Projet Long

Ait benali Faycal Bouchachia Ayoub, Boucherir Mohamed Zineddine

February 18, 2022

Contents

1 Introduction générale

Depuis les deux dernières décennies, internet est devenu un outil incontournable de recrutement.Les méthodes de recherche et les relations avec les candidats ont profondément changé. Le e-recrutement se structure et évolue. En effet , ce média puissant offre à l'entreprise, quelle que soit sa taille, un potentiel important de candidats qui apportent au recruteur : ciblage pertinent, rapidité, coûts maitrisés et ouverture sur le monde, pour des recrutements ponctuels, récurrents ou de masse tant en France qu'à l'international. De nombreux sites emploi sont, aujourd'hui, la première source de CV qualifiés , ainsi le é-recrutement est devenu un des services en ligne les plus populaires pour les demandeurs d'emploi ainsi que pour les employeurs. Néanmoins, le processus de recrutement et de recherche d'emploi n'est pas facile, notamment en ce qui concerne la recherche de profils et de talents, car de nombreuses approches se limitent actuellement à la recherche par mots-clé, qui n'est plus efficace lorsque la taille des données devient énorme.

Dans ce projet, nous présentons SEEKJOBS un système de recommandation d'offre d'emploi basé sur les méthodes d'apprentissage automatique et du traitement du langage naturel. Ce système permet au recruteur de trouver les meilleurs candidats à un emploi et aide aussi les chercheurs d'emploi à trouver les emplois appropriés à leurs profiles les présentés dans les CVs. La recommandation est basé sur un modèle de classifcation des textes et le calcul de similarités entre l'offre d'emploi et les CVs des candidats. Dans ce rapport , nous présenterons un aperçu des données utilisées, leur collections et les différentes traitements faites sur le corpus d'entraînement. l'extraction d'informations à partir des CVs, les démarches de la formation d'un modèle de classification et le modèle de similarité sont aussi présentés dans ce chapitre. En fin , une conclusion contenant le bilan et les différentes perspectives pour améliorer ce travail.

2 Collecte des données

Nous avons utilisé les techniques du web scrapping pour collecter les données du site Indeed.

2.1 web scrapping

C'est une technique utilisée pour extraire des données des sites web. Il s'agit d'un processus automatisé dans lequel une application traite le code HTML d'une page web afin d'extraire des données Les étapes de la collection des CVs

1. L'examination de l'URL et la page contenant les résultats de la recherche.
2. L'extraction des liens des CVs.
3. L'extraction des données des CVs.
4. La segmentation des CVs en blocs.

5. Le stockage des données collectées dans un fichier ”json”.


```
{
    "Title":{"0":"SOFTWARE ENGINEER, AUTHOR-IT"},
    "Location":{"0":"Seattle, WA"},
    "Work Experience":{"0":"SOFTWARE ENGINEER, AUTHOR-ITAuthor ItNovember 2016 to October 2018Translated UX wireframes\/mockups into responsive a
    "Education":{"0":" in HEALTH INFORMATICS AND MANAGEMENTUNIVERSITY OF WASHINGTON June 2011"},
    "Skills":{"0":"Ajax (Less than 1 year), AngularJS (3 years), Backbone (2 years), Bootstrap (3 years), Django (Less than 1 year), Python (Less
    "Links":{"0":"https:\/\/www.linkedin.com\/in\/vuhdinhhttps:\/\/github.com\/vhdinh"}
}
```


Figure 1: Un cv segmenté (Corpus de CV)

3 L’architecture générale

L’architecture générale de notre système de recommandations est donnée par la figure

