

PROJET FIN D'ÉTUDE : Système recommandation des métiers à partir du site indeed

Réalisée par: ABDI GHAVIDEL Azadeh

Master2 MIASHS (parcours: WEB ANALYSE)

Année: 2020-2021

Établissement: Université de LILLE



SOMMAIRE

CAHIER DES CHARGES	3
<i>LES CONTRAINTES.....</i>	<i>3</i>
<i>LES PRESTATIONS ATTENDUES</i>	<i>3</i>
<i>LES BESOINS FONCTIONNELS.....</i>	<i>3</i>
<i>DÉLAI DE PROJET</i>	<i>3</i>
INTRODUCTION ET OBJECTIF.....	4
LES Étapes du projet	4
<i>PREMIÈRE ÉTAPE « Préparation les machines virtuelles »</i>	<i>4</i>
<i>DEUXIÈME ÉTAPE «scraper les données »</i>	<i>4</i>
LES FONCTIONNES.....	5
<i>TROISIÈME ÉTAPE « Pre-processing»</i>	<i>6</i>
LES FONCTIONNES.....	6
ÉXPORTATION LES DONNÉES NETTOYÉES	6
<i>QUATRIÈME ÉTAPE « Prédiction».....</i>	<i>6</i>
TF-IDF	6
KMEANS.....	7
FLASK	7
<i>CINQUIÈME ÉTAPE «Application web».....</i>	<i>7</i>
PARTIE HTML.....	7
PARTIE JAVASCRIPT ET AJAX.....	8
Conclusion.....	9
Remerciement.....	9



CAHIER DES CHARGES

LES CONTRAINTES

- Le projet doit prendre la forme d'un site internet soit héberger sur un serveur de votre choix soit héberger sur une machine virtuelle du service OpenStack de l'université.
- Les codes sources doivent être disponible via git.
- Des dumps de la base de données doivent être téléchargeable simplement.

LES PRESTATIONS ATTENDUES

- Un rapport (quelques pages) expliquant:
 - o Les objectifs du projet,
 - o Le déroulement de sa réalisation,
 - o La structure du code source,
 - o L'adéquation du projet avec les objectifs d'évaluation.

LES BESOINS FONCTIONNELS

- Machine virtuelle
- Python
- Scraping
- NLP
- Clustering
- SQL

DÉLAI DE PROJET

A finaliser pour la date : le mercredi 20 Janvier 2021.



INTRODUCTION ET OBJECTIF

L'idée de départ de mon projet était de scraper le site indeed pour certains métiers et d'identifier les compétences plus demandées pour certains types de poste.

Au fur et au mesure dans mon cursus et avec les conseils de mes professeurs , je suis allée vers un application web , qui fonctionne comme un système de recommandation des jobs.

Objectif de projet est de faire un clustering en fonctionne de compétences demandées pour chaque une des postes pour identifier les différents types de jobs et si une personne entre des mots clés avec les compétences à elle, de lui donner les offres qui sont plus proche à ses compétences.

LES ÉTAPES DU PROJET

PREMIÈRE ÉTAPE « PREPARATION LES MACHINES VIRTUELLES »

La première étape du travail a consisté à créer une machine virtuelle sur le cloud d'université, pour le stockage de base de données , mettre les codes et la page de la partie visuelle du projet.

L'IP de mes instances :

Serveur : 172.28.100.229

Web : <http://172.28.100.229/projet/>

DEUXIÈME ÉTAPE « SCRAPER LES DONNEES »

Après la création mon instance, j'ai créé une base de données qui s'appelle projet, dans cette base de données j'ai créé une table « annonces » afin d'importer les données que j'ai scrapé du site « fr.indeed.com », ma table contient de 9 champs (id, motcle, metier, entreprise, location, datedannoe, lien, sommaire et description).

Pour la partie scraping j'ai utilisé « BeautifulSoup ».

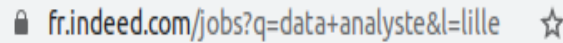


LES FONCTIONNES

J'ai créé 4 fonctionnes pour pouvoir scraper le site fr.indeed.com

1- `recupere_url(n_metier, lieu)` : est une fonctionne qui prend le nom de métier et le lieu pour modéliser et retourner un URL

Les URL dans le site indeed sont sous le format : 'https://fr.indeed.com/emplois?q={}&l={}' dont q représente le nom de métier et « l » représente la location.

