

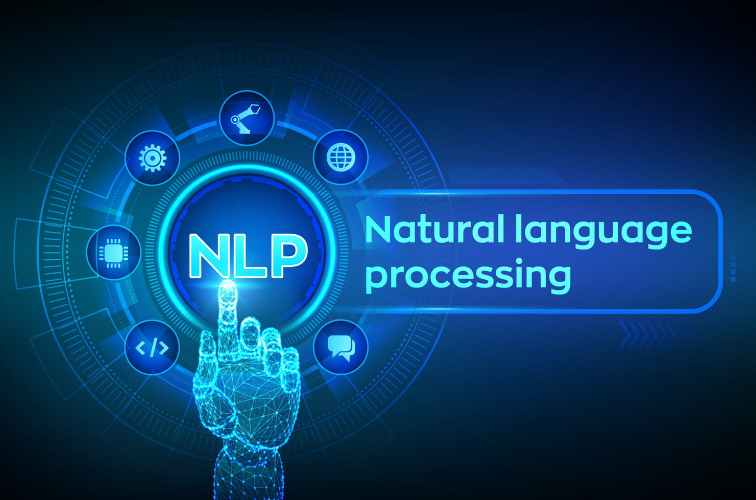
Projet - Traitement Automatique du Langage

**Fouilles d’offre d’emploi sur une plateforme dédiée - Seek**

------------------------------------------------

Floriane Mézirard & Lucie Raimbault

------------------------------------------------



M2 MAS

F. Moreau

2023-2024

Table des matières

[Table des figures 2](#_Toc149246228)

[Introduction 3](#_Toc149246229)

[I – Collecte des données 3](#_Toc149246230)

[II – Pré-Traitement 4](#_Toc149246231)

[III – Application de méthode de fouille 6](#_Toc149246232)

[Conclusion 7](#_Toc149246233)

# Table des figures

[Figure 1 : Nuage de mots description des postes 5](#_Toc149246197)

* Table des figures

*Consigne*

*Pour chaque étape du projet (choix de la problématique, des données, des traitements opérés et de la méthode de fouille appliquée), vous justifierez vos choix, et présenterez les principaux problèmes rencontrés et les solutions mises en œuvre pour les contourner. Enfin, vous pourrez en guise de conclusion donner votre ressenti sur l’intérêt des méthodes d’analyse textuelle dans le cadre d’un projet de fouille de données.*

# Introduction

La définition donnée par la CNIL sur le Traitement Automatique du Langage est la suivante :

« Le traitement automatique du langage naturel est un domaine multidisciplinaire impliquant la linguistique, l'informatique et l'intelligence artificielle.

Il vise à créer des outils capables d’interpréter et de synthétiser du texte pour diverses applications. »

En connaissant cela et en étant nous même en fin de cursus scolaire nous avons décidé de nous intéresser aux offres d’emplois. Pour cela nous avons choisi un site référençant les offres. Nous sommes dans un cursus basé autour de la data alors nous nous concentrerons sur les offres de ce milieu.

Le site référence sera « Seek ». Ce site fait partie des leaders du marché de l’emploi en ligne. Il couvre la zone géographique de l’Océanie et de l’Asie du Sud-Est essentiellement. Etant nous-même parties en Océanie l’an dernier pour notre année ce césure cela fait un petit clin d’œil.

Nous commencerons par collecter nos données en utilisant la méthode de Web Scraping. Nous poursuivrons par un pré-traitement des textes. Puis nous terminerons par utiliser différentes méthodes statistiques pour étudier et sortir des informations de nos données.

# I – Collecte des données

Nous avons décidé de collecter nous-même nos données. Pour cela nous avons choisi le seek. La première étape est de trouver une API afin de pouvoir récupérer les informations souhaitées. L’url du site web est la suivante : <https://www.seek.com.au/>. Après avoir dans la page que nous cherchons des offres dans le domaine de la data l’url devient : <https://www.seek.com.au/data-jobs>. Si nous souhaitons changer de domaine il suffit de modifier l’url comme voici https://www.seek.com.au/{domaine} -jobs.

En analysant cette nouvelle page, nous constatons plusieurs informations à récupérer. Premièrement il y a plusieurs pages d’offres d’emploi. Nous devons donc parcourir toutes ces pages. Deuxièmement pour chaque page nous devons collecter tous les liens pour les offres.

Une fois tous les liens pour chaque offre d’emplois récupérés nous pouvons scraper une page métier pour en collecter les informations voulues. Nous garderons les données suivantes :

* Intitulé du travail
* Entreprise
* Lieu de l’emploi
* Domaine
* Type de contrat
* Salaire (si présent dans l’annonce)
* Date de parution de l’offre
* Description de l’emploi

Si nous disposons de données sur l’entreprise nous récupèrerons également :

* Note moyenne de l’entreprise
* Lien sur les commentaires de l’entreprise

Une fois toutes ces données collectées pour chaque offre d’emplois nous créons un dataframe (jeu de données).

*???? Mettre un screen des premières lignes du jeu de données ????*

A partir de toutes ces informations nous allons pouvoir commencer à étudier notre jeu de données. Le travail principal consistera à analyser les descriptifs des offres.

