**Nom et Prénom : BOUYAHYAOUI Soufiane**

# Test d'évaluation candidature Data Engineer spécialisé en web scraping

## Section 1 : Extraction de données de comptes Twitter sans utiliser l'API de Twitter

**Réponse :**

Pour résoudre cet exercice de web scraping et extraire les données des comptes Twitter sans utiliser l'API de Twitter, on peut suivre les étapes et la stratégie d'attaque suivantes :

**Étape 1 :** Installation des bibliothèques nécessaires

- Installer Python sur notre système.
- Utiliser pip pour installer les bibliothèques suivantes : BeautifulSoup, requests et pandas.

**Requests :** La bibliothèque requests sert à envoyer des requêtes HTTP en utilisant Python. Elle est simple et facile à utiliser avec de nombreuses fonctionnalités allant de la transmission de paramètres dans les URL à l'envoi d'en-têtes personnalisés et à la vérification SSL.



**BeautifulSoup :** BeautifulSoup fournit des méthodes simples pour naviguer, rechercher et modifier un arbre d'analyse dans des fichiers HTML ou XML. Il transforme un document HTML complexe en un arbre d'objets Python. Il convertit aussi automatiquement le document en Unicode, de sorte que vous n'avez pas à penser aux encodages. Cet outil vous aide non seulement à scraper, mais aussi à nettoyer les données. BeautifulSoup prend en charge l'analyseur HTML inclus dans la bibliothèque standard de Python, mais aussi plusieurs analyseurs Python tiers comme lxml ou html5lib.


**Pandas :** La bibliothèque logicielle open-source Pandas est spécifiquement conçue pour la manipulation et l'analyse de données en langage Python. Elle est à la fois performante, flexible et simple d'utilisation.


**Étape 2 :** Analyse de la structure de la page Twitter

- Ouvrir la page Twitter d'un compte spécifique que nous souhaitons extraire.
- Utiliser l'inspecteur d'éléments de notre navigateur pour analyser la structure HTML de la page.
- Identifier les balises HTML contenant les données que nous souhaitons extraire, telles que les tweets, les métriques, etc.

**Étape 3 :** Récupération des données des tweets

- Utiliser la bibliothèque requests pour envoyer une requête HTTP GET à la page du compte Twitter.
- Analyser la réponse HTML en utilisant BeautifulSoup pour extraire les balises contenant les tweets.
- Parcourir les balises des tweets pour extraire le contenu textuel, les médias, les liens, les documents et les hashtags.

**Étape 4 :** Extraction des métriques du compte Twitter

- Rechercher les balises HTML contenant les informations sur les métriques du compte, telles que le nombre de followers et son évolution sur 7 jours.

- Utiliser BeautifulSoup pour extraire ces métriques et les enregistrer dans une structure de données appropriée.

**Étape 5 :** Extraction des métriques des tweets

- Rechercher les balises HTML contenant les informations sur les métriques des tweets, telles que le nombre de likes, de retweets et de partages.
- Utiliser BeautifulSoup pour extraire ces métriques pour chaque tweet et les enregistrer dans une structure de données appropriée.

**Étape 6 :** Suivi des métriques dans le temps

- Pour suivre l'évolution des métriques dans le temps, on doit collecter les données sur une période de 7 jours.
- On va répéter les étapes 3 à 5 quotidiennement pour extraire les nouvelles métriques et les ajouter à notre ensemble de données existant.
- Utiliser la bibliothèque pandas pour stocker les données dans un dataframe et analyser les tendances au fil du temps.

**Section 2 : Web Scraping et traitement de données**

**3) Un schéma qui montre le stockage des données extraites dans une database RavenDB et comment servir les données via l'API :**