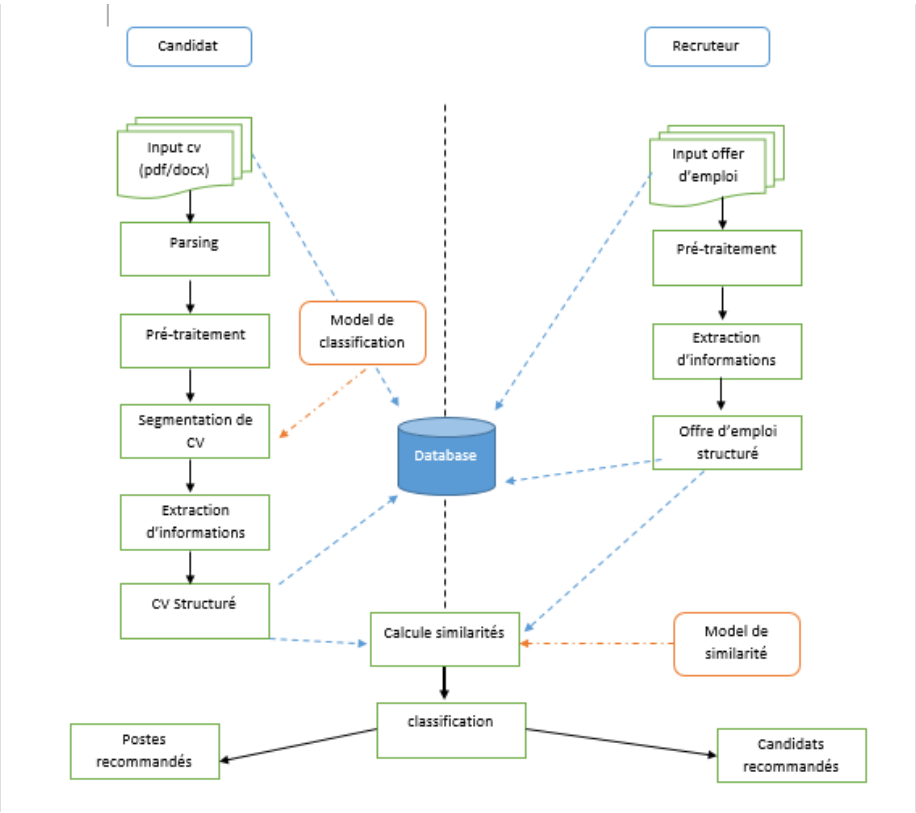

Figure 2: Architecture de SEEKJOBS

4 Ségmentation du CV

Pour la segmentation des CVs nous avons utilisé un classificateur de texte pour identifier la classe des différents blocs (skills, education, work experience). Nous avons utilisé les CVs collectés pour l'entrainement du modèle.

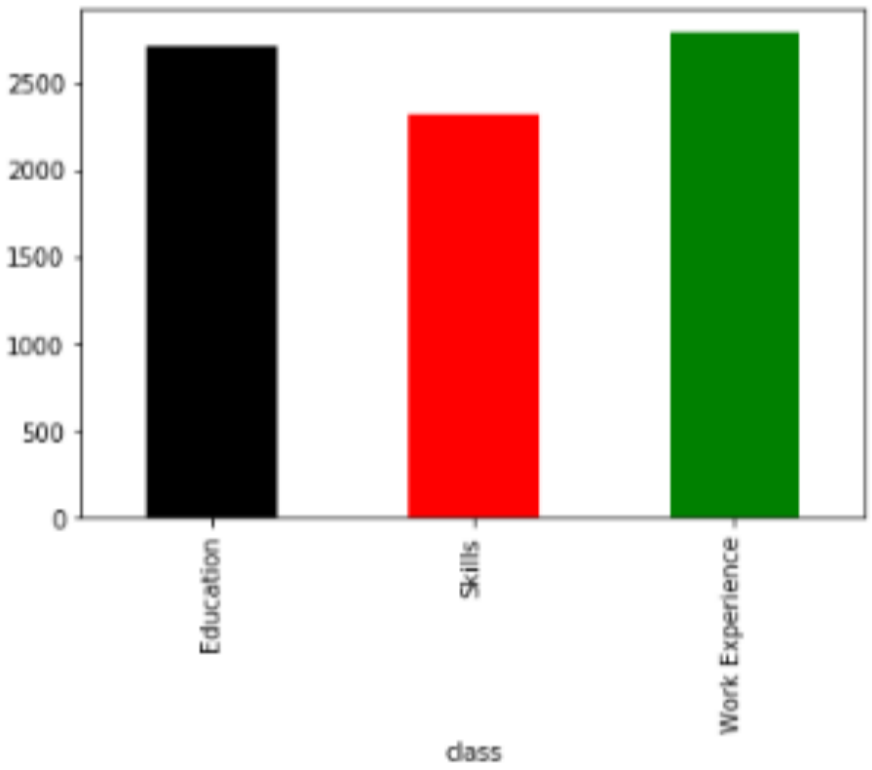

Figure 3: distribution des données selon les classes

4.1 Pré-traitement

Les données textuelles brutes exprimées en langage naturel ne sont pas utilisables directement par la machine.

