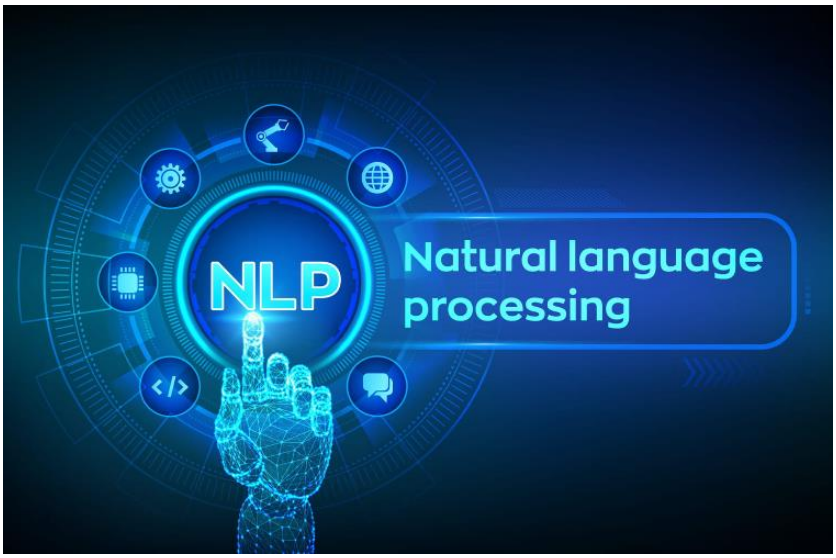
Projet - Traitement Automatique du Langage

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

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 interpréter 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. Ci-dessous voici une capture d'écran pour une offre.

