir la dimension par rapport au champ Company  
var company = ndx.dimension(function(d) return d.company_name; );  
// Définir le groupement pour la dimension ainsi créée  
var offersBycompany = company.group();  
// Dessiner le graphique avec spécification des différents attributs  
var companyChart = dc.rowChart("#company-chart");  
companyChart  
.width(300)  
.height(300)  
.colors("#6baed6")  
.dimension(company)  
.group(filtered_companies)  
.gap(5)  
.ordering(function(d) return -d.value; )  
.data(function (group) return group.top(10); )  
.elasticX(true)  
.xAxis().ticks(10)  
.tickFormat(d3.format("d"))  
.tickSubdivide(0);
```

Figure 3.4 – Utilisation de Crossfilter et dc.js pour la conception du graphique *Company*

Six graphiques sont conçus pour le baromètre, en plus d’une liste déroulante où sont listés les univers. Ces graphiques représentent les champs décrits précédemment, à savoir l’univers, le secteur, la localisation géographique, la date de publication, le salaire annuel moyen, les compagnies qui recrutent et les compétences requises. La figure 3.8 illustre

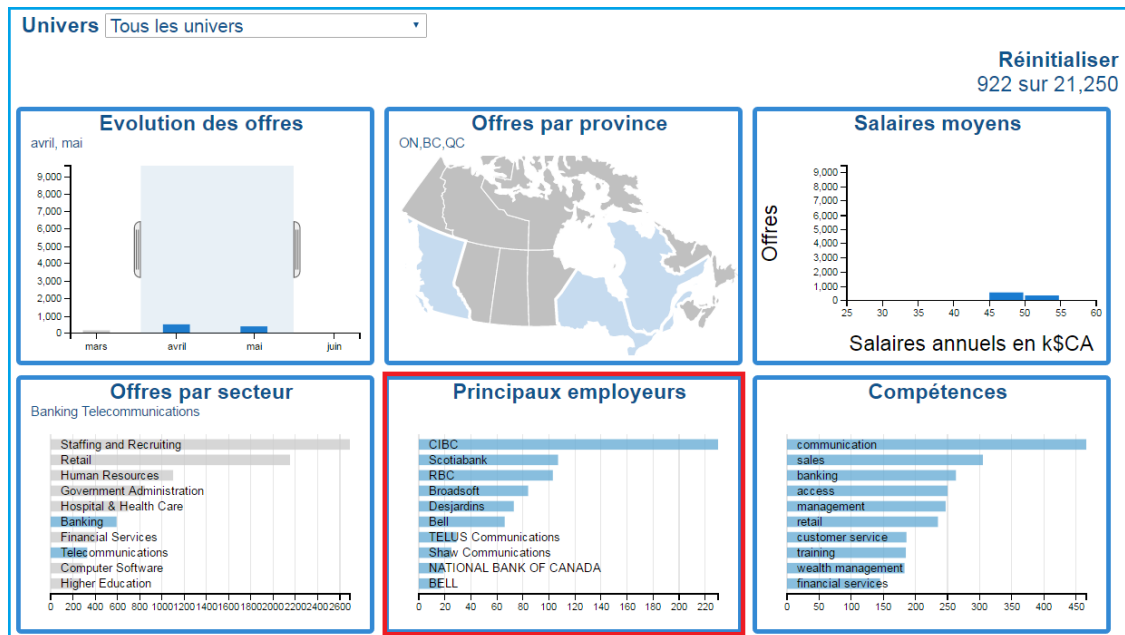


Figure 3.5 – Contrôle du graphique Company

une vue d'ensemble du baromètre qui englobe les six graphiques ainsi décrits.

3.2 Choix du type des graphiques

Selon Zelazny [39], pour choisir le type de graphique pertinent, l'essentiel est d'abord d'avoir, en tant que concepteur, une idée claire du message à transmettre. Le message ainsi défini comportera toujours l'un des cinq types de comparaisons suivants :

- La décomposition (component) : pour montrer la taille de chaque fraction d'un total, cette fraction est exprimée en pourcentage.
- La position (item) : qui vise à comparer comment les éléments se classent les uns par rapport aux autres, s'ils sont à peu près égaux, ou bien l'un d'eux représente-il plus ou moins que les autres .
- L'évolution (time series) : où on ne s'y intéresse pas à la taille de chacune des parties, ni à leur classement, mais à la façon dont elles varient dans le temps, que

la tendance à travers le temps soit en hausse, en baisse ou stable.

- La répartition (frequency distribution) : cette comparaison montre combien d'éléments se répartissent dans chaque intervalle d'une série numérique continue.
- Ou la corrélation (correlation) : qui montre si une relation entre deux variables se comporte ou non comme on pourrait s'y attendre.

Chaque comparaison conduit, à son tour, à l'un des cinq types de graphiques : camembert, graphe à barres, graphe à colonnes, courbe ou points. La figure 3.6 montre les graphiques recommandés pour chaque comparaison.

Pour le baromètre, il est demandé d'illustrer l'évolution des offres d'emploi à travers le temps (par mois), les dix employeurs qui recrutent le plus, les secteurs qui présentent le plus d'offres (secteurs tous confondus par défaut, ou les secteurs appartenant au même univers dans le cas de la sélection d'un univers particulier), les compétences les plus demandées, la répartition des salaires et le nombre d'offres par province. Selon la classification décrite ci-haut, nous avons, au niveau du baromètre décrit en figure 3.8, trois graphiques à barres : *compétences*, *principaux employeurs*, et *offres par secteur* où il s'agit de classer les éléments ; un graphe à colonnes *évolution des offres* ; et un autre graphe à colonnes *salaires moyens* pour la répartition des salaires. Pour ce qui est de la répartition des *offres par province*, illustrée par la figure 3.7, il s'agit de données géographiques. Hardin et al. [16] recommandent l'utilisation de cartes géographiques (type qui ne fait pas partie des cinq types proposés par Zelazny) pour ce type de visualisation. Pour la présentation des données sur une carte, Harrower et Brewer [17] proposent plusieurs schémas de couleurs, dont le schéma de couleurs séquentielles utilisé pour la carte géographique du baromètre. Ce schéma est utilisé pour représenter les données qui impliquent un ordre et qui varient selon une échelle ordinale ou numérique ; Les couleurs claires sont associées aux valeurs faibles et les couleurs sombres sont associées



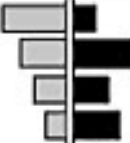


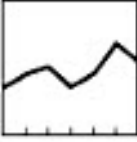
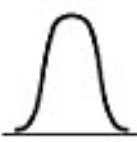

		KINDS OF COMPARISON				
		COMPONENT	ITEM	TIME SERIES	FREQUENCY	CORRELATION
BASIC CHART FORMS	PIE					
	BAR					
	COLUMN					
	LINE					
	DOT					

Figure 3.6 – Matrice du choix du type du graphique (Zelazny)

aux valeurs de données élevées.

La sélection d'un ou plusieurs éléments de données dans chacun des graphiques qui composent le baromètre, correspond en effet à l'application d'un filtre par rapport à ce champ de données, tout en gardant l'interactivité avec les autres graphiques. Ainsi, à chaque sélection, la liste des éléments sélectionnés est affichée au-dessus du graphique correspondant.

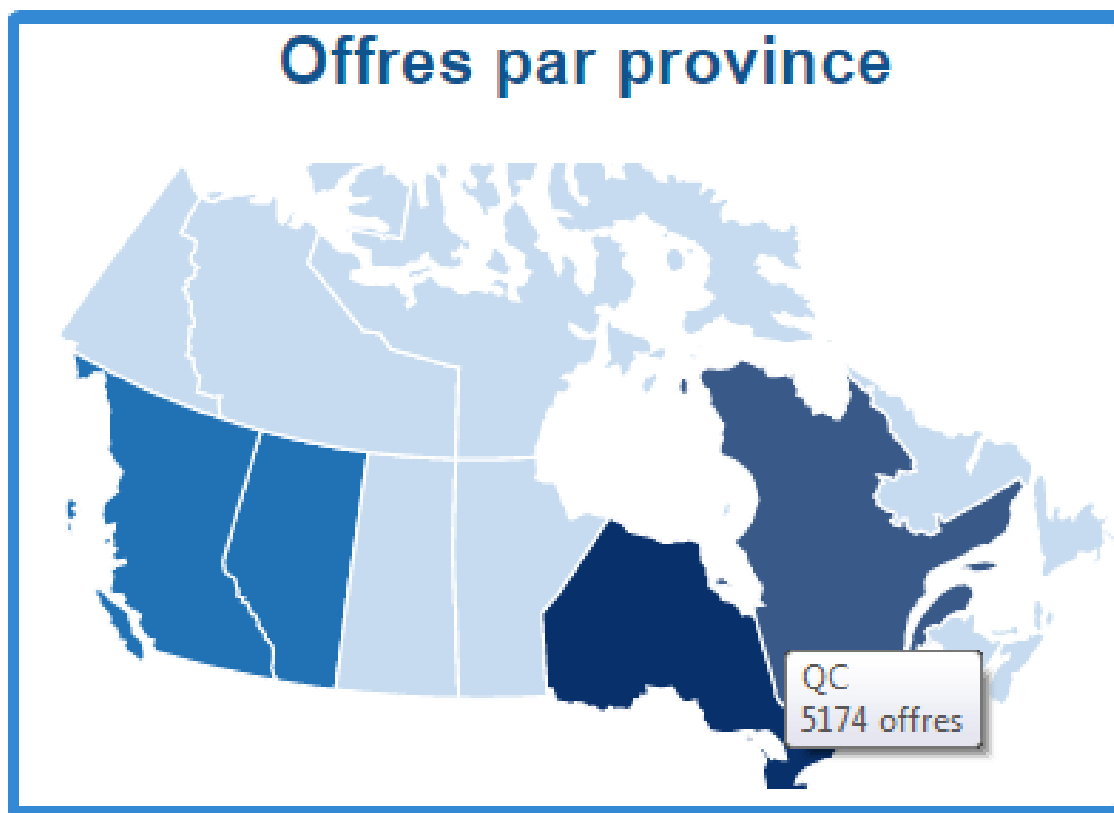


Figure 3.7 – Visualisation du nombre d’offres par province

A travers ce chapitre, nous avons présenté les motivations derrière le recours à la visualisation, la description des données utilisées et les choix faits quant à la nature des graphiques qui constituent le baromètre d’emploi. Actuellement nous nous sommes contentés de la visualisation des offres, sachant que les données concernant les candidats n’ont pas le même format que celles des offres, ce qui nécessite un effort supplémentaire pour les intégrer dans la même page du baromètre. Le choix de D3 pour réaliser ces visualisations est justifié par la puissance de cette librairie, la flexibilité qu’elle offre en termes de manipulation des données multidimensionnelles. Le chapitre suivant est consacré à l’extraction des compétences, qui font partie du baromètre, à partir d’une offre d’emploi.