 `fr.indeed.com/jobs?q=data+analyste&l=lille` ☆

2- `scraper_lien(n_metier, lieu)` : une fonctionne pour scraper la page de résultat de métiers et récupérer les blocs des annonces en utilisant « BeautifulSoup »

3- `recupere_donnees(bloque)` : La page de résultat contient de nombres annonces, chaque un dans un bloque, `recupere_donnees` est une fonctionne qui prend une annonce à la fois pour récupérer ces données.



4- `recupere_decription(lien)` : chaque annonce a un lien spécifique qui nous dirige vers sa description complémentaire, `recupere_decription` est une fonctionne qui prends le lien d'annonce et récupère sa description à l'aide de « BeautifulSoup »

Et en utilisant MySQL, j'ai sauvegardé une annonce à la fois dans ma base de données.

À la fin en utilisant une boucle for j'ai scrapé les annonces de métiers : ['data analyste', 'informatique', 'seo', 'marketing', 'python', 'js', 'cloud', 'web']

TROISIÈME ÉTAPE « PRE-PROCESSING »

Dans la partie preprocessing, je vais préparer mes données brutes au meilleur format afin d'être utilisées dans la partie « prédiction »

J'ai commencé par importation les librairies dont j'aurai besoins dans mes démarches comme « nltk » pour tokenisation et normalisation les corpus, « langdetect » pour détecter le langage du corpus, TfidfVectorizer et Kmeans pour clustering mes annonces.

LES FONCTIONNES

J'ai créé 4 fonctionnes pour cette étape :

- 1- « normalisationCorpus » est une fonctionne qui va ajouter la colonne 'phrase_token' qui contiendra le nom de métier et sa description en minuscule et trois colonnes vides 'phrases_cible', 'resume' et 'category' afin d'être utilisé dans le preprocessing
- 2- « langueDetecte » est une fonctionne qui va détecter les annonces anglais et puis il les supprime
- 3- « nettoyage » est une fonctionne afin de chercher quelques mots ciblés qui peuvent contenir les compétences
- 4- « stopMots » est une fonctionne qui va supprimer les mots fréquents comme (de, le, la, je, avec , ...) dans le corpus pour qu'il soit plus propre pour l'étape de Machine Learning.

EXPORTATION LES DONNÉES NETTOYÉES

Et à la fin j'ai exporté mes données traitées à la format « csv » afin d'être utilisée dans la partie prédiction je l'ai stocké dans ma base de données dans la table « final ».

QUATRIÈME ÉTAPE « PREDICTION »

Après la récupération mes données j'ai utilisé la technique « TF-IDF ».

TF-IDF

« TF-IDF » qui signifie "Term Frequency - Inverse Document Frequency", est une technique pour quantifier un mot dans les documents, nous calculons généralement un poids à chaque mot qui signifie l'importance du mot dans le document et le corpus.

Cette méthode est une technique largement utilisée dans la recherche d'informations et l'exploration de textes.



KMEANS

Et enfin avec Kmeans j'ai produit 20 clusters afin de clustering les annonces plus proches.

FLASK

J'ai utilisé flask pour pouvoir envoyer le texte de la page index vers mon fichier python et récupérer le résultat de la prédiction.

CINQUIÈME ÉTAPE «APPLICATION WEB»

Pour illustrer mon application web j'ai créé un fichier index.html et un fichier style.css.

Ma page index.html est construit par deux parties :

1. HTML
2. JavaScript et Ajax

PARTIE HTML

Dans cette partie j'ai créé :

1. Une balise <div> afin de récupérer le texte saisi par utilisateur.
 - Une balise <input> à remplir par internaute.
 - Une balise <button> pour commencer la recherche en utilisant JavaScript dès qu'internaute clique sur ça.

Recommande moi :

Mon metier préféré ...