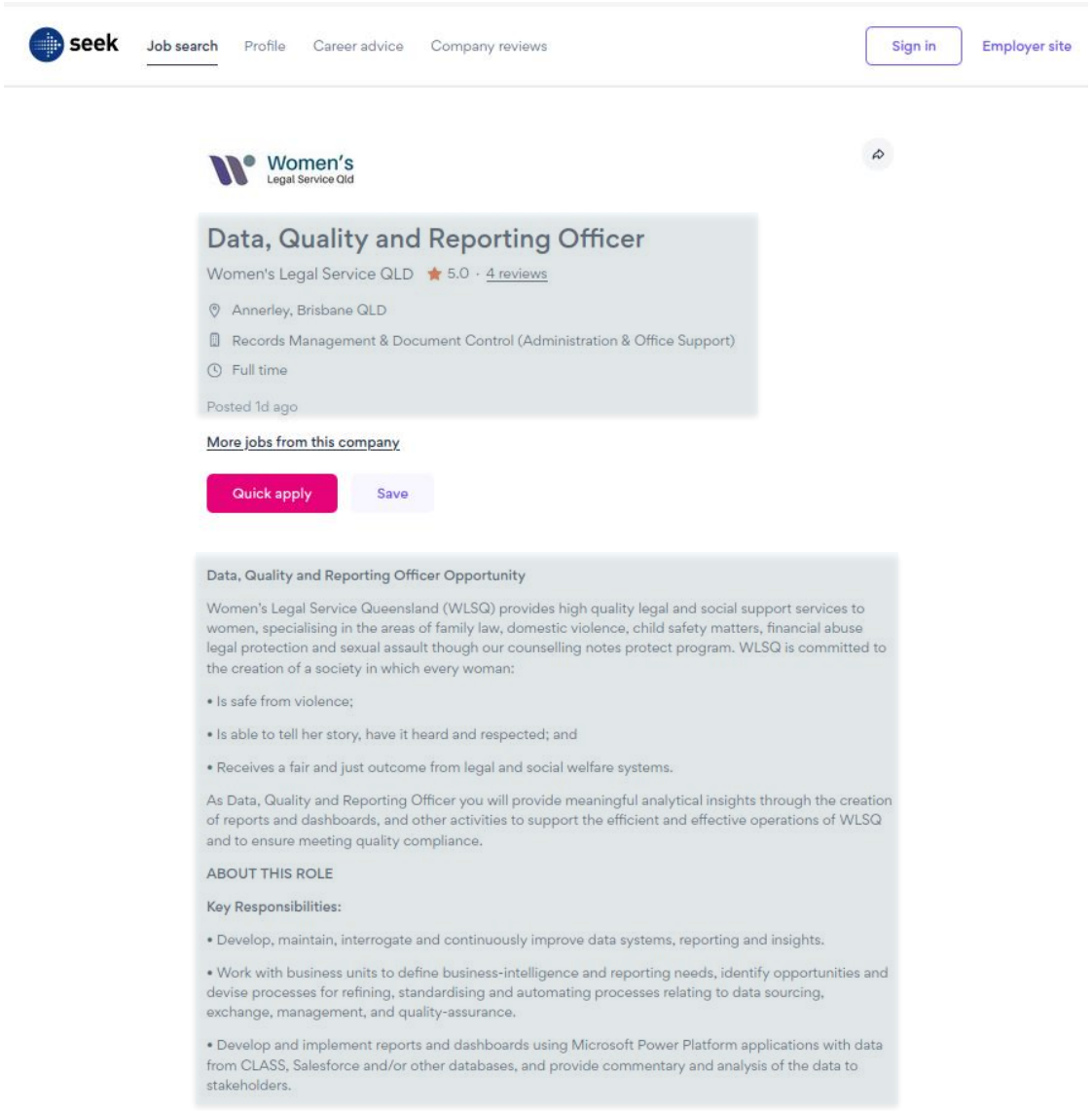


Figure 1: Capture d'écran Offre d'emploi Data - Seek

Nous garderons les données suivantes :

- Intitulé du poste
- Nom de l'entreprise
- Lieu de l'emploi
- Domaine
- Type de contract (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 récupérerons 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 154 emplois lorsque nous avons collecté les données le 26 octobre 2023. Ci-dessous un extrait des informations récupérées.

	A	B	C	D	E	F	G	H	I	J
1	titre	entreprise	lieu	categorie	contrat	salaire	parution	description	note	review
2	OT Data Support Engineer	Talison Lithium	Greenbushes, Bunbury	Other (Information & Com	Full time		Posted 19h ago	Operational Technology		
3	Campaign and Data Operations Manager	Melbourne Grammar School	South Yarra, Melbourn	Fundraising (Community S	Full time		Posted 15h ago	Campaign and Data Ope	3.9	/companies/melbourne-gramma
4	Data Analyst (Paid Internship)	Police Credit Union Ltd	Adelaide SA	Business/Systems Analysts	Part time		Posted 13h ago	A role you'll love, with t	4.0	/companies/police-credit-union-
5	Junior Data Analyst	Add Staff Recruitment Pty Ltd	Fortitude Valley, Brisb	Business/Systems Analysts	Full time	$55,000 – $60,000 per year	Posted 4d ago	Solar Service Guys (SSG		
6	Data Science Cadet	Jemena	Melbourne VIC	Mathematics, Statistics & I	Contract/Temp		Posted 1d ago	About UsJemena is an A	3.0	/companies/jemena-432985/rev
7	Head of Data Strategy, Analytics & Insights	AustralianSuper	Melbourne VIC	Other (Consulting & Strate	Full time		Posted 3d ago	AustralianSuper is on ar	4.0	/companies/australiansuper-813
8	Data Analyst, Modeller	Ignite	Canberra ACT	Other (Information & Com	Contract/Temp	AUD 125 - 140 per hour	Posted 3d ago	Location: Canberra,Dur		
9	Junior Data Entry	Energy Advance Australia Pty Ltd	Wangara, Perth WA	Data Entry & Word Proces	Casual/Vacation	$29 – $44 per hour	Posted 17h ago	POSITION - JUNIOR DA		
10	Solution Architect, Data	Paxus	Sydney NSW	Architects (Information & I	Contract/Temp	$1200 - $1400 p.d. inc. supe	Posted 14h ago	About the role Solution		
11	Data Analyst	Ignite	Sydney NSW	Other (Information & Com	Contract/Temp	AUD 100 - 115 per hour	Posted 2d ago	Location: ACT, NSW, QL		