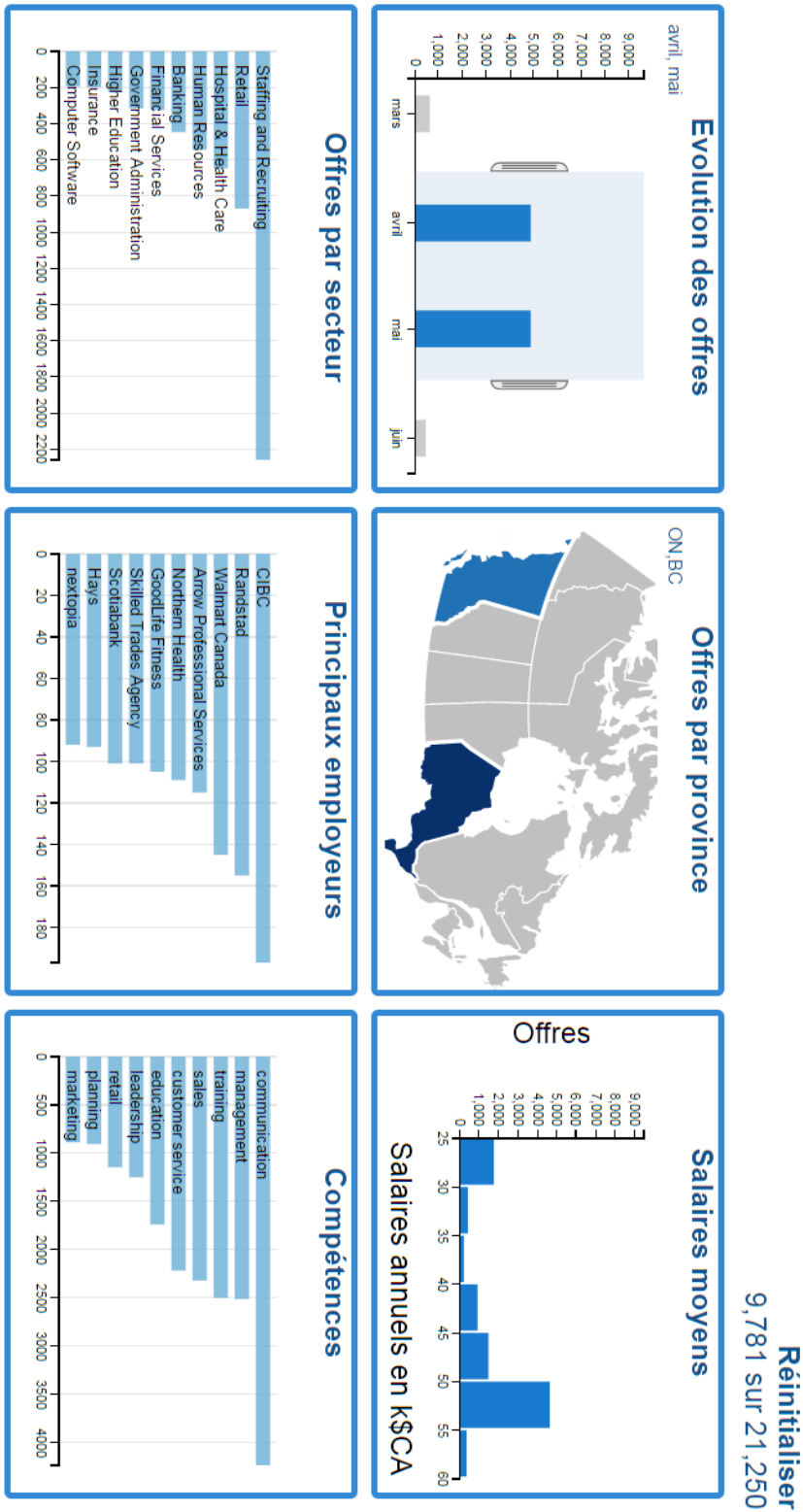


Figure 3.8 – Vue d’ensemble du baromètre avec sélection de mois et provinces

CHAPITRE 4

EXTRACTION DES COMPÉTENCES

L'objectif de cette partie est de décrire les étapes du processus de l'extraction des compétences, nécessaires à l'alimentation du baromètre de l'emploi qu'illustre la figure 3.8, à partir de la description d'une offre donnée. LinkedIn, le plus grand réseau social professionnel a introduit en 2012 la section *Compétences* [2]. Cette fonctionnalité permet aux membres de lister leurs compétences et leurs domaines d'expertise mais aussi de recommander ou avoir des recommandations de membres de leur réseau. D'autres travaux se sont penchés sur le même sujet [41], qui proposent une approche pour l'extraction de l'éducation, l'expérience et les compétences à partir d'un cv, puis l'attribution d'un score pour identifier les candidats pertinents. Kivimäki et al. [25] se basent sur une liste de compétences issues de LinkedIn ainsi que les articles et le graphe des liens hypertextes de Wikipédia, pour trouver les articles correspondants aux compétences de la requête. L'extraction des compétences dans le cadre de ce projet se fait dans un contexte différent. La tâche consiste à parcourir toutes les offres d'emploi et à y identifier les compétences pour permettre leur visualisation. Nous relatons dans ce qui suit la méthodologie suivie pour élaborer cette tâche.

4.1 Données

4.1.1 Description des données

Les données utilisées dans le cadre des expérimentations sont collectées par LBJ au format JSON. Il s'agit de 118 269 offres d'emploi (version collectée en février 2016), dont chacune est composée principalement des champs suivants : l'id de l'offre, la date de publication, la description, le titre du poste, la place et le nom de la compagnie. La

figure 4.1 illustre un exemple d'une offre collectée, et la figure 4.2 la même offre au format JSON.

Data Quality Analyst
SI Systems - Toronto, ON
Contract
3 Positions to Fill.

Responsibilities
Interpret Data Quality business requirements and identify technical deliverables
Develop technology solution/deliverables documentation for the Data Quality implementation
Drive discussions with technical teams to agree processes that deliver business Data Quality results
Develop relevant test cases and facilitate discussions with QA teams to develop QA activity planning.
Develop and maintain traceability documentation through the SDLC lifecycle

Skills
Excellent problem-solving skills
Familiarity with Risk Data reporting, Data Warehouse design/development environments
Communication/detailed documentation skills
Data Quality profiling
Data reconciliation Tools - mandatory
MS Office suite, including Excel macros
MS Visio Tools - desirable
SQL knowledge (basic/intermediate)
Cognos (basic/intermediate)
Specialization and Skills :
Business Analysis
Business Analyst

Additional Requirements :
None
Work Environment :

Figure 4.1 – Exemple d'une offre d'emploi

Dans cette partie, nous allons nous focaliser sur le champ description, où sont spécifiées les compétences requises pour le poste en question. On distingue deux types de compétences : les *soft skills* et les *hard skills*. Les *hard skills* comprennent les connaissances spécifiques requises pour réussir dans un emploi. Elles sont acquises lors d'une

```

{
  "_id" : { "$oid" : "5556891903be984ac58027ae" },
  "url_secondary" : "http://ca.indeed.com/XXXXXX",
  "date_update" : "15/05/15 20:02",
  "description" : "3 Positions to Fill. <br> Responsibilities <br> <br> Interpret Data
Quality business requirements and identify technical <br> deliverables <br> <br>
Develop technology solution/deliverables documentation for the Data Quality <br>
implementation <br> <br> Drive discussions with technical teams to agree processes
that deliver business <br> Data Quality results <br> <br> Develop relevant test cases
and facilitate discussions with QA teams to develop <br> QA activity planning. <br>
<br> Develop and maintain traceability documentation through the SDLC lifecycle
<br> Skills <br> <br> Excellent problem-solving skills <br> <br> Familiarity with
Risk Data reporting, Data Warehouse design/development <br> environments <br>
<br> Communication/detailed documentation skills <br> <br> Data Quality profiling
<br> <br> Data reconciliation <br> Tools - mandatory <br> <br> MS Office suite,
including Excel macros <br> <br> MS Visio <br> Tools - <br> desirable <br> <br>
SQL knowledge (basic/intermediate) <br> <br> Cognos (basic/intermediate) <br> <br>
<b>Specialization and Skills:</b> <br> Business Analysis <br> Business Analyst
<br> <br> <b>Additional Requirements:</b> <br> None",
  "title" : "Data Quality Analyst",
  "url" : "http://ca.indeed.com/job/Data-Quality-Analyst-at-XXXXXX-in-Toronto,-ON-
XXXXXX",
  "ref_external" : "2549add1e9b60c95",
  "place" : "Toronto, ON",
  "company_name" : "XXXXXX",
  "job_id" : "INDEED2549add1e9b60c95"
}

```

Figure 4.2 – Même offre au format JSON

expérience professionnelle ou via l'éducation et les formations. Des exemples de ces compétences comprennent les diplômes et les certificats obtenus, la maîtrise des langues étrangères, les langages de programmation. Elles peuvent être évaluées et mesurées et sont le plus souvent utilisées au cours du processus d'embauche pour comparer les compétences des candidats vis-à-vis d'une offre d'emploi. Les *soft skills*, par contre, sont des compétences plus subjectives et moins quantifiables que les *hard skills*. Il s'agit d'attributs personnels, traits de personnalité, indices sociaux inhérents, ou des capacités de communication nécessaires pour réussir dans un travail. Des qualités telles que la patience, le travail en équipe, la résolution de problèmes et la communication font partie

de cette catégorie de compétences.

