**Projet d’évaluation : Machine Learning**

**4 Groupes de 5 :**

Groupe 1 : Radia, Maxence, Philippe B, Martial et David

Groupe 2 : Catherine, Bachir, Samir, Billel et Kamus

Groupe 3 : Celia, Guillaume, Hugh, Philippe P et Soukeye

Groupe 4 : Chloé, Boris, Anthony, Sami, Arnaud

**Rendu :** Vos scripts (avec docstrings) et un mail tous les lundi matin à 9h30 intitulé « Dev & data job market analysis – dd/mm/yyyy » sur l’adresse [bflevet.ext@simplon.co](mailto:bflevet.ext@simplon.co)

avec le document « Job Market Analysis »

**Deadline :** Fin mars

**Votre mission :**

1/ Faire un script de scraping sur indeed qui permette de spécifier à l’user le type d’annonces qu’il souhaite récupérer :

* Métier (développeur, data scientist…)
* Type de contrat recherché (CDI, CDD, freelance…)
* Lieu de recherche (Paris, Toulouse, …)

Les infos à scraper :

* Titre
* Nom de la boite
* Adresse
* Salaire
* Descriptif du poste
* Date de publication de l’annonce

Vous allez vous concentrer uniquement sur les annonces :

* Métiers : développeur, data scientist, data analyst, business intelligence.
* Localisation : Paris, Lyon, Toulouse, Nantes et Bordeaux.
* Type de contrat : tous

2/ Prévoir un script qui permette de créer puis de stocker automatiquement les infos scrapées dans une bdd Mongo (le script devra prendre en compte le fait de remplacer ou de ne pas tenir compte d’une annonce si cette dernière est déjà dans la bdd).

3/ Récupérer les annonces pour lesquelles on a un salaire. (Il faudra un peu cleaner)

Sur ces annonces l’objectif est de prédire le salaire en fonction des features : localisation et/ou titre et/ou descriptif (à vous de tester ce qui est pertinent)

Pour ce faire, vous devrez utiliser (et comparer) le modèle random forest et kernel rbf.

4/ Ajouter dans la bdd les champs salaire\_forest et salaire\_rbf pour chaque annonce.

Sur les annonces pour lesquels il n’y a pas de salaire, déduire et compléter ces 2 champs, pour les autres laisser ces 2 champs vides.

5/ Faire un script qui permette d’actualiser automatiquement la bdd toutes les semaines. (ie scrap automatiquement les nouvelles annonces liées à nos critères de recherche + déduit salaire\_forest & salaire\_rbf)

6/ Je suis CEO d’une boite qui s’occupe de faire des statistiques sur l’emploi dans le secteur du développement informatique et de la data à Paris, Lyon, Toulouse, Nantes et Bordeaux.

Toutes les semaines, j’aimerai recevoir un document sur ma boite mail comprenant différentes statistiques / graphiques pertinents sur ces marchés.

7/ BONUS 1: Pour chacune des entreprises scrapées, créer un script qui permette de récupérer sur LinkedIn toutes les infos disponibles sur ces dernières (taille de l’entreprise, spécialité, adresse mail, site, description, nombre d’employés etc.).

Ajouter ces informations dans la bdd.

ATTENTION DE NE PAS VOUS FAIRE BAN VOTRE COMPTE (Tips : Avec Selenium, chercher l’entreprise dans les suggestions et non sur la page de résultat)

8/ BONUS 2 : Avec le nom de l’entreprise (ou l’url du site récupéré via profil LinkedIn), rechercher sur le site de l’entreprise le mail générique ainsi que leurs potentiels recherche de poste (onglet « On recrute », « Nous rejoindre » etc..).

Compléter la bdd mongo avec ces informations.

Pour push votre projet :

* <https://towardsdatascience.com/learn-enough-git-to-be-useful-281561eef959>
* <https://towardsdatascience.com/build-your-first-open-source-python-project-53471c9942a7>

A vous de jouer !!