

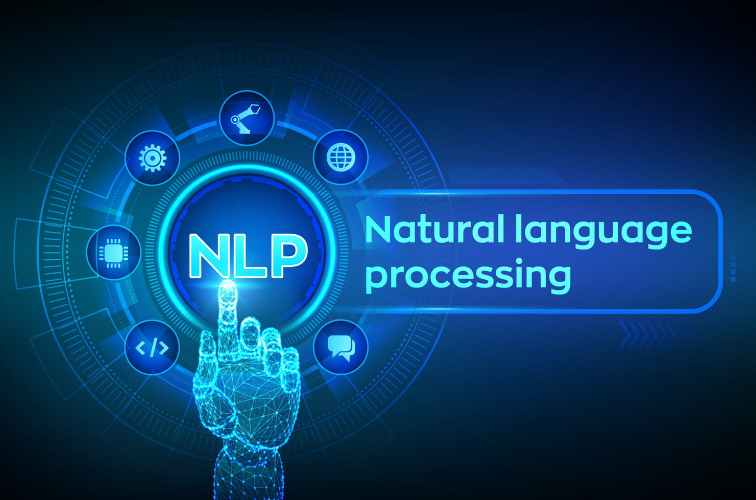
Projet - Traitement Automatique du Langage

**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 matières 2](#_Toc151555335)

[Table des figures 2](#_Toc151555336)

[Introduction 3](#_Toc151555337)

[I – Collecte des données 3](#_Toc151555338)

[II – Pré-Traitement et Méthode de fouilles de texte 4](#_Toc151555339)

[Conclusion 8](#_Toc151555340)

# Table des figures

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

*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 ». Il 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 de césure cela fait un petit clin d’œil.

Nous commencerons par collecter nos données en utilisant la méthode de Web Scraping et d’API. 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 site 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 inscrit 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 poste
* Nom de l’entreprise
* Lieu de l’emploi
* Domaine
* Type de contrat (Full Time, Part Time, Casual/Vacation, Contract/Temp)
* Salaire (si présent dans l’annonce)
* Date de parution de l’offre (par rapport au moment de la collecte d’informations)
* Description de l’emploi

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

* La note moyenne de l’entreprise
* Le 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). Nous avons en notre possession … emplois lorsque nous avons collecté les données le … .

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

Pour les entreprises dont nous disposons de liens vers les avis des employés, nous allons récupérer celle-ci grâce à l’API suivante :

[https://api-seek.prod.companyreview.co/companies/{identifiant\_entreprise}/company-reviews?page=1&sort=&api\_key=jwt\_prodSeekAuBrowserKey](https://api-seek.prod.companyreview.co/companies/%7bidentifiant_entreprise%7d/company-reviews?page=1&sort=&api_key=jwt_prodSeekAuBrowserKey)

Attention il faut bien entendu remplacer la partie {identifiant\_entreprise} par la valeur correspondante.

Suite à cela nous récupérons dans une autre base de données les informations suivantes :

* Nom de l’entreprise
* Identifiant de l’entreprise
* Avantages
* Inconvénients
* Job de la personne ayant posté l’avis
* Lieu de l’entreprise
* Note globale
* Date de parution de l’avis

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 notre première base de données et des avis (et notes) pour notre seconde base de données.

# II – Pré-Traitement et Méthode de fouilles de texte

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, nous avons lemmatiser nos descriptifs. Nous réalisons également un arrangement sur la colonne des noms de poste afin de pouvoir regarder quel emploi est le plus fréquent dans notre jeu de données. – Lucie si tu veux ajouter un truc sur la fonction que tu as crée. Nous obtenons le résultat suivant.

Tableau répartition

Dans notre base le métier de data analyst est prédominant. Ce métier est majoritairement recherché par les employeurs. Nous constatons également que le métier de data scientist est bien moins recherché.

Nous réalisons ensuite un nuage de mots sur les descriptifs d’emplois que nous possédons. Le résultat est le suivant. Peut etre à mettre le dernier crée car presence de the par exemple ici



Figure : Nuage de mots description des postes

Nous observons sans grande surprise que les mots les plus fréquents sont data et role. Pour une recherche d’emploi dans le domaine de la data cela semble plutôt rassurant. Nous constatons également que les termes d’experience, team, work, ou skill sont important dans les descriptions.

Nous pouvons également se focaliser sur une entreprise et un métier en particulier. Pour cela nous créons une autre fonction qui nous permet de sortir des nuages de mots en choisissant l’entreprise et le poste souhaité. Attention le métier doit être présent dans l’entreprise pour avoir un résultat positif.

Par exemple pour l’entreprise Ignite et le métier data analyst nous obtenons deux nuages de mots. En effet cette entreprise possède deux postes pour le métier de data analyst.