Les *soft skills* sont des compétences exigées par les employeurs indépendamment du secteur d'activité. On les qualifie également de « compétences transversales ». Il est donc normal qu'elles soient citées plus que les *hard skills*. La figure 4.3¹ illustre la loi de Zipf appliquée aux compétences du corpus, qui consiste à classer les mots d'un texte par fréquence d'occurrence en ordre décroissant (rang) et à représenter cette distribution par la courbe des fréquences en fonction du rang. Nous constatons que des compétences comme *communication*, *problem solving*, *interpersonal skills*, qui sont des *soft skills*, se situent parmi les compétences les plus fréquentes, et à partir d'un certain rang (rangs supérieurs à celui de *risk management*), nous observons l'absence des *soft skills*. Dans l'exemple de la figure 4.1, nous trouvons à la fois des *soft skills* : *problem-solving*, *communication* ; et des *hard skills* : *risk data reporting*, *datawarehouse*, *detailed documentation*, *data quality profiling*, *data reconciliation*, *MS Office*, *MS Visio*, *SQL*, *Cognos*, *business analysis*.

Le travail de Bastian et al. [2], à l'origine de l'introduction de la section recommandations des compétences de LinkedIn, a eu comme objectif la création d'une « folksonomie » pour permettre aux membres de sélectionner les compétences à partir d'une liste normalisée, puis, la réalisation d'un système de recommandation pour leur proposer d'étendre leurs compétences. Cette folksonomie qui pourrait servir comme liste de référence de compétences a été retirée par LinkedIn. Nous nous servons donc d'une liste moins exhaustive de 6000 compétences, également issue de LinkedIn, disponible en ligne ², dont une partie est listée à la figure 4.4. Afin de mesurer l'exhaustivité de cette liste, nous avons pris un ensemble de 100 offres, étiqueté d'abord sur la base des 6000 compétences. Un script développé à cet effet (utilisé également pour extraire les compétences représentées dans la figure 4.3) transforme la liste des compétences en un arbre,

¹<https://gist.github.com/fnielsen/7102991#file-brownzipf-py>

²<http://free-db.com/download/4/list-of-skills-from-linkedin.html>

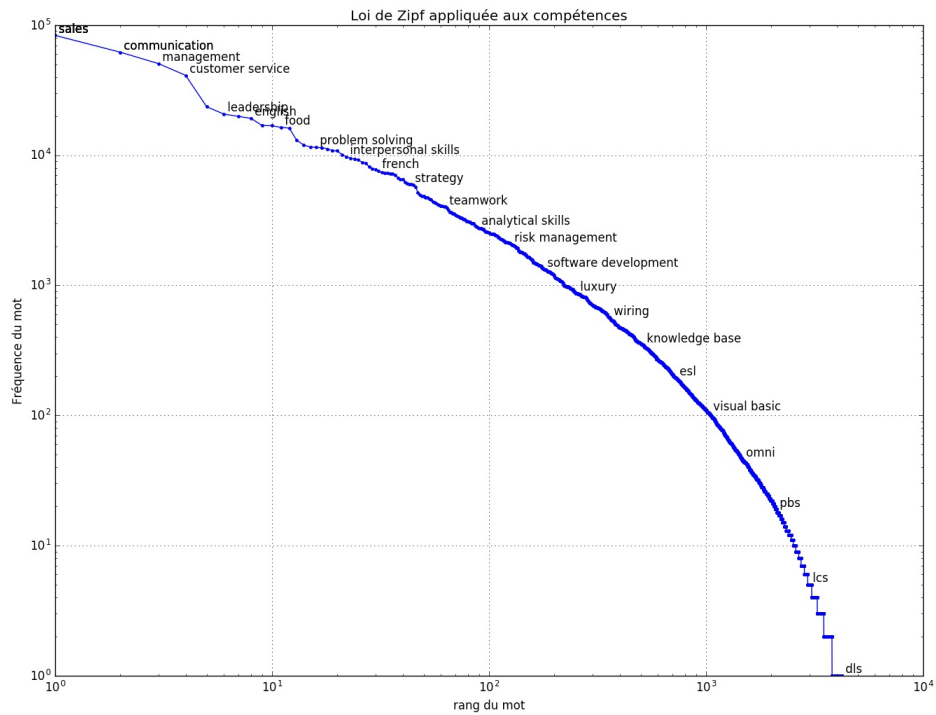


Figure 4.3 – Loi de Zipf appliquée aux compétences du corpus des offres

puis parcourt l’offre, une fois le début d’une compétence est trouvé, on cherche dans l’arbre construit la plus grande chaîne de caractères correspondante à celle de l’offre, pour s’assurer de l’étiquetage de la compétence en entier.

Puis, le même ensemble d’offres est annoté manuellement, c’est-à-dire pour chaque offre, nous annotons tous les termes que nous estimons être des compétences. Le tableau 4.I illustre le nombre de compétences annotées automatiquement sur la base des 6000 compétences, et celui obtenu suite à l’annotation manuelle.

	Annotation par script	Annotation manuelle
Nombre de compétences	432	608

Tableau 4.I – Compétences annotées

L'annotation manuelle a permis de passer de 432 à 608 compétences étiquetées, c'est-à-dire 40% de plus. Toutefois, la comparaison des deux annotations montre que 90 termes (environ 20%) ont été faussement annotés par le script comme étant des compétences. Dans cet extrait d'une description d'offre : *Full-time permanent employees receive a Health Care package which includes extended medical and dental insurance, short and long-term disability and Provincial Health Care insurance*, les termes *insurance* et *Health Care*, bien qu'ils fassent partie de la liste des compétences, réfèrent aux avantages sociaux offerts à l'employé, leur annotation dans ce contexte est inadéquate. Il est donc difficile de dresser une liste complète de toutes les compétences existantes. De plus, de nouvelles compétences apparaissent continuellement dans le monde du travail, et il est impossible de les recenser toutes. Plutôt que d'essayer d'avoir une liste exhaustive, il est préférable d'analyser la structure du texte et d'en cibler les termes qui sont des compétences.

4.1.2 Prétraitement des données

Avant de procéder aux expérimentations, il s'avère nécessaire de nettoyer le corpus de données. Ce processus se déroule comme suit :

1. Vu que la liste des 6000 skills contient des compétences uniquement en anglais, toutes les offres rédigées dans d'autres langues sont rejetées. Pour identifier la langue du texte, nous utilisons la librairie LangDetect³ ; LangDetect implémente un modèle de Bayes naïf pour classifier le texte selon la langue où il est écrit.
2. Il arrive qu'il y ait des descriptions d'offres dupliquées dans le corpus, ceci pour deux raisons principales : soit plusieurs occurrences de la même offre ont été récupérées lors de la collecte, soit il s'agit d'une compagnie qui cherche le même

³<https://pypi.python.org/pypi/langdetect>

aac
eoo/aa compliance
aaus scientific diver
aaahc
aar
abap
abaqus ableton live
substance abuse prevention
ab initio
abstract paintings
abap-oo
artistic abilities
abap web dynpro
study abroad programs
sickness absence management
atomic absorption
working abroad
numerical ability
account management
accounting
key account management
mergers and acquisitions
access
financial accounting
accounts payable
active directory
talent acquisition

Figure 4.4 – Extrait de la liste des 6000 compétences

profil dans différents endroits, la description dans ce cas est la même, c'est juste le champ *place* de l'offre au format JSON qui varie. Dans ce cas, une seule occurrence de la description est retenue.

4.1.3 Identification de la section des compétences

Une première approche consiste à parcourir la description de l'offre, et à étiqueter les termes susceptibles d'être une compétence par l'utilisation du script d'extraction décrit

à la section 4.1.1. Un exemple de sortie de ce script est illustré dans la figure 4.5.