Figure 2: Extrait de notre jeu de données - emploi

Pour les entreprises dont nous disposons d'un lien vers les avis des employés, nous allons récupérer d'autres informations grâce à l'API suivante :

https://api-seek.prod.companyreview.co/companies/{identifiant_entreprise}/company-reviews?page=1&sort=&api_key=jwt_prodSeekAuBrowserKey

Attention il faut bien entendu remplacer la partie {identifiant_entreprise} par la valeur correspondante.

Les avis sur le site de seek sont décrits en 2 parties, une partie avantage et une partie sur les défis, ainsi qu'une note que l'employé attribue à l'entreprise. Ci-dessous une capture d'écran pour l'entreprise Women's avec les avis et les notes des employés.

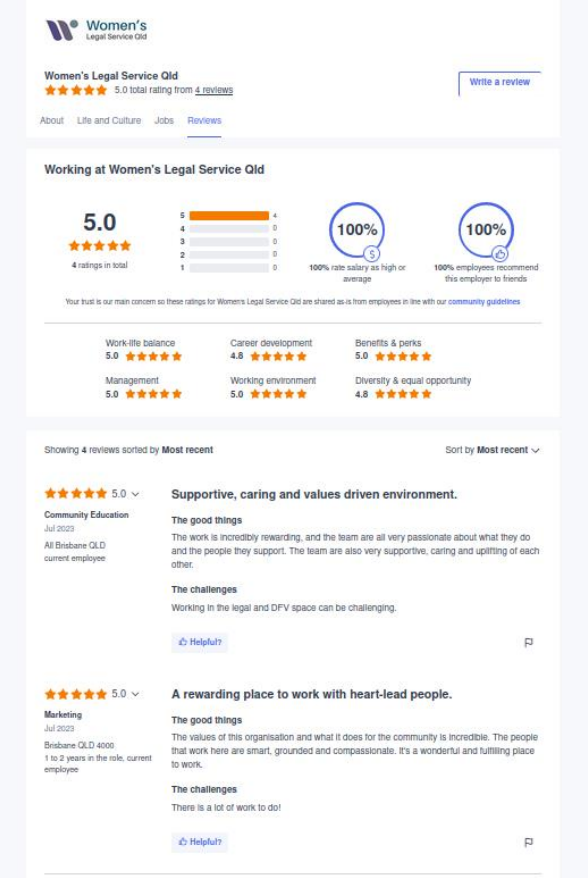


Figure 3: Capture d'écran Review - Seek

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

- Nom de l'entreprise
- Identifiant de l'entreprise
- Avantages
- Défis / Inconvénients
- Job de la personne ayant posté l'avis
- Lieu de l'entreprise
- Note
- 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 les avis (et notes) pour notre seconde base de données.

II – Pré-Traitement et Méthode de fouilles de texte

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, 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. Nous avons aussi créé une fonction qui nous permet de regrouper les noms de métier en des catégories plus larges, et avoir au final 8 catégories de métiers. Nous obtenons le résultat suivant.

Métiers	Nombre d'offres
Data Analyst	67
Data Manager	25
Data Engineer	21
Data Entry	18
Data Scientist	10
Data Architect	5
Business Analyst	5
Other Data Job	3

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. Cependant la différence entre ces métiers n'est pas si évidente.

Offre d'emploi

Nuage de mot

Nous réalisons ensuite un nuage de mots sur les descriptifs de tous les emplois que nous possédons dans notre base. Le résultat est le suivant.

