

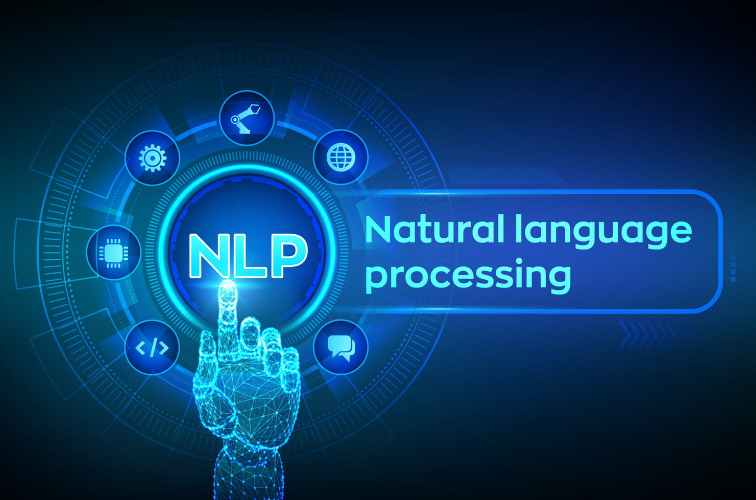
Projet - Traitement Automatique du Langage

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

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

Floriane Mézirard & Lucie Raimbault

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



M2 MAS

F. Moreau

2023-2024

Table des matières

[Introduction 3](#_Toc149162983)

[I – Collecte des données 3](#_Toc149162984)

[II – Pré-Traitement 4](#_Toc149162985)

[III – Application de méthode de fouille 5](#_Toc149162986)

[Conclusion 5](#_Toc149162987)

* 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
* Date de parution de l’offre
* Description de l’emploi

!!!!!!!!!!!!!!!!!!!!!!!! Je viens de voir que parfois il y a le salaire à ajouter !!!!!!!!!!!!!!!!!!!!!!

Ex avec cet url : [Data Entry Clerk Job in Waroona, Mandurah & Peel WA - SEEK](https://www.seek.com.au/job/71027142?type=promoted#sol=c537ffa3b1629ea11bf6be89b0b5a0eaa93d2d07)

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

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

Nuage de mot pour le corpus

Méthode TD-IFD (je ne sais plus ce que c’est ..) – 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