```
<p> For over 40 years Winters has connected highly trained personnel with reputable employers in the engineering and construction, aerospace, automotive, electronic , manufacturing, pharmaceutical , supply chain industries . <br> Our busy client in Kitchener is looking for CNC Machinist. <br> Shift is 6 :00 am – 6 :30 pm Friday , Saturday and Sunday </p> <p> Responsibilities : </p> <ul> <li> Set up and operate 2D and 3D milling machines </li> <li> Will monitor feed and speed of machines during the machining process </li> <li> Run first offs and prototypes </li> <li> Use verniers , calipers and jigs </li> <li> Must be able to work on your own </li> </ul> <p> Qualifications : </p> <ul> <li> Must be proficient in CNC machining and have at least 5 years </li> <li> Have a strong knowledge of G & M Codes </li> <li> Knowledge of Fanuc controls </li> <li> Ability to work independently </li> <li> Flexibility to work overtime when required </li> <li> Have a positive attitude </li> </ul> <p> If you have the skills and experience that we are looking for to be successful in this role please respond to this job posting with this resume or fax to 519-578-0918 </p>
```

Figure 4.5 – Exemple de d'étiquetage de compétences

Le point faible de cette méthode est que l'étiquetage se fait à travers toute la description de l'offre. A titre d'exemple, dans l'offre de la figure 4.5, il s'agit d'une compagnie active dans plusieurs domaines : la construction, l'aéronautique, l'industrie automobile, l'électronique, l'industrie manufacturière, l'industrie pharmaceutique, le supply chain ; qui cherche un machiniste MOCN (Machine-outil à commande numérique, en anglais *computerized numerical control (CNC)*). CNC est la seule compétence requise dans l'offre mais étant donné que toutes les autres compétences (ou plus précisément les domaines d'expertise) sont présentes dans la liste de référence, elles sont alors toutes étiquetées bien qu'elles ne correspondent pas à ce qui est demandé, puisqu'elles sont citées dans le contexte de la présentation de la compagnie.

Ceci nous amène à explorer la distribution des compétences à travers le texte de l'offre afin de détecter une éventuelle tendance à citer les compétences dans une zone donnée du texte. A cette fin, le texte est divisé en plusieurs segments de taille identique en termes de nombre de mots (le nombre de segments est un paramètre qu'on ajuste) pour calculer la proportion des compétences présentes dans chaque segment selon le script

décrit plus haut. Le tableau 4.II illustre les proportions moyennes des compétences sur tous le corpus des offres.

Segments	Proportion moyenne des compétences par segment									
2	0.44	0.55								
3	0.27	0.35	0.36							
4	0.20	0.24	0.29	0.25						
5	0.16	0.17	0.21	0.24	0.19					
6	0.13	0.13	0.17	0.18	0.19	0.16				
7	0.11	0.11	0.13	0.15	0.16	0.16	0.13			
8	0.1	0.1	0.11	0.12	0.13	0.15	0.14	0.11		
9	0.09	0.08	0.09	0.1	0.11	0.12	0.13	0.12	0.1	
10	0.08	0.07	0.08	0.09	0.1	0.10	0.11	0.11	0.10	0.08

Tableau 4.II – Distribution des compétences dans une offre

Les résultats obtenus montrent que la distribution des compétences est quasi uniforme à travers le texte et que les variations de la proportion des compétences entre les segments sont légères en moyenne. Par conséquent, la position d'une phrase dans l'offre ne semble pas être un indicateur pertinent. Il faut donc repenser la méthodologie de segmentation, en explorant d'autres pistes plus efficaces. A mentionner que le champ description de la plupart des offres d'emploi est en format HTML, il est composé, en général, de plusieurs sections : la présentation de la compagnie, les responsabilités et les tâches relatives au poste, l'éducation exigée, les qualités et compétences recherchées, etc. La figure 4.6 illustre un exemple d'offre avec différentes sections. Notre approche consiste à extraire de toutes les offres les titres des sections en se servant des balises HTML. Un seuil est établi afin de retenir les titres les plus fréquents dans le corpus (un seuil de 100 est défini). Cette liste est ensuite filtrée manuellement pour ne garder que les titres des sections où les compétences requises sont citées, à savoir l'éducation et les compétences. Ces titres retenus sont détaillés dans l'annexe II. La deuxième étape de la segmentation consiste à définir la fin de la section. Ceci est réalisé en faisant appel à des

expressions régulières afin de détecter les balises qui mentionnent la fin de la section. La figure 4.6 illustre, en rouge, la portion de l'offre détectée par l'application de ce processus (*Qualifications* fait partie de la liste des titres tandis que *Join us* sert de marqueur de fin de section).

Five reasons to join XXXX
Be part of a team of seasoned professionals recognized for their extraordinary experience and expertise.
Advance and develop in a fun and stimulating workplace.
Reach your full potential in a firm that values the human element of doing business.
Work in a firm that actively promotes an entrepreneurial culture.
Take advantage of competitive benefits and exceptional working conditions.

Your role
Verifying large corporate tax returns and identifying planning opportunities ;
Researching tax issues.
Assist with corporate reorganization, estate and other tax planning ;
Preparation of steps planning memorandum ;
Very comfortable interpreting and applying tax legislation ;
Train and supervise junior team members ;

Your qualifications
CPA-CA, CGA, or CMA, lawyer, CICA In-Depth Tax Course or a Master of Tax completed ;
Minimum 3 years of experience in a tax manager role working with private business ;
Strong focus on tax planning
Have proven problem solving capabilities including ingenuity, analytical skills and creativity
Possess exceptional technical skills and a fine attention to detail ;
Computer proficient ; knowledge of MS Outlook, Word, Excel, TaxPrep, TaxnetPro/CCH or similar tax research software.
Strong interpersonal and organizational skills and the ability to organize, prepare and clearly present information to our clients and team
Must be able to handle multiple projects and meet tight deadlines ;
The ability to have fun and a commitment to providing exceptional client service

Join us
The human resources team thanks all applicants for their interest in the position.
Following the receipt of your application, a pre-selection based on the study of your qualifications will be performed. All candidates will be contacted.

Figure 4.6 – Exemple de structure d'une offre d'emploi

Il faut mentionner que le processus de segmentation n'aboutit pas tout le temps, il existe en effet des offres qui dérogent à la règle, soit parce qu'elles ne contiennent pas

de titres (absence de structure HTML) ou bien parce qu’il n’est pas possible de déterminer la fin de la section des compétences (toutes les offres n’ont pas la même structure HTML). Pour cela, l’étape de segmentation a largement réduit le nombre d’offres retenues, passant de 86 000 offres à environ 10 000 offres. Le tableau 4.III résume les phases de prétraitement.

Traitement	Nombre d’offres retenues
Initial	118 269
Offres en anglais retenues	97 829
Offres après suppression des duplications	86 001
Offres après segmentation	10 137

Tableau 4.III – Étapes de prétraitement des données

4.2 Approche basée sur l’apprentissage automatique

4.2.1 Motivations du choix des CRF

La tâche consiste à parcourir une offre d’emploi afin d’en déterminer les termes pouvant être qualifiés de compétences. Ceci peut être réalisé à l’aide de l’étiquetage de chaque mot du texte. Plusieurs techniques sont appliquées dans le domaine du traitement automatique des langues naturelles pour ce type de tâches. Dans le cadre de ce projet nous avons choisi comme modèle d’apprentissage les champs markoviens aléatoires (Conditional Random Fields ou CRF) et plus particulièrement les champs markoviens aléatoires linéaires (linear-chain CRF).

Les CRF sont des modèles probabilistes discriminatifs [20] introduits par Lafferty et al. [26] qui ont fait leur preuve dans de nombreux problèmes de prédiction structurée et se comparent habituellement bien à d’autres modèles. Ils ont donné de meilleurs résultats que des modèles tels que SVM et naive bayes pour la classification des dialogues en ligne ou live chat [24] puisque la séquence des mots est prise en compte. Les modèles

de Markov cachés (HMM) qui font aussi partie des modèles graphiques probabilistes nécessitent la définition d'une probabilité jointe $P(X, Y)$, ceci présente des limitations : non seulement la dimensionnalité de X est souvent très grande, mais les traits (features) ont des dépendances complexes, de sorte que la construction d'une distribution de probabilité sur les observations X et les étiquettes Y devient difficile (à moins d'une hypothèse d'indépendance). La modélisation des dépendances entre les entrées peuvent conduire à des modèles difficiles à résoudre, mais les ignorer peut conduire à une réduction des performances.

A l'inverse, les CRF, qui combinent les avantages de la classification et de la modélisation graphique, présentent une solution à ce problème en modélisant la distribution conditionnelle $P(X|Y)$ directement, ce qui leur donne la capacité de modéliser de façon compacte des données multivariées avec la possibilité de tirer parti d'un grand nombre de traits (features) d'entrée pour la prédiction [34].

Si X et Y sont deux vecteurs et $\theta = \{\theta_k\} \in R^k$ un vecteur de paramètres et $\{f_k(y, y', x_t)\}_{k=1}^K$ un ensemble de fonctions de traits de valeurs réelles. Alors les CRF définissent la distribution conditionnelle de la forme :

$$P(x|y) = \frac{1}{Z(X)} \prod_{t=1}^T \exp\left(\sum_{k=1}^K \theta_k f_k(y_t, y_{t-1}, x_t)\right)$$

Avec $Z(X)$ une fonction de normalisation :

$$Z(X) = \sum_y \prod_{t=1}^T \exp\left(\sum_{k=1}^K \theta_k f_k(y_t, y_{t-1}, x_t)\right)$$

La figure 4.7 est la représentation graphique d'un CRF linéaire où les X_i désignent la séquence des observations et les Y_i leurs étiquettes correspondantes.