Figure 4: Nuage de mots description des postes

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

Nous pouvons également nous 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. Si l'entreprise propose plusieurs mêmes postes, un wordcloud ressortira pour chaque annonce.

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.

Pour la première annonce, le résultat est ci-dessous.

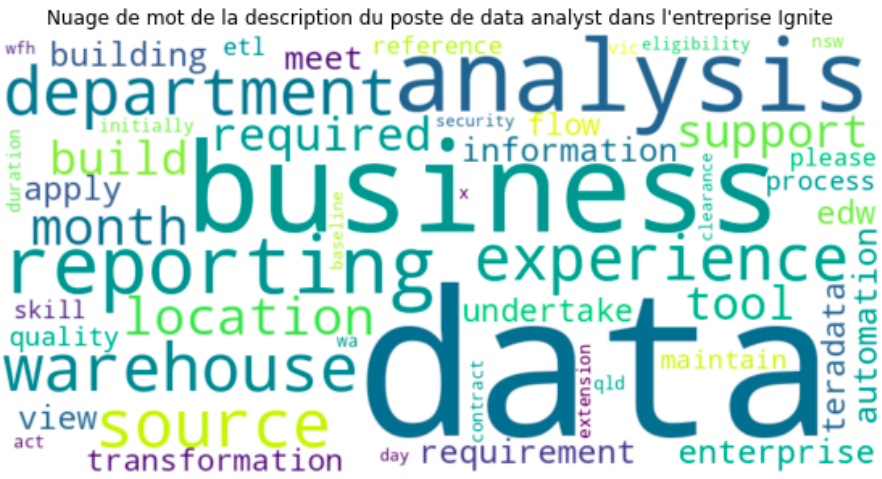


Figure 5: Nuage de mots de la description du poste de data analyst chez Ignite

Pour le premier poste, nous observons que les mots les plus fréquents pour la première annonce sont ‘data’, ‘business’ et ‘analysis’. Nous observons aussi que l’annonce précise les outils et logiciels comme ‘reporting’, ‘etl’.

Le nuage de mots pour la seconde annonce est le suivant.

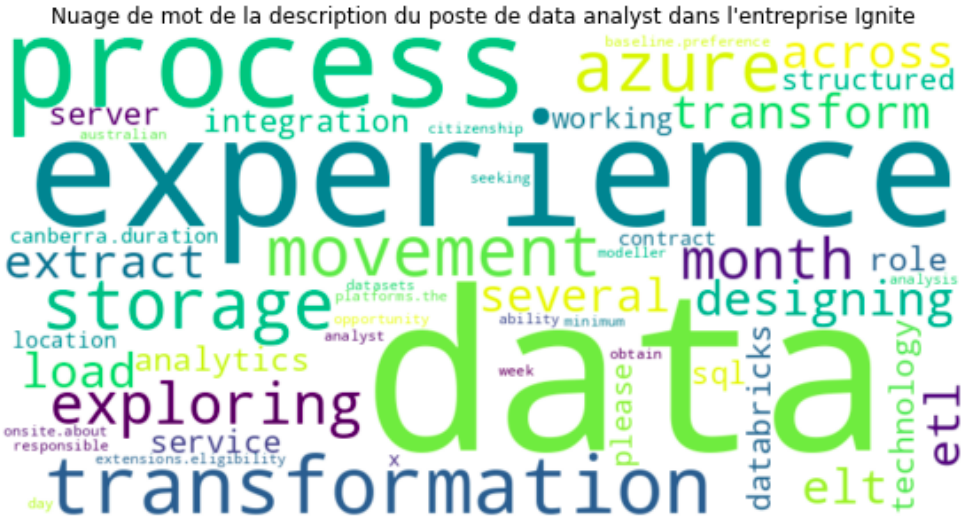


Figure 6: Nuage de mots de la description du poste de data analyst chez Ignite

Pour ce nouveau poste, les mots ‘transformation’, ‘experience’, ‘process’, ‘exploring’, ‘movement’, ‘storage’ semblent prédominants après le mot ‘data’. Sur ce poste un profil ayant déjà une expérience dans le domaine semble attendu ainsi qu’une expérience avec le logiciel ‘azure’. Un junior ne sera potentiellement pas retenu.