L'étape de prétraitement a donc a pour objectif de préparer le texte pour la phase d'apprentissage. Techniquement, c'est l'étape de nettoyage du texte (éliminer les termes et caractères non signicatifs), l'extraction des occurrences des termes définissant les segments et le calcul des fréquences de ces derniers.

Ensuite, tous les mots sont convertis en minuscules puis traités à l'aide de Porter Stemmer du package NLTK. Le but est de diminuer le nombre de mots flexibles apparaissant dans les CVs. Cela permet de réduire la taille du vocabulaire et d'améliorer le volume de l'espace des fonctions dans le corpus.

4.2 Apprentissage

Lors de la construction du modèle, nous pouvons choisir parmi une large gamme d'algorithmes de classication. Nous avons testé quatres algorithmes diérents : KNN, multinomial naïf Bayes, linéaire SVC.

Avant de comparer les performances des différents algorithmes de classification, nous nous sommes basés sur des mesures de statistiques communes tels que la précision et le rappel. Après avoir testé les algorithmes, les expérimentation ont montré que l'algorithme de du classificateur naif de bayes est simple et facile à deployer par rapport aux autres.

4.2.1 classifieur Naïve Bayes

Cette méthode se base sur le théorème de Bayes qui permet de calculer les probabilités conditionnelles.

Dans le cas de la classification du texte, la méthode Naïve Bayes est utilisée comme suit : on cherche la classification qui maximise la probabilité d'observer les mots du document. Lors de la phase d'entraînement, le classifieur calcule les probabilités qu'un nouveau document appartient à telle catégorie à partir de la proportion des documents d'entraînement appartenant à cette catégorie. Il calcule aussi la probabilité qu'un mot donné soit présent dans un texte, sachant que ce texte appartient à une telle classe. Quand un nouveau document doit être classé, on calcule les probabilités qu'il appartient à chacune des classes à l'aide de la règle de Bayes et on prend la classe la plus probable.

Soit le jeu de données D (corpus) composé de n paires (x,c), avec x un document qui contient un ensemble de mots w et c la classe auquel ce texte appartient parmi q classes possibles :

$$D = \{(x_1,c_1),(x_2,c_2),...,(x_n,c_n)\}\forall(x_i,c_i)_{1\leq i\leq n}, c_i \in \{y_1,y_2,...,y_q\}, x_i = \{w_1,w_2,...,w_{di}\}$$

Le modèle Naïve Bayes modélise la probabilité jointe d'un texte x et de sa catégorie y $P(X = x, Y = y)$. Par définition :

$$P(X = x, Y = y) = P(Y = y)P(X = x|Y = y)$$

Prenons comme hypothèse que mots dépendent seulement de la classe et il sont indépendants entre eux, on peut réécrire $P(X = x, Y = y)$ comme le produit des probabilités conditionnelles des mots.

$$P(X = x|Y = y) = P(X = [w_1, w_2, ..., w_d]|Y = y) = \prod_{k=1}^{d} P(W = w_k|Y = y)$$

Ainsi, on obtient :

$$P(X = x, Y = y) = P(Y = y) \prod_{k=1}^{d} P(W = w_k|Y = y)$$

La classe d'une nouvelle observation z est déterminée selon la règle décision suivante :

$$\hat{y} = argmax_{j \in \{1,2,...,q\}} \{P(Y = y_j|X = z)\}$$

En appliquant le théorème de Bayess :

$$\hat{y} = argmax_{j \in \{1,2,...,q\}} \{\frac{P(X=z|Y=y_j)P(Y=y_j)}{P(X=z)}\}$$

