RAPPORT DE PROJET

PROJET DE WEB-SCRAPING DE DONNEES D'EMPLOI EN FRANCE

Auteur : GANIYU Isaac, étudiant en 2e année du cycle ingénieur à l'ENSAI de Rennes

27 décembre 2022

Table des matières

INTRODUCTION

Ce projet a été réalisé en réponse au challenge de web-scraping pour le stage de chargé d'études sur les salaires à la Banque de France.

Le but de ce rapport est d'expliciter aux lecteurs intéressés la démarche utilisée pour acquérir les données d'emploi en France par une méthode de web-scraping et aussi, d'éclairer sur la structure de la base de données finale obtenue. Ce projet a été réalisé dans un cadre purement académique. Il pourrait toutefois servir l'étude de la Banque de France sur les salaires en cas de réussite au challenge. De plus, les données scrapées n'ont pas un caractère personnel, ce qui, implicitement, permet leur extraction et leur manipulation.

Dans ce rapport, nous commencerons par donner l'objectif du projet ainsi que les raisons qui ont motivé le choix du site utilisé pour le web-scraping, ensuite nous expliciterons la démarche utilisée, puis nous éclairerons les lecteurs sur les étapes de la préparation des données en vue d'une analyse future. Enfin, nous étalerons les difficultés rencontrées dans la réalisation de ce projet et donnerons une piste d'amélioration.

1. OBJECTIF ET CHOIX DU SITE

Pour ce projet, nous avons décidé de construire une base de données d'emploi en France. Nous avions le choix entre plusieurs pays à savoir : la France, le Royaume-Uni, l'Allemagne, l'Italie, l'Espagne, les Pays Bas et la Belgique. Le choix de la France est motivé par le fait que nous nous mettons dans une optique d'analyse. En effet, l'objectif final de la base de données construite pourrait être le suivi du marché des offres d'emploi en France. Il est donc essentiel de se placer dans une optique d'analyse. Et vu que l'environnement français et les réalités sur l'emploi en France sont plus ou moins connues comparées aux autres pays, il était préférable de collecter des données sur la France, ce qui pourrait théoriquement faciliter nos analyses futures.

Il nous était ensuite indispensable de choisir un site regroupant des offres d'emploi en France pour y collecter les données. Dans ce projet le site choisi est **france-emploi.com.**

Ce site a été choisi pour les raisons suivantes :

- Premièrement, ce site renferme de nombreuses offres d'emploi avec des informations très détaillées sur ces offres. En particulier, on peut retrouver, le lien internet de l'annonce, le poste proposé, la localisation de l'emploi, le salaire proposé, le statut de l'emploi et le descriptif de l'annonce.
- Deuxièmement, pour la structure du site. En effet, la structure du site est assez simple. Techniquement, le code source du site est fait purement de code html, il n'y a pas de code javascript injecté dans le code source. Ce cas est relativement peu complexe. De plus il est aisé de passer d'une page d'offres d'emploi à une autre.
- Enfin, les conditions générales d'utilisation du site permettent de collecter les données tant qu'il ne s'agit pas de données à caractère personnel. Et dans notre cas, les données collectées ne sont pas des données à caractère personnel.

2. DEMARCHE DE COLLECTE DES DONNEES

En vue de collecter les données du site par web-scraping, nous avons élaboré un programme Python basé sur le package ***bs4*** et sa librairie ***BeautifulSoup*** mais aussi sur le package ***requests***.

L'algorithme de web-scraping a été lancé le 26 décembre 2022 à 14h45. La base obtenue est donc constituée des annonces présentes sur le site à cet instant précis.

Le programme élaboré contient une boucle principale qui se charge d'itérer sur les différentes pages du site qui renferment des offres d'emploi. Nous en avons recensé 500. Pour chaque page, le code html, de la page est « parsée », puis pour chaque annonce de la page, nous avons pu récupérer les éléments suivants :

- Le poste proposé ;
- Le lien de l'annonce, obtenu après concaténation du lien de base du site et d'une référence spécifique à l'annonce ;
- La date de publication de l'annonce ;
- La localisation de l'emploi ;
- Le contrat (CDI, CDD, intérim, alternance, stage, franchise ou saisonnier) ;
- Les informations sur le salaire. Il s'agissait d'un texte sous une structure particulière. De ce texte nous avons récupéré : le type de salaire (annuel, mensuel ou horaire), le salaire minimum et le salaire maximum. Nous avons extrait ces informations du texte brut initial grâce à des méthodes de manipulation de texte, notamment grâce à la méthode *split()*. Nous avons jugé important de le faire, dans le but de faciliter les analyses futures sur cette base de données ;
- La description de l'annonce.