L'application de ces modèles couvre plusieurs domaines, nous pouvons citer, à titre d'exemple, les travaux sur les entités nommées [28], l'étiquetage des rôles sémantiques

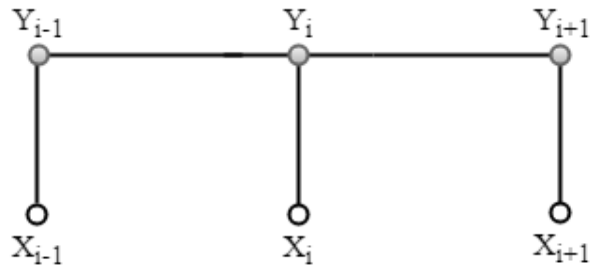


Figure 4.7 – Modèle graphique d’un CRF linéaire

dans un texte [31], l’alignement des mots dans la traduction automatique [3]. Dans le domaine de la bio-informatique, la prédiction de la structure des protéines [27], l’alignement des structures ARN [32]; ainsi que dans le domaine médical, pour la reconnaissance des symptômes dans les textes biomédicaux [19].

4.2.2 Protocole expérimental

L’annotation manuelle du corpus d’entraînement s’avère une tâche coûteuse quand il s’agit de données de grande taille, nous allons donc procéder par annotation automatique, en nous basant sur la liste de référence des compétences décrite à la section 4.1.1. Le script, dont le fonctionnement est décrit à la même section, est utilisé pour faire l’annotation des descriptions des offres. Il s’agit d’étiqueter les descriptions où les X_i correspondent aux mots de la description, les Y_i leurs étiquettes correspondantes. Ces dernières prennent l’une des valeurs suivantes :

- $Y_i = BS$: pour désigner le premier terme d’une compétence.
- $Y_i = IS$: pour chacun des autres termes de la compétence dont l’étiquette est différente de BS.
- $Y_i = O$: pour tout autre terme qui n’est pas une compétence.

Afin de mesurer le pouvoir de généralisation du modèle, autrement dit, sa capacité de prédiction avec un ensemble d’entraînement partiellement étiqueté, nous divisons la

liste des 6000 compétences en deux sous-ensembles de taille identique :

- Le premier sous-ensemble étant celui de référence (**compétences vues**), qui est utilisé pour annoter le corpus d’entraînement.
- Le deuxième sous-ensemble contient la liste des compétences que le modèle devrait identifier (**compétences à trouver**), c’est sur la base de cet ensemble que nous allons mesurer la capacité de généralisation du CRF.

Ainsi, nous allons utiliser trois ensembles de test, il s’agit des mêmes données étiquetées avec trois listes de référence différentes : le premier est étiqueté par la liste **compétences vues**, ce qui permet de comparer les performances du modèle par rapport à une recherche simple des termes dans le document. Le deuxième est étiqueté par la liste complète des **6000 compétences**, ce qui correspond à simuler le cas réel (des compétences pourraient être vues lors de l’entraînement, toutefois il reste toujours à identifier de nouvelles non existantes dans la référence). Enfin, un ensemble étiqueté par la liste **compétences non vues** ce qui permet de mesurer le pouvoir de généralisation du modèle.

L’ensemble de test étiqueté avec compétences vues est appelé **Test vu**, celui étiqueté avec la liste complète des 6000 compétences est désigné par **Test total**. Enfin, l’étiquetage du même ensemble de test avec compétences non vues donne **Test à trouver**. Le tableau 4.IV détaille la taille de l’ensemble d’entraînement ainsi que les trois ensembles de test avec les trois annotations, à savoir le nombre de tokens et de skills étiquetés.

	Entraînement	Test vu	Test total	Test à trouver
Nombre de mots	1 219 417	11 263	11 263	11 263
Compétences annotées	28 178	267	432	189

Tableau 4.IV – Nombre de mots et de compétences

L'exemple de la figure 4.8 montre l'annotation d'une offre de l'ensemble de test avec les 6000 compétences. Les compétences en rouge font partie du premier sous ensemble de compétence (**compétences vues**) utilisé lors de l'entraînement ; celles en bleu sont les compétences du sous ensemble **compétences à trouver**, non vues lors de l'entraînement, que le modèle CRF doit identifier en plus de celles vues lors de l'entraînement. En plus, il existe des compétences que nous n'avons pas dans la liste de référence des 6000 compétences (annotées en vert) qui ne sont pas prises en compte lors de la mesure des performances du CRF. Ce sont des compétences annotées manuellement. Une évaluation prenant en compte l'annotation manuelle est expliquée à la section 4.2.3.

Mesures de performances :

Pour mesurer les performances du modèle CRF, trois métriques sont utilisées : la précision, le rappel et la F-mesure. La précision est le ratio des entités pertinentes extraites sur toutes les entités extraites. Le rappel est le ratio des entités pertinentes extraites sur les entités de la référence (entités pertinentes). La F-mesure est une moyenne pondérée harmonique de la précision et du rappel. Si nous nous référons au tableau 4.V :

	Pertinent	Non pertinent
Trouvé	Vrai positif (tp)	Faux positif (fp)
Non trouvé	Faux négatif (fn)	Vrai négatif (tn)

Tableau 4.V – Matrice de confusion

Alors :

$$\text{Précision} = P(\text{pertinent} \mid \text{trouvé}) = \frac{tp}{(tp + fp)}$$

$$\text{Rappel} = P(\text{trouvé} \mid \text{pertinent}) = \frac{tp}{(tp + fn)}$$

$$\text{F-mesure} = \frac{(\beta^2 + 1) * \text{Précision} * \text{Rappel}}{(\beta^2 * \text{Précision}) + \text{Rappel}} = \frac{2 * \text{Précision} * \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

($\beta = 1$ pour une moyenne pondérée harmonique, où on accorde le même poids à la précision et au rappel)

ESSENTIAL QUALIFICATIONS

To be up to the exciting challenges of this Project Manager role, you must have a professional profile that includes :

- University degree in a relevant discipline such as **Architecture, Engineering** (Mechanical, Electrical or Building Sciences)
- 5 to 10 years of **project management** experience in **construction** and tenant set-up.
- **Project Management** Accreditation (preferred).
- Be comfortable working in a highly restricted environment.
- Strong **organizational, documentation, and project management** skills.
- **Self-motivation** and **proactive** approach.
- Effective **client management, interpersonal, and verbal and written communication** skills.
- Specific experience and skill related to the assignment – **churn, facility infrastructures, new construction, furniture, engineering, client industry** sector, **government** sector, etc.
- Knowledge of **construction management**, depending on assignment.
- Estimating and/or **forecasting** skills.
- **Problem-solving** and conflict management skills.
- Knowledge and experience with respect to industry standards and regulations.
- Software proficiencies depending on assignment, e.g., **Excel, AutoCAD, MS Project**, email, etc.
- Bilingualism (**English/French**)

Figure 4.8 – Annotation selon les ensembles de test

Extraction des traits :

Le bon choix des traits (le terme caractéristiques et l'anglicisme *features* sont également utilisés) est une étape importante lors de la conception des modèles d'apprentissage. Pour chaque terme de l'offre, des traits correspondants sont extraits, pour les expérimentations déroulées, nous avons choisi les familles de traits suivantes :

- Le mot lui-même est considéré comme trait lexical.
- Le Part of Speech (POS) : qui pourrait être traduit de plusieurs façons : nature

du mot, partie du discours, classe grammaticale, etc⁴. En anglais, on définit typiquement une dizaine de parties du discours à savoir : le nom, le verbe, l'adjectif, l'adverbe, le pronom, la préposition, la conjonction, l'interjection et l'article (ou déterminant). La librairie NLTK ⁵ est utilisée pour extraire le POS de chaque terme.

- Les suffixes : dans nos expérimentations, nous considérons deux types de suffixes : les suffixes à deux caractères et les suffixes à trois caractères. Les suffixes caractérisent certaines compétences composées d'un seul mot comme *forecasting*, *coaching*, ou de plusieurs mots : *Mechanical engineering*, *Electrical engineering*.
- La longueur du mot : le nombre de caractères du mot.
- La casse du mot : le format du mot peut être soit minuscule, majuscule ou mixte s'il contient à la fois des minuscules et des majuscules. La longueur et le format des mots aident à distinguer certaines compétences exprimées par exemple sous forme d'abréviation.
- Liste des termes les plus fréquents avant et après : les compétences sont souvent citées conjointement avec d'autres termes qui peuvent se situer avant ou après. Il pourrait s'agir par exemple de qualificatifs comme *strong*, *excellent*, de conjonctions pour plusieurs compétences successives, de substantifs pour mentionner l'éducation ou l'expérience *degree*, *certification*, *background*, ou des tags HTML comme par exemple ** lors de la description d'une liste de compétences. Deux listes sont ainsi constituées, chacune contient les termes les plus fréquents cités avant et après les compétences dans le corpus des offres. Ce trait est binaire, il s'agit d'un flag mis à l'état vrai dans le cas de la présence d'un des termes des

⁴[https://fr.wikipedia.org/wiki/Nature_\(grammaire\)](https://fr.wikipedia.org/wiki/Nature_(grammaire))

⁵<http://www.nltk.org/>

deux listes au voisinage d'une compétence. La liste de ces termes est détaillée dans l'annexe III.