Comme le dénominateur, $P(X = z)$, ne dépend pas de Y, sa valeur est constante et minimiser l'expression ci-dessus revient à maximiser :

$$\hat{y} = argmax_{j \in \{1,2,...,q\}} \{P(X = z|Y = y_j)P(Y = y_j)\}$$
$$\hat{y} = argmax_{j \in \{1,2,...,q\}} \{P(Y = y_j) \prod_{k=1}^{d_z} P(W = w_k|Y = y_j)\}$$

Si on pose $P(Y = y_j) = g(y_j)$ et $P(W = w_k|Y = y_j) = g_k(w_k|y_j)$, le modèle Naïve Bayes s'écrit comme suit :

$$\hat{y} = argmax_{j \in \{1,2,...,q\}} \{g(y_j) \prod_{k=1}^{d_z} g_k(w_k|y_j)\}$$

Où $g(y_j)$ et $g_k(w_k|y_j)$ sont des paramètres à estimer.

On estime la probabilité apriori d'observer la classe y en mesurant sa fréquence relative dans D (le nombre de documents qui ont la classe y_j comme catégorie sur le nombre total de documents):

$$g(y_j) = \frac{\sum_{i=1}^{n} (y_j = y_i)}{n}$$

On estime la probabilité d'observer le mot w_k conditionnellement à la classe y_j en calculant le rapport entre le nombre de fois que w_k apparaît dans les documents de la catégorie y_j ,et le nombre de mots total dans les documents de la catégorie y_j:

$$g(w_k|y_j)_k = \frac{\sum_{i=1}^{n} [(y_j = y_i) \sum_{r=1}^{d_i} (w_k = w_r)]}{\sum_{i=1}^{n} [(y_j = y_i) d_i]}$$

	Précision	Recall	F1-score
Naive Bayes	0.89	0.90	0.87

Figure 4: la performance des modèles de classi

5 Diffcultés retrouvées

Durant la réalisation , nous avons retrouvé quelques difficultés techniques (liées au materiel) nottament le stockage , en effet , nous n'avons pas pu faire l'apprentissage , sur toutes les données qu'on a collecté , comme elle ont une taille très grande , ce qui nous a poussé à diminuer la taille des données pour pouvoir avancer

6 Fonctionnalité non encore implémentées

A cette étape, nous avons réaliser une partie importante du projet , mais il reste encore des taches à faire, en l'occurence la plateforme Web. En plus , il reste la partie de calcul de similarité entre une offre d'emploi et un CV, ce qui nous permet de réaliser la recommandation.

7 Outils

En ce qui concerne les outils et langage de programmation que nous avons jugé utiles pour notre réalisation nous avons utilisé pour réaliser le porjet : Python, Selenium, BeautifulSoup, NLTK, scikit-learn et django, et concernant la création de design de site web du côté client nous avons opté pour Bootstrap .

8 Conclusion

Nous nous sommes intéressés dans ce travail à concevoir et implémenter un système de recommandation d'emploi qui facilite le processus de recrutement et de la recherche d'emploi. Nous avons choisi de travailler sur les données d'un seul domaine de travaille.

Nous avons choisi une approche basée sur le contenu des CVs et des offres d'emploi pour concevoir notre système.

Nous avons utilisé les différentes techniques du TALN afin d'extraire les informations pertinentes à partir des CVs et les offres d'emploi tels que les compétences et la formation.

Un modèle de classification de textes a été entrainé qui permet de mieux segmenter le CV selon la catégorie des informations qu'il présente.

il nous reste l'implémentation d'un modèle de similarité pour permettre de faire la correspondance entre les CVs et les offres d'emploi, et la plateforme web afin de mettre en œuvre notre système de recommandation.

Nous citons comme perspectives les questions suivantes qui pourraient améliorer notre système :

- ajouter un questionnaire personnalisé selon le CV, pour évaluer les compétences du candidat.
- Généraliser notre modèle pour travailler dans un contexte multilingue et trouver un benchmark permettant de meilleurs tests.
- Utiliser l'approche hybride pour améliorer les recommandations et tester d'autres modèles d'apprentissage automatique.
- Automatiser l'ajout de nouveaux domaines de travaille et l'entrainement des modèles de classification et similarité selon les données de ce domaine.