Pour certaines annonces, le salaire ou le type de contrat n'est pas renseigné. Nous avons donc renseigné ces données comme *None* dans Python. Il en résulte naturellement que, pour ces données, les colonnes correspondantes sont vides.

Une fois ces données récupérées via le site, nous les mettons sous la forme d'un *DataFrame* que nous enregistrons ensuite sous un format (*.csv*). Il est ensuite aisé d'importer ces données dans Excel.

3. MISE EN FORME ET NETTOYAGE DES DONNEES

Comme spécifié dans le challenge, nous avons remplacé le cases vides du fichier Excel (correspondantes aux valeurs manquantes) par des «.».

Nous obtenons un fichier Excel final de 11 colonnes et 10 000 lignes de données. Les colonnes sont (sur la base des éléments récupérés cités ci-dessus) :

- **id** : un identifiant unique à chaque offre d'emploi. Ces identifiants vont de 0 à 9999. Cette colonne a été mise sous format *nombre.*
- **Emploi** : L'intitulé du poste. Cette colonne a été mise sous format *texte*.
- **Lien de l'annonce** : Le lien renvoyant directement à l'annonce sur le site utilisé. Cette colonne a été mise sous format *texte*.
- **Date** : La date de publication de l'annonce. Cette colonne a été mise sous format *date*.
- **Localisation** : Il s'agit de la localisation de l'emploi proposé. Cette colonne a été mise sous format *texte*.
- **Contrat** : Le type de contrat proposé (CDI, CDD, stage, etc.). Cette colonne a été mise sous format *texte*.
- **Salaire** : le texte de base indiquant les informations sur le salaire relatives au poste proposé. Cette colonne a été mise sous format *texte*.
- **Type de salaire** : La périodicité du salaire (annuel, mensuel ou horaire). Cette colonne a été mise sous format *texte*.
- **Salaire minimum** : Le salaire minimum proposé. Il faut noter que dans certains cas, ce salaire minimum n'est pas spécifié et seul le salaire maximum l'est. Dans ces cas, la valeur est manquante. Cette colonne a été mise sous format *nombre*.
- **Salaire maximum** : Le salaire maximum proposé. Il faut noter que dans certains cas, ce salaire maximum n'est pas spécifié et seul le salaire minimum l'est. Dans ces cas, la valeur est manquante. Cette colonne a été mise sous format *nombre*.
- **Descriptif** : La description de l'annonce telle qu'elle est sur le site. Cette colonne a été mise sous format *texte*.

4. DIFFICULTES RENCONTREES ET PISTE D'AMELIORATION

Durant ce projet, nous avons rencontré quelques difficultés. La difficulté principale a été la recherche du site web adéquat. En effet, trouver un site web, correspondant aux informations recherchées et permettant de scraper les données (que ce soit au niveau technique, comme au niveau de l'autorisation) n'a pas été facile. Une fois le site web trouvé, le projet a avancé plutôt rapidement. Toutefois nous avons noté certaines difficultés techniques : les salaires étaient sous

une forme texte contentant le type de salaire, le salaire minimum et/ou le salaire maximum (selon le cas). De plus les prix (salaires) avaient un encodage non adéquat. Il a donc fallu surmonter toutes ces difficultés et “décrypter“ le message pour construire les colonnes « type de salaire », « salaire minimum » et « salaire maximum ».

Enfin, une dernière difficulté a été le temps. N’ayant pas rapidement pris connaissance du challenge et compte tenu d’autres contraintes, nous avons effectué ce challenge en un temps court et n’avons peut-être pas abordé des aspects plus poussés du projet. Nous aurions par exemple pu rechercher d’autres sites, appliquer la méthode de web-scraping sur ces sites et joindre les bases issues des différents sites utilisés.