Pour les entreprises dons nous disposons d’informations supplémentaires (notes, commentaires) nous les récupérerons dans un nouveau dataframe.

Après analyse des pages de commentaires sur les entreprises …

# II – Pré-Traitement

Avec nos deux bases de données récupérées nous commençons par un prétraitement. Nos données étant en anglais nous utiliserons la librairie nltk.

Concernant la base de données sur les offres d’emplois…..

Nuage de mot pour le corpus



Figure 1 : Nuage de mots description des postes

Concernant la base de données sur les avis des entreprises, nous avons lemmatisé les commentaires avantages et les inconvénients. Nous avons également concaténé les deux sous-catégories pour avoir un avis global sur l’entreprise. Nous appliquons le même traitement sur cette nouvelle colonne. Parfois des émojis sont présents il faut donc les supprimer. Nous supprimons également les mots vides et la ponctuation.

Nous avons ensuite créé une fonction permettant d’afficher un nuage de mots en fonction de l’entreprise. L’utilisateur n’a qu’à indiquer l’entreprise dont il souhaite avoir un affichage. Deux nuages de mots sortiront : un pour les avantages et un pour les inconvénients.

Voici un exemple pour l’entreprise ….

*Nuage de mots avantages*

*Nuage de mots inconvénients*

Suite à ces premières analyses nous allons étudier plus approfondissement nos données.

*Utiliser les méthodes traitements de textes*

* *Supprimer les mots vides, mettre en minuscules, suppression de la ponctuation ?*
* *Lemmatiser*
* *Extraction de certains types de mot – Ex : nom propre ?? A voir en fonction de ce qui est intéressant*
* *Faire des extraction bi-grams*

*Faire des graphiques de bases pour connaitre le jeu de données.*

*Ex : diagramme en barres sur le nom des métiers*

*Diagramme circulaire – répartition des offres en fonction de l’endroit*

*Entreprises pour lesquelles on dispose des notes*

*Faire un classement des meilleurs ?*

# III – Application de méthode de fouille

Lorsque l’on recherche un emploi nous regardons des offres d’emplois ayant des similarités ou étant les plus proches de nos critères. Il est donc intéressant de créer une similarité entre documents. L’utilisateur pourra inscrire ses critères et trouver l’offre (dans sa globalité : nom de l’entreprise, lieu, description, …) qu’il lui correspond le mieux. Nous avons donc réalisé cela.

*Avec des descriptions d’offres d’emplois, il est intéressant de les classifier en groupe ??. Redondance similarité mais possibilité d’avoir un dendrogramme + nombre de groupe potentiel --- complémentaire*

Il est également intéressant de connaitre le nom du poste si nous disposons d’une description. Nous appliquons donc une forêt aléatoire à nos données. Nos variables explicatives sont nos descriptions et la variable à prédire est le nom du poste.

Notre algorithme nous donne une accuracy de plus de 67% cela n’est pas si mal. Cependant nous disposons de plus de données avec comme nom de métier data analyst. L’algorithme aura tendance à prédire plus souvent ce métier. De plus nous avons fait le choix de nous consacrer sur le domaine de la data et les différences entre métiers ne sont pas forcément très clairs. Il y a des similarités ce qui rend plus complexe la bonne prédiction.

….

+ ajout sur les titres d’emplois

En étude de textes il est possible de connaitre les ressentis (plutôt positif ou négatif) d’un texte. Avec nos commentaires il est donc intéressant d’appliquer la méthode de polarité du sentiment (-1 pour négatif, 1 pour positif et 0 pour neutre). En étudiant l’avis global de l’utilisateur (avantage et inconvénients) nous obtiendrons un score que nous pourrons comparer à la note attribuée par l’individu à l’entreprise. Nous appliquons d’abord la polarité sur tous les avis. Nous obtenons un score compris entre -1 et 1. Il faut donc le modifier pour avoir une note comparable à celle mise par l’utilisateur qui est comprise entre 0 et 5. Une fois les scores sur la même échelle nous pouvons les comparer.

*Méthode TD-IFD – faire un moteur de recherche pour données l’offre la plus similaire avec les informations données au départ par l’utilisateur*

*Classifier les offres en groupes*

*Si suffisamment d’info de salaire – FA pour trouver les salaires manquants ???*

*Pour les commentaires – ressortir le/les sentiments qui ressortent en fonction de l’entreprise*

# Conclusion

Nous avons récupéré différentes offres d’emplois sur le site seek qui constituent notre base de données. Nous avons également des avis d’employés sur les entreprises. Après traitement et analyse des données, il y a beaucoup de choses à faire. Nous nous sommes concentrées sur la prédiction du nom du métier en cas de présence d’une description métier ou d’une note attribuée sur les entreprises par un employé. Il y a de nombreuses possibilités d’amélioration comme

* la récupération en temps réel des données,
* la création d’une application où l’utilisateur pourrait choisir tout ce dont il souhaite comme trouver un emploi en sélectionnant un lieu par exemple…