La figure 4.9 montre un extrait de la description d'une offre après la tokenization (nous utilisons NLTK), l'étiquetage des compétences et l'extraction de traits. Chaque ligne est constituée d'un terme de l'offre suivi des différents traits décrits ci-haut et son étiquetage en fin de ligne. Les lignes vides servent à séparer les offres les unes des autres.

```
<b> TAG 1 3 L N N O
Preferred VBN ed red 1 9 M N N O
Qualifications NNS ns ons 1 14 M N N O
</b> TAG 1 4 L N N O
<br> TAG 1 4 L N N O
<ul> TAG 1 4 L N N O
<li> TAG 1 4 L N N O
Completion NNP on ion 2 10 M N N O
of IN of of 2 2 L N N O
a DT a a 2 1 L N N O
Certificate NNP te ate 2 11 M N N O
in IN in in 2 2 L N N O
Survey NNP ey vey 2 6 M N N BS
or CC or or 2 2 L N N O
Civil NNP il vil 3 5 M N N BS
Engineering NNP ng ing 3 11 M N N IS
Technology NNP gy ogy 3 10 M N N O
from IN om rom 3 4 L N N O
a DT a a 3 1 L N N O
</li> TAG 4 5 L N N O
</ul> TAG 4 5 L N N O
recognized VBD ed zed 4 10 L N N O
Technical NNP al cal 4 9 M N N O
Institute NNP te ute 4 9 M N N O
```

Figure 4.9 – Format des données préparées pour l'entraînement du CRF

Pour chacune des familles de traits citées précédemment, nous considérons des traits relatifs aux termes voisins ainsi qu’aux deux termes précédents. A titre d’exemple, pour les mots, nous prenons le mot courant ainsi que les deux mots précédents comme traits. Les abréviations suivantes sont utilisées pour désigner chacune des familles de traits :

- w (word) : pour les mots ;
- pos (Part of Speech) : pour la classe grammaticale ;
- sfx2 (suffix) : pour les suffixes à deux caractères ;
- sfx3 (suffix) : pour les suffixes à trois caractères ;
- len (length) : pour la longueur du mot ;
- case : si le mot est en minuscules, majuscules ou mixtes ;
- before : pour les termes avant la compétence ;
- after : pour les termes après la compétence.

L’entraînement des modèles est réalisé grâce au toolkit crfsuite⁶ qui implémente les CRF linéaires (first-order Markov) en utilisant l’algorithme d’optimisation L-BFGS ou limited-memory BFGS [29] qui est une variante de l’algorithme quasi newtonien BFGS. LBFSG est efficace dans le cas des problèmes d’optimisation avec un nombre élevé de variables, grâce à l’utilisation d’un espace limité de mémoire en ne considérant que les itérations récentes pour la construction de l’approximation de la hessienne [38]. Le plus simple modèle comprenant seulement le trait lexical (mot) génère 57 751 traits et l’entraînement converge au bout de 302 itérations ; la combinaison de toutes les familles de traits génère 131 176 traits et l’entraînement converge au bout de 468 itérations.

⁶<http://www.chokkan.org/software/crfsuite/>

4.2.3 Résultats

Nous allons, dans un premier temps, entraîner des modèles par famille de trait, dont les résultats sont illustrés dans le tableau 4.VI, afin de mesurer le pouvoir de prédiction de chaque famille de trait séparément, puis, sur la base des performances obtenues, nous choisissons parmi les combinaisons des traits celles qui sont optimales, dont les résultats sont détaillés dans le tableau 4.VII.

- **Par trait**

Trait	Test vu			Test total			Test à trouver		
	P	R	F	P	R	F	P	R	F
w	97%	87%	92%	95%	52%	67%	1%	0.5%	0.7%
pos	0%	0%	0%	0%	0%	0%	0%	0%	0%
case	0%	0%	0%	0%	0%	0%	0%	0%	0%
len	100%	2%	4%	100%	1%	0.3%	0%	0%	0%
sfx2	74%	32%	45%	74%	20%	32%	4%	1%	2%
sfx3	79%	49%	60%	83%	31%	45%	4%	3%	3%
before	59%	20%	30%	58%	17%	26%	62%	11%	19%
after	67%	24%	35%	69%	21%	33%	79%	18%	30%

Tableau 4.VI – Performances par trait

Nous constatons d’après les résultats du tableau 4.VI que le trait lexical (w) et les suffixes donnent les meilleures performances, et ce, pour l’ensemble de test étiqueté avec les compétences vues (**Test vu**) et avec la liste totale des compétences (**Test total**). C’est normal que la F-mesure relative au trait lexical (w) soit de 0.7% puisqu’on est dans le cas le plus difficile où il faut trouver les compétences non vues lors de l’entraînement. La prise en compte du Part of Speech, la casse, la longueur, à eux seuls donne une F-mesure très faible voire nulle quel que soit l’ensemble de test, à titre d’exemple, dans la liste des 6000 compétences, la longueur des compétences varie entre 2 et 49 caractères avec une moyenne de 10 caractères et un écart-type de 7 caractères. Ceci montre que le trait longueur du mot à lui

seul n'a pas la capacité discriminative permettant de distinguer une compétence des autres mots du texte de l'offre. Pour le test annoté avec la liste **compétences à trouver**, ce sont les familles *before*s et *after*s qui donnent la F-mesure la plus élevée étant donné que la liste des termes les plus fréquents avant et après une compétence est indépendante des trois annotations adoptées.

• **Combinaison des traits**

Combinaison	Test vu			Test total			Test à trouver		
	P	R	F	P	R	F	P	R	F
w	97%	87%	92%	95%	52%	67%	1%	0.5%	0.7%
+afters	98%	90%	94%	90%	60%	72%	11%	13%	12%
+before s	98%	91%	94%	91%	66%	77%	23%	26%	24%
+len	98%	92%	94%	91%	66%	76%	23%	25%	24%
+sfx2	98%	93%	95%	91%	66%	77%	23%	25%	24%
+sfx3	98%	93%	95%	92%	66%	77%	22%	25%	23%
+case	98%	93%	95%	91%	66%	76%	22%	25%	23%
+pos	98%	93%	95%	90%	66%	77%	23%	25%	24%
All \ w ⁷	75%	39%	51%	74%	35%	47%	77%	29%	42%

Tableau 4.VII – Performances des combinaisons des traits

Les métriques montrent que la différence de la capacité de prédiction des compétences d'une famille de trait à une autre, dépend de la liste des compétences utilisées en plus du trait lexical qui est le mot lui-même, nous constatons que la liste des termes et balises HTML les plus fréquents au voisinage de la compétence ont considérablement amélioré les performances et ce, indépendamment de l'ensemble de test pris. C'est pour cela nous avons considéré la dernière combinaison du tableau 4.VII qui ne prend pas en compte le trait mot (w), ce qui a permis d'améliorer la F-mesure en passant de 24% à 42% pour le troisième ensemble de test (**A trouver**). Cependant, la F-mesure est passée de 95% à 51% pour le pre-

⁷La dernière ligne du tableau 4.VII désigne la combinaison de tous les traits sauf le mot : **afters + before**s + **len** + **sfx2** + **sfx3** + **case** + **pos**

mier cas du **Test vu** et de 77% à 47% pour le deuxième cas **total**, c'est-à-dire qu'en l'absence d'une liste initiale de compétences, la dernière combinaison qui exclut le mot est le meilleur modèle, sinon, il faut toujours prendre le trait lexical. A noter aussi que la précision est plus élevée que le rappel dans le cas des deux ensembles de test (**vu et total**), c'est-à-dire que le nombre des faux positifs (termes faussement identifiés par le modèle comme étant des compétences) est inférieur aux faux négatifs (compétences que le modèle n'a pas trouvées).

- **Annotation manuelle de l'ensemble du test**

Afin d'évaluer la qualité de l'étiquetage réalisé par le script décrit dans la section 4.1.1 et l'exhaustivité de la liste des compétences, nous avons effectué l'annotation manuelle de l'ensemble de test utilisé auparavant, étiqueté avec la liste des 6000 compétences (**Test total**).

Le modèle qui donne les meilleures performances (F-mesure) d'après le tableau 4.VII (plus précisément pour l'ensemble **Test total**) est utilisé pour évaluer l'ensemble de test annoté manuellement. Le tableau 4.VIII illustre les résultats de cette évaluation.

Combinaison	Test total			Test annoté		
	P	R	F	P	R	F
All ⁸	90%	66%	77%	73%	37%	49%

Tableau 4.VIII – Performances de l'annotation manuelle

L'annotation manuelle a permis de passer de 432 à 608 skills étiquetés, c'est-à-dire 40% de plus. Certes, la liste de référence des skills utilisée ne couvre pas toutes les compétences, par exemple, *integration access policy* est étiqueté par le script comme trois skills distincts. Toutefois nous ne pouvons pas considérer ce point comme seul facteur

⁸All réfère à la combinaison de tous les traits : **w + afters + befores + len + sfx2 + sfx3 + case + pos**

pour interpréter cette variation de 40%. En effet, lors de l'annotation manuelle, nous avons remarqué plusieurs défis auxquels se heurte le processus de l'étiquetage automatique : d'abord, les différentes formulations utilisées par les recruteurs pour désigner les mêmes compétences : *excellent oral and written communication skills* vs *communicate effectively both verbally and in writing*. Le niveau de détail lors de la description d'une compétence : *management* vs *behaviour management, maintenance management, client management, etc.* Les compétences exprimées sous forme de phrases : *effectively manage statistical programming activities and integrate them with the entire clinical trial operations* ou encore *installing replacement windows and doors*.