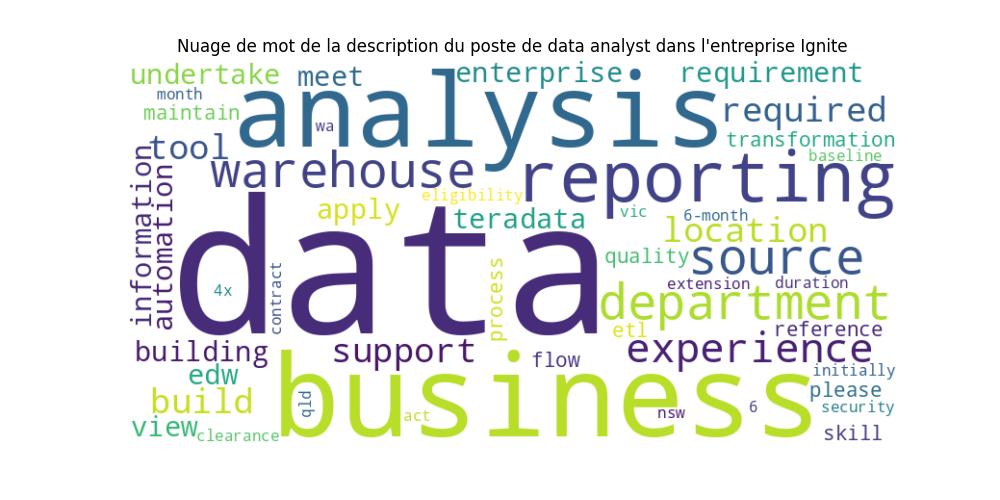


Figure : Nuage de mots de la description du poste de data analyst chez Ignite

Pour ce poste nous observons que les mots les plus fréquents sont data, analysis et business. Cela semble cohérent par rapport au poste.

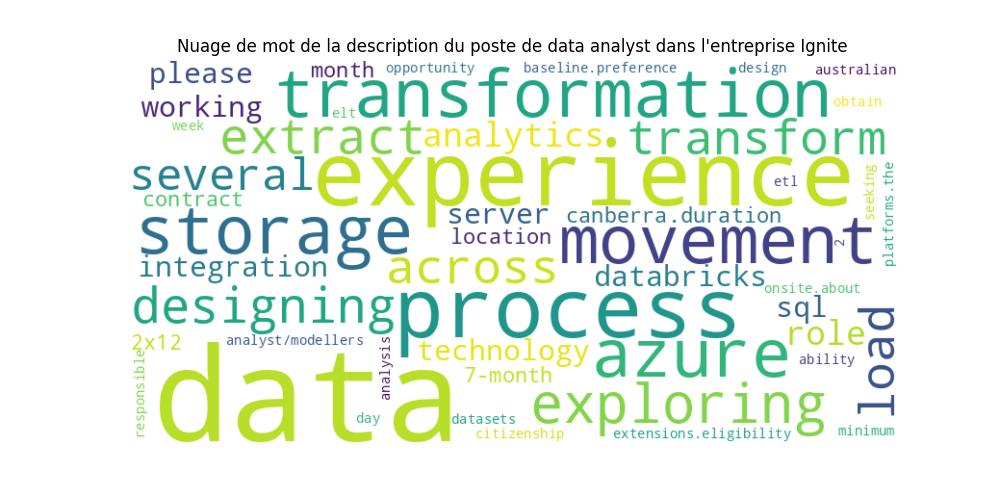


Figure : Nuage de mots de la description du poste de data analyst chez Ignite

Pour ce nouveau poste il n’y a pas de mots qui ressortent le plus qu’un autre. Les mots transformation, experience, process, exploring, movement, storage semblent tout de même prédominant après le mot data. Sur ce poste un profil ayant déjà une expérience dans le domaine semble attendu. Un junior ne sera potentiellement pas retenu.

Nous avons également créé une fonction permettant de rechercher l’offre d’emploi la plus similaire à un emploi que nous souhaiterions. L’utilisateur doit indiquer sa recherche et une sortie (nom du poste, entreprise, lieu, description, …) sera affichée pour indiquer l’offre d’emploi qui est la plus proche de ses envies.

Exemple de sortie ??? --- faire graphique de la CAH ? pour donner un nombre de groupe potentiel qui doit correspondre à notre nombre de nom de postes.

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

Nous nous sommes également intéressées au problème suivant : en étant en présence d’une annonce d’emploi quel serait le nom du poste associé ?

Nos variables explicatives sont nos descriptions et la variable à prédire est le nom du poste.

Nous avons ainsi fait de la classification de description d’emplois. Nous avons utilisé la méthode de Forêt aléatoire sur nos données. Notre modèle possède une accuracy (un taux de bonne prédiction) de plus de 67%. Cela semble raisonnable par rapport à nos données et à notre sujet. En effet les frontières des métiers de la data ne sont pas très tracées. Il y a des similarités entre les emplois. En étant data scientist nous pouvons être amené à faire des choses qu’un data engineer peut effectuer ou qu’un data analyst peut faire. De plus dans notre base nous disposons principalement d’exemple de data analyst donc le modèle aura tendance à prédire plutôt des métiers de data analyst.

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. Nous affichons un nuage de mots global, c’est-à-dire en prenant en compte les avis positifs et ceux négatifs.