Modèle sur les données

Similarité des documents

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 en sortie (nom du poste, entreprise, lieu, description, …) nous affichons l’offre d’emploi qui est la plus proche de ses envies.

Nous avons également regardé pour classifier les offres d’emploi avec la matrice calculée pour la similarité. Nous avons appliqué une CAH, nous avions deux possibilités pour choisir le nombre de groupes, soit le nombre de catégorie de métiers (8), soit le nombre de groupes trouvé par un critère comme celui Silhouette. Dans notre notebook, le nombre de groupes est de 5 (celui pour lequel le coefficient Silhouette est le plus faible). Nous pouvons comparer nos groupes avec nos 8 catégories. Nous observons qu’il n’y a pas de réelle distinction pour nos catégories. Les descriptifs entre les différents métiers sont tellement similaires que la répartition en cluster ne se fait pas sur une catégorie mais plutôt sur les besoins (description) dans un métier. Ils peuvent être les mêmes pour un data analyst et un data scientist. Cela explique le résultat observé.

Prédiction du poste

Nous nous sommes également intéressées au problème suivant : étant en présence d'une annonce 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.

Avis des entreprises

Nuage de mot

Concernant la base de données sur les avis des entreprises, nous avons lemmatisé les commentaires sur les 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 :

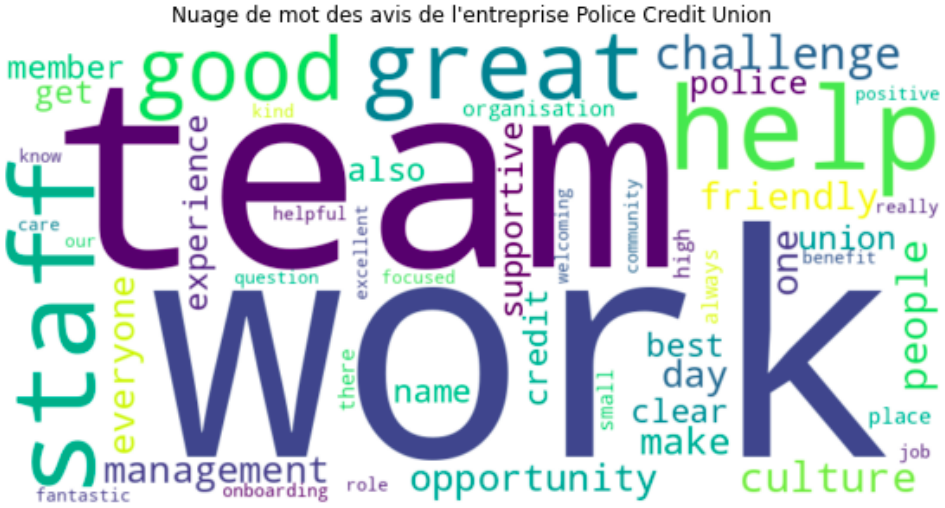


Figure 7 : Nuage de mots des avis sur une entreprise

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. On voit également que plusieurs mots positifs ressortent comme : 'good' et 'great'.

Nous avons créé une fonction qui ressort pour une entreprise le nuage de mot associé aux avis positifs et un pour les avis négatifs.

Voici un exemple pour l'entreprise Marathon Health.

Figure 8 : Nuage de mots des avis positifs pour Marathon Health

Figure 9 : 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.

Modèle sur les avis

Suite à ces premières analyses nous allons étudier de manière plus approfondie nos données sur les avis.

En étude de textes il est possible de connaître 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 (avantages 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.