Il faut noter toutefois que, la comparaison des deux ensembles de test montre que 90 termes (environ 20%) ont été faussement annotés par le script comme des skills. A titre d'exemple, le terme *Sound* qui fait partie de la liste des skills ne peut pas être considéré comme compétence dans le cas de *Sound working knowledge and experience with human resources*. Le contexte où le terme est cité est également à prendre en considération, par exemple dans *experience in partnering influencing coaching and building credibility with managers and human resources team*, le terme *human resources* réfère à une équipe ou un département et non pas à une compétence. Notre liste n'étant pas bruitée, ce qui corrobore le fait que la précision soit meilleure que le rappel, elle est néanmoins insuffisante pour garantir un étiquetage automatique complet. Ce qui explique le passage de la F-mesure de 77% à 49% après l'annotation manuelle. Enfin, il est important de souligner la difficulté de définir de manière précise les qualités humaines qu'on peut qualifier de *soft skill* dans la description d'une offre : *sense of initiative, high energy level, strong work ethic, ability to deal with others effectively, etc.*

A travers ce chapitre, nous avons relaté les étapes de l'extraction des compétences, à savoir le prétraitement des données des offres, puis les différentes approches pour aborder le problème de l'extraction dont les limites nous ont poussé à segmenter les

offres pour minimiser le bruit lors de l'annotation des compétences. Ensuite le recours aux CRF pour réaliser l'extraction. Les résultats de l'extraction montrent la capacité de notre modèle à trouver de nouvelles dont la F mesure est de 42%. Après l'annotation manuelle, les performances liées à l'ensemble de test annoté avec les 6000 compétences ont passé de 77% à 49%.

CHAPITRE 5

CONCLUSION

Le projet BPP réconcilie la recherche et l'industrie. Le volet visualisation de données a comme objectif l'implémentation d'un tableau de bord qui offre une vue globale sur les tendances du marché de l'emploi. Pour ce faire, nous avons considéré chacune des composantes d'une offre à savoir le nom de la compagnie, son secteur d'activité, sa localisation géographique, la date de publication de l'offre et les compétences requises pour ledit poste ; chacune de ces dernières est présentée sous forme de graphe qui fait partie du tableau de bord. Le choix de la librairie D3 est justifié par le fait qu'elle est libre de droits et qu'elle offre une large panoplie de visualisations mais aussi pour la diversité des librairies, comme dc.js et Crossfilter qui ont été développées pour tirer parti de la puissance de D3 en incorporant d'autres aspects comme l'interactivité et l'exploration multidimensionnelle des données. La visualisation montre par exemple que *recruiting and staffing* se place en tête des secteurs qui recrutent le plus, ceci est dû au fait que beaucoup de recruteurs restent anonymes et font appel aux services des agences de recrutement ou de placement dont le nom figure dans l'offre. L'autre défi a été de faire correspondre une société à son secteur d'activité : beaucoup de sociétés ne sont pas référencées où leurs noms sont formulés de diverses façons. L'extraction des salaires est difficile pour deux raisons : les salaires sont simplement cités dans 20% des offres et ne suivent pas un format standard.

Pour l'extraction des compétences, nous avons pu, grâce aux modèles de prédiction CRF, prédire 77% des compétences dans le cas d'un ensemble d'entraînement à moitié annoté, et 42% en l'absence d'une information (le mot comme trait) lexical préalable. L'annotation manuelle montre les contraintes de l'annotation automatique en l'occur-

rence la polysémie, le contexte de citation de la compétence, les différentes formulations de la même compétence. Ce qui impacte négativement la qualité de l'annotation des ensembles d'entraînement et par conséquent la qualité des modèles générés par le CRF.

Perspectives

Au niveau du tableau de bord, nous estimons important d'intégrer les fonctions (métiers) qui font partie de la hiérarchie de l'ontologie BPP développée durant ce projet qui lie un secteur à plusieurs métiers et ces derniers à leur tour aux compétences. Le tableau de bord peut être amélioré en ajoutant les profils des candidats afin d'avoir un aperçu sur les tendances de l'offre vis-à-vis de la demande.

Pour améliorer les modèles CRF, il serait intéressant d'avoir une liste plus exhaustive de référence ainsi nous pouvons exploiter la section compétences des profils issus de LinkedIn. L'inconvénient de cette option est que la saisie libre des compétences offertes rend la liste des compétences bruitée.

Enfin, vu que les compétences utilisées pour annoter les données d'entraînement de notre modèle sont incomplètes et que l'évolution des besoins du marché de l'emploi implique l'apparition de nouvelles compétences, l'apprentissage semi supervisé serait alors une piste à explorer, et plus particulièrement l'apprentissage actif (Active Learning) qui vise à optimiser la qualité de la prédiction en minimisant l'effort de l'étiquetage. Son principe, à l'inverse des modèles d'apprentissage supervisé qui utilisent les données complètement, consiste à choisir un ensemble initial de données étiqueté, puis par plusieurs itérations, faire une requête pour choisir à partir des instances non étiquetées celles considérées plus informatives qui sont ensuite ajoutées à l'ensemble d'entraînement.

BIBLIOGRAPHIE

- [1] Russell L. Ackoff. From data to wisdom. *Journal of applied systems analysis*, 16 (1):3–9, 1989.
- [2] Mathieu Bastian, Matthew Hayes, William Vaughan, Sam Shah, Peter Skomoroch, Hyungjin Kim, Sal Uryasev et Christopher Lloyd. LinkedIn skills : large-scale topic extraction and inference. Dans *Proceedings of the 8th ACM Conference on Recommender systems*, pages 1–8. ACM, 2014. ISBN 1-4503-2668-4.
- [3] Phil Blunsom et Trevor Cohn. Discriminative word alignment with conditional random fields. Dans *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 65–72. Association for Computational Linguistics, 2006.
- [4] Michael Bostock et Jeffrey Heer. Protovis : A graphical toolkit for visualization. *IEEE transactions on visualization and computer graphics*, 15(6):1121–1128, 2009.
- [5] Michael Bostock, Vadim Ogievetsky et Jeffrey Heer. D3 : data-driven documents. *IEEE transactions on visualization and computer graphics*, 17(12):2301–2309, 2011.
- [6] Stuart K. Card, Jock D. Mackinlay et Ben Shneiderman. *Readings in information visualization : using vision to think*. Morgan Kaufmann, 1999. URL https://books.google.ca/books?hl=en&lr=&id=w dh2gqWfQmgC&oi=fnd&pg=PR13&dq=Readings+in+information+visualization:+using+vision+to+think&ots=omELavvLPv&sig=bB_TfDD0zqcLtwSxEmP4uCqmW94.

- [7] ULRIKA WISS DAVID CARR. A cognitive classification framework for 3-dimensional information visualization. 1998.
- [8] Ed Huai-hsin Chi. A taxonomy of visualization techniques using the data state reference model. Dans *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on*, pages 69–75. IEEE, 2000. ISBN 0-7695-0804-9.
- [9] Ed Huai-hsin Chi et John T. Riedl. An operator interaction framework for visualization systems. Dans *Information Visualization, 1998. Proceedings. IEEE Symposium on*, pages 63–70. IEEE, 1998. ISBN 0-8186-9093-3.
- [10] Jeffrey Dean et Sanjay Ghemawat. MapReduce : simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.
- [11] Mamadou Alimou Dieng. Développement d’un système d’appariement pour l’embauchement. 2016.
- [12] Selan Rodrigues Dos Santos. *A framework for the visualization of multidimensional and multivariate data*. Thèse de doctorat, The University of Leeds, 2004.
- [13] Stephen Few et PERCEPTUAL EDGE. Data visualization : past, present, and future. *IBM Cognos Innovation Center*, 2007.
- [14] Michael Friendly et Daniel J. Denis. Milestones in the history of thematic cartography, statistical graphics, and data visualization. *URL <http://www.datavis.ca/milestones>*, 2001.
- [15] Philippe Grand’Maison. Génération automatique de lettres de recrutement. 2016.
- [16] Maila Hardin, Daniel Hom, Ross Perez et Lori Williams. Which chart or graph is right for you ? *Tell Impactful Stories with Data. Tableau Software*, 2012.

- [17] Mark Harrower et Cynthia A. Brewer. ColorBrewer.org : an online tool for selecting colour schemes for maps. *The Cartographic Journal*, 2013.
- [18] Jeffrey Heer, Stuart K. Card et James A. Landay. Prefuse : a toolkit for interactive information visualization. Dans *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 421–430. ACM, 2005. ISBN 1-58113-998-5.
- [19] Pierre Holat, Nadi Tomeh, Thierry Charnois, Delphine Battistelli, Marie-Christine Jaulent et Jean-Philippe Métivier. Fouille de motifs et CRF pour la reconnaissance de symptômes dans les textes biomédicaux.