Par exemple pour l’entreprise Police Credit Union nous obtenons le nuage de mots suivant :

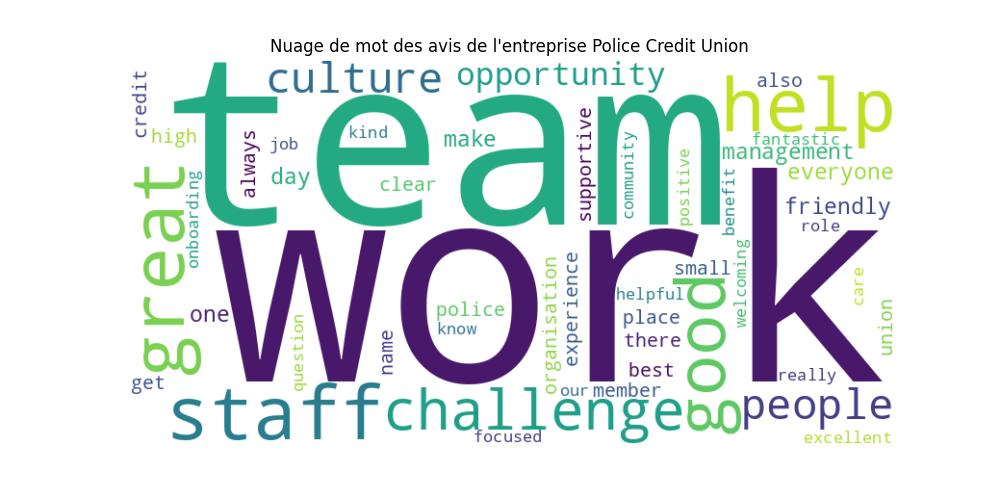


Figure 4: Nuage de mots des avis sur une entreprise

Nous observons clairement ici que les maîtres mots sont work et team. Nous pouvons penser qu’une personne aimant le travail d’équipe se plaira plutôt dans cette entreprise.

Nous avons également sorti deux nuages de mots pour une entreprise : un pour les avantages et un pour les inconvénients.

Voici un exemple pour l’entreprise Marathon Health.

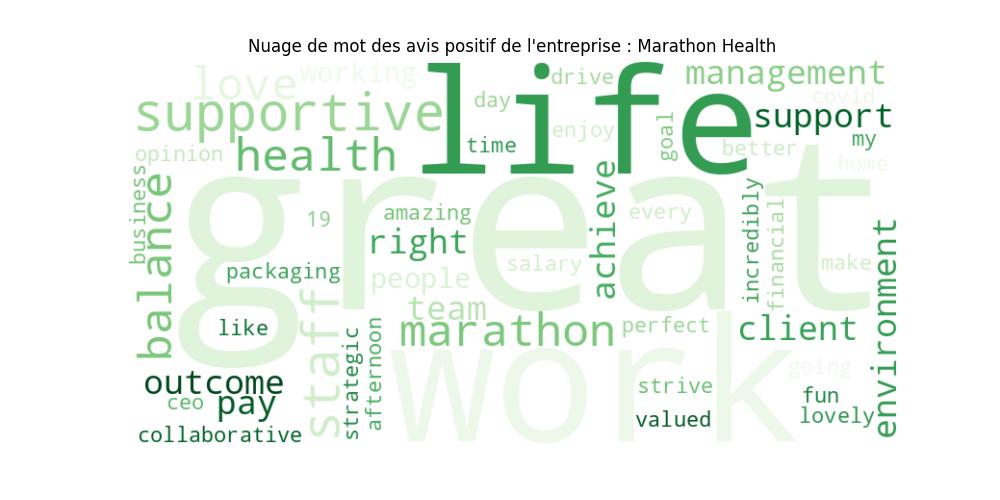


Figure : Nuage de mots des avis positifs pour Marathon Health



Figure : Nuage de mots des avis négatif pour Marathon Health

Nous observons ici des adjectifs plutôt positifs sur l’entreprise comme great, life, perfect, amazing. Concernant le nuage de mots des avis négatifs les mots revenants le plus sont management, work, challenge.

L’avis global sur l’entreprise semble tout de même relativement positif. La moyenne globale pour cette entreprise est de 3.6. Cela n’est pas très élevé mais plus que la moyenne qui est à 2.5. Notre ressenti semble cohérent.

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

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 ainsi 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.

Exemple sortie – tableau comparaison + petite analyse rapide en plus de ce qu’il y a dessous.

Nous observons que notre méthode a donné des notes similaires à celles données par les utilisateurs. Cela est très positif. En étant en présence de texte nous pouvons donc prédire la note qu’un utilisateur pourrait donner à une entreprise.

*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*

*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 ?*

# 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 et de possibilités. Tout dépend de ce que nous souhaitons savoir sur le sujet. Nous nous sommes concentrées sur la prédiction du nom du métier en cas de présence d’une description métier ou d’une prédiction de note attribuée sur les entreprises par un employé en présence d’avis sur l’entreprise. 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,
* Estimer le salaire d’un poste,