Recherche



2. Une balise `<div id="search-list" class="search-list">` pour montrer les résultats de recherche (les métiers prédits par la machine)

Recommande moi :

Développeur Back-end Java DEVOPS Confirmé – Lille (H/F)

Filiale privée du groupe SNCF, e. Voyageurs SNCF est une entité créée en octobre 2018, qui rassemble les compétences digitales client du groupe SNCF.

OUI.sncf

Ingénieur Etude et Développement Track & Trace (F/H)

Worldline [Euronext: WLN] est le leader du marché européen dans le secteur des services de paiement et de transaction. Poste de Développeur Track & Trace (F/H).

Worldline

- 3- Une balise `<div id="popup" class="popup">` afin de montrer la description de métier sélectionné par utilisateur.

Elle va apparaître dès qu'internaute clique sur le bouton « Plus d'infos ».

Lead developpeur Front / Freelance X

Coriom Conseil, réseau collaboratif d'indépendants spécialisés dans les missions de conseil et d'expertise auprès des organisations métiers, recherche pour un de ses clients de la métropole lilloise un Lead développeur Front-end

MISSIONS : En tant que lead développeur front, vous interviendrez dans tout le processus d'un projet agile - Contribuer à la définition des architectures techniques en collaboration avec toute l'équipe - Collaborer efficacement avec les designers UI / UX pour créer/transformer les maquettes en code - Leader de la montée en compétence des autres membres de l'équipe sur angular - Participer à la conception des features et à l'ergonomie des applications et de l'interface utilisateurs - Développer (avec les autres membres de l'équipe) en utilisant le langage Angular tout en respectant les bonnes pratiques inner source et devops - Concevoir et participer aux Tests unitaires et fonctionnels

ENVIRONNEMENT TECHNIQUE : - Angular 8+ - Programmation réactive / RxJS - Javascript (ES6+) - Typescript 3+ - CSS, SASS Un plus : Docker / Kubernetes Gitlab CI

PROFIL : - Expérience utilisateur first - Pédagogue, partager vos savoirs avec les autres - Soucieux de la qualité et savoir la mesurer (Webperf, Accessibilité, Sémantique, Normes de code,) - Curieux, intéressez aux problématiques métier, challenger les demandes dans un objectif d'efficacité et de pragmatisme

- Autonomie, partager ses choix et bonnes pratiques avec le reste de l'équipe

LANGUES : Anglais professionnel (impératif)

[Lien vers l'annonce](#)

PARTIE JAVASCRIPT ET AJAX

Dans cette partie j'ai créé une fonction nommée « search » en utilisant JavaScript et Ajax pour envoyer des données à mon fichier « prediction.py » en arrière-plan et récupérer le résultat afin de le monter dans ma page « index.html »

J'ai déclaré un objet « XMLHttpRequest » qui permet d'obtenir des données au format JSON, ou même un simple texte à l'aide de requêtes HTTP.

Et j'ai utilisé la méthode GET pour envoyer mon texte à le fichier python.

Pour chaque recherche je vais prendre quinze annonces (obj) prédits par mon modèle Kmeans que j'ai utilisé dans « prediction.py »,

Chaque obj a les attributs « id, métier, entreprise, sommaire, description et son vrai lien représenté sur le site indeed dont je vais les utiliser dans ma page web.

CONCLUSION

En faisant ce projet, j'ai appris beaucoup de choses comme utilisation de la librairie BeautifulSoup, NLTK, TF-IDF, Kmeans, Ajax, JavaScript, etc.

J'ai appris beaucoup plus de connaissance sur python et Machine Learning .

J'avais beaucoup de problème sur nettoyage de mes données, j'ai pu trouver les solutions pour avancer dans mon projet.

Comme j'avais noté dans la partie d'introduction, je voulais compter le nombre de compétences plus demandés mais je suis allée vers un système de recommandation , faire de la fouille de textes sur les textes français était un vrai challenge, je pense que mon modèle peut améliorer encore.

REMERCIEMENT

Je voudrais remercier Madame Petra Rahme, Monsieur Louis Bigo, Monsieur Dorian Baudry, Monsieur Mohamed Elati et Monsieur Charles Paperman, d'avoir enrichi mes connaissances et de m'avoir guidé durant toute cette année.

J'ai grandement apprécié votre soutien, votre implication et votre expérience tout au long de ce projet.