- [20] A. Jordan. On discriminative vs. generative classifiers : A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14: 841, 2002. URL https://books.google.ca/books?hl=en&lr=&id=GbC8cqxGR7YC&oi=fnd&pg=PA841&dq=On+Discriminative+vs.+Generative+Classifiers&ots=ZwO7J3Zvx9&sig=5ydLGxZN_Zt1MecZCihm4Laf_vc.
- [21] Daniel A. Keim. Designing pixel-oriented visualization techniques : Theory and applications. *IEEE Transactions on visualization and computer graphics*, 6(1): 59–78, 2000.
- [22] Daniel A. Keim. Information visualization and visual data mining. *IEEE transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.
- [23] Rémy Kessler, Guy Lapalme et Eric Tondo. Génération d’une ontologie dans le domaine des ressources humaines.
- [24] Su Nam Kim, Lawrence Cavedon et Timothy Baldwin. Classifying dialogue acts in one-on-one live chats. Dans *Proceedings of the 2010 Conference on Empirical*

- Methods in Natural Language Processing*, pages 862–871. Association for Computational Linguistics, 2010.
- [25] Ilkka Kivimäki, Alexander Panchenko, Adrien Dessy, Dries Verdegem, Pascal Francq, Cédric Fairon, Hugues Bersini et Marco Saerens. A graph-based approach to skill extraction from text. *Graph-Based Methods for Natural Language Processing*, page 79, 2013.
- [26] John Lafferty, Andrew McCallum et Fernando Pereira. Conditional random fields : Probabilistic models for segmenting and labeling sequence data. Dans *Proceedings of the eighteenth international conference on machine learning, ICML*, volume 1, pages 282–289, 2001.
- [27] Yan Liu, Jaime Carbonell, Peter Weigle et Vanathi Gopalakrishnan. Protein fold recognition using segmentation conditional random fields (SCRFs). *Journal of Computational Biology*, 13(2):394–406, 2006.
- [28] Andrew McCallum. Efficiently inducing features of conditional random fields. Dans *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 403–410. Morgan Kaufmann Publishers Inc., 2002. ISBN 0-12-705664-5.
- [29] Jorge Nocedal. Updating quasi-Newton matrices with limited storage. *Mathematics of computation*, 35(151):773–782, 1980.
- [30] Jose F. Rodrigues, Agma JM Traina, Maria Cristina Ferreira de Oliveira et C. Traina. Reviewing data visualization : an analytical taxonomical study. Dans *Tenth International Conference on Information Visualisation (IV'06)*, pages 713–720. IEEE, 2006. ISBN 0-7695-2602-0.

- [31] Dan Roth et Wen-tau Yih. Integer linear programming inference for conditional random fields. Dans *Proceedings of the 22nd international conference on Machine learning*, pages 736–743. ACM, 2005. ISBN 1-59593-180-5.
- [32] Kengo Sato et Yasubumi Sakakibara. RNA secondary structural alignment with conditional random fields. *Bioinformatics*, 21(suppl 2):ii237–ii242, 2005.
- [33] Ben Shneiderman. The eyes have it : A task by data type taxonomy for information visualizations. Dans *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336–343. IEEE, 1996. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=545307.
- [34] Charles Sutton et Andrew McCallum. An introduction to conditional random fields. *arXiv preprint arXiv :1011.4088*, 2010.
- [35] Melanie Tory et Torsten Moller. Rethinking visualization : A high-level taxonomy. Dans *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, pages 151–158. IEEE, 2004. ISBN 1522-404X.
- [36] Edward R. Tufte et P. R. Graves-Morris. *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT, 1983.
- [37] Matthew O. Ward, Georges Grinstein et Daniel Keim. *Interactive data visualization : foundations, techniques, and applications*. CRC Press, 2010. ISBN 1-4398-6554-X.
- [38] Stephen Wright et Jorge Nocedal. Numerical optimization. *Springer Science*, 35: 67–68, 1999.
- [39] Gene Zelazny. *Say it with charts*. McGraw Hill, 2001. ISBN 0-7863-0894-X.

- [40] Nick Qi Zhu. *Data visualization with D3.js cookbook*. Packt Publishing Ltd, 2013. ISBN 1-78216-217-8.
- [41] Tim Zimmermann, Leo Kotschenreuther et Karsten Schmidt. Data-driven HR-Resumé Analysis Based on Natural Language Processing and Machine Learning. *arXiv preprint arXiv :1606.05611*, 2016.

Annexe I

Liste des univers

Id univers	Libellé
1	Administration publique, territoriale et internationale
2	Aéronautique et aérospatiale
3	Agriculture, viticulture, élevage, pêche
4	Agroalimentaire
5	Art, design, culture et artisanat d'art, musique, musée
6	Associations, syndicats, fondation, social, humanitaire, religions
7	Assurances, mutuelles, prévoyance
8	Audiovisuel, cinéma, spectacles, média, publicité, événementiel, divertissement, communication
9	Industries automobiles
10	Banque, finance, capital risque, fonds privés
11	Bois, papier, imprimerie
12	Chimie, caoutchouc, plastique
13	Conseil et services informatiques, édition de logiciels
14	Conseil stratégie et organisation, prestations intellectuelles pour les entreprises
15	Construction, Architecture, urbanisme, BTP
16	Cosmétique
17	Défense et armement, police, sécurité, transport de fonds
18	Digital, e-commerce, big data, jeux électronique
19	Édition, journalisme, presse

20	Énergie, eau, nucléaire, pétrole, gaz
21	Environnement, gestion des déchets
22	Équipements électriques et électroniques, composants, matériel informatique
23	Ferroviaire (matériel et équipements)
24	Formation initiale et continue, enseignement, éducation
25	Hôtellerie - restauration
26	Immobilier
27	Industries pharmaceutiques, biotechnologies, équipements médicaux
28	Industries manufacturières, mobiliers, textiles
29	Ingénierie - R&D
30	Juridique, droit et fiscalité
31	Matériel de construction
32	Matières premières, extraction, transformation, mines, métaux hors énergie
33	Métallurgie et mécanique, outils
34	Mode, luxe
35	Naval
36	Négoce B2B, distribution professionnelle, import-export
37	Retail, grand distribution, distribution généraliste et spécialisée
38	Santé, action sociale, hôpitaux, soins, bien-être
39	Services aux entreprises (Maintenance, entretien, sécurité, travail temporaire...)
40	Services aux particuliers (cours, ménage...)
41	Sports

42	Télécoms, hébergement, internet
43	Transports marchandises, logistique, stockage, emballage, conteneurs
44	Voyages, tourisme, loisirs, jeux d'argent
45	Ressources humaines
46	Informatique
47	Support (accueil, assistanat, services généraux, ...)
48	Comptabilité, gestion, audit

Annexe II

Titres utilisés pour la segmentation

Liste des titres retenus :

- abilities
- ability to
- about the role
- about you
- additional requirements
- applicant requirements
- application requirements
- are you ?
- are you an individual who
- as our ideal candidate
- assets
- attributes
- background
- basic
- basic/minimum qualifications
- basic qualifications
- candidate profile
- candidate requirements
- competencies
- core competencies
- credentials

- demonstrates the following
- desired skills and experience
- education and experience
- education required discipline
- employment requirements
- entry requirements
- required qualifications
- essential criteria
- essential qualifications
- essential skills
- experience
- experience/knowledge required
- experience required
- experience / skill
- experience, skills, academic
- here's what we need from you
- here's what we're looking for
- job experience
- job knowledge/qualifications
- job qualifications
- job-related experience
- job requirements
- job skills
- key competencies
- key qualifications
- key requirements

- knowledge
- knowledge and experience
- knowledge and skills
- mandatory
- mandatory qualifications
- minimum experience
- minimum qualifications
- minimum requirements
- must have
- must have skills
- needs to be
- needs to know
- other requirements
- our ideal candidate
- personal requirements
- position qualifications
- position requirements
- profile
- qualification requirements
- qualifications
- qualifications and experience
- qualifications and skills
- qualifications include
- qualifications required
- qualifications / requirements
- qualifications/skills

- required
- required competencies
- required education
- required experience
- required knowledge
- required language
- required qualifications
- required skills
- requirement note
- requirements
- requirements of the position
- requirements/qualifications
- skill requirements
- skills
- skills and abilities
- skills and experience
- skills and knowledge
- skills and qualifications
- skills/qualifications
- skills required
- specialization and skills
- specific skills
- successful candidates require
- supervisory experience
- the candidate
- the ideal candidate

- the ideal candidate must have
- the ideal candidate profile
- the ideal candidate will
- the ideal candidate will have
- we are looking for
- we require
- what we are looking for
- what you bring
- what you bring to this role
- what you need
- what you need to work with us
- who we want
- who you are.
- work experience
- you
- you are
- you have
- you have for us
- you must also have
- you must have
- you possess
- your experience includes
- your profile
- your qualifications
- your qualifications include

Annexe III

Mots fréquemment cités autour des compétences

Termes cités fréquemment avant une compétence :

- related
- prior
- senior
- solid
- work
- your
- within
- experience
- reliable
- using
- basic
- as
- oral
- including
- demonstrated
- superior
- and/or
- exceptional
- effective
- years
- proven

- verbal
- previous
- good
- written
-

- with
- or
- strong
- excellent
- of
-
- in
- and

Termes cités fréquemment après une compétence :

- degree
- knowledge
- abilities
- certification
- required
- role
- language
- preferred
- and/or
- in
- environment

- with
- industry
- to
- of
- is
-

- or
-
- experience
- skills
- and