Nous agrégeons par entreprises pour comparer les résultats et avoir moins de lignes à comparer. Ci-dessous les résultats de notre étude.

entreprise	note	Note_calcule	Comparaison	différence
ALS Limited	2.3	2.8	Plus grand	-0.5
Accolade Wines	3.0	3.1	Plus grand	-0.1
Austin Health	2.9	3.1	Plus grand	-0.2
Australian Clinical Labs	1.5	2.9	Plus grand	-1.4
Australian Department of Defence	3.3	3.0	Plus petit	0.3
Australian Retirement Trust	4.5	3.2	Plus petit	1.3
AustralianSuper	3.3	3.0	Plus petit	0.3
Auto One	4.0	3.6	Plus petit	0.4
BAE Systems	2.7	3.0	Plus grand	-0.3
Bank of Queensland	2.1	2.5	Plus grand	-0.4
BlueScope Steel	2.7	2.8	Plus grand	-0.1
Border Express	4.6	3.5	Plus petit	1.1
Brisbane North PHN	5.0	3.4	Plus petit	1.6
Bunzl Australasia	2.8	2.4	Plus petit	0.4
Charles Darwin University	4.5	3.5	Plus petit	1.0
City of Sydney	3.1	2.7	Plus petit	0.4
Department of Training and Workforce Development WA	3.0	2.6	Plus petit	0.4
Energy Australia	2.5	3.0	Plus grand	-0.5

Figure 10 : Tableau des notes calculées et de l'écart

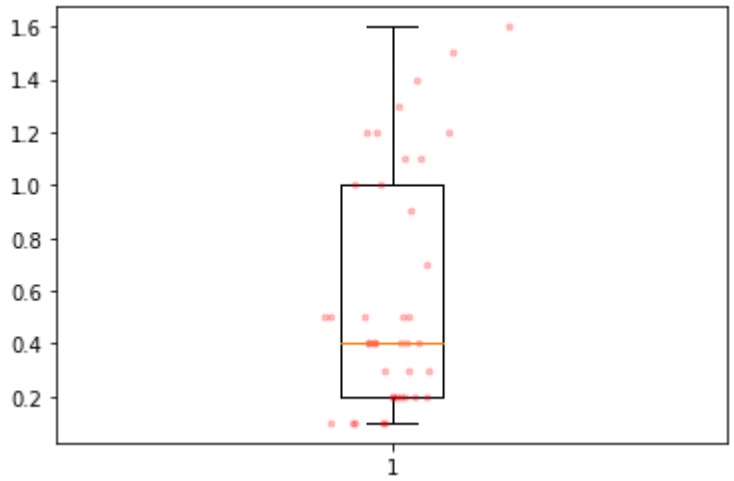


Figure 11 : Distribution de l'écart entre la note réelle et la note calculée

Nous observons que notre méthode a donné des notes similaires à celles données par les utilisateurs. Cela est très positif. Etant en présence de texte nous pouvons donc prédire la note qu'un utilisateur pourrait donner à une entreprise. Notre plus grand écart est de 1.6, mais en moyenne nous avons un écart de 0.5. Une piste d'amélioration pourrait être de récupérer plus d'avis sur les entreprises, car nous avons seulement récupéré les 10 derniers avis de l'entreprise. Nous pouvons donc se dire que si les 10 derniers avis sont très positifs (ou négatifs) mais que les autres avis étaient mauvais (ou bons) cela fausse notre estimation. Mais de manière générale nous arrivons bien à retrouver le sentiment des avis.

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 pour une nouvelle annonce, la similarité des annonces et la prédiction de la note de l'entreprise. Étant donné que les métiers liés aux données présentent peu de barrières entre eux, nos modèles auraient pu être encore meilleure si les annonces de postes étaient plus différentes. 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, par exemple, le lieu ou le salaire et sortir les annonces qui correspondent à la recherche ainsi que les wordclouds associé.
- Estimer le salaire d'un poste.
- Comparer à l'aide des 'llm' pourquoi un employé a mis une note à 4 plutôt que 5…