DATA ENGINEER EN WEB SCRAPING

Section 1 :

Étape 1 : Installation des bibliothèques

Nous installons les bibliothèques nécessaires pour extraire automatiquement les données depuis les comtes de twitter tels que la bibliothèque BeautifulSoup et requests.

Étape 2 : Analyse de la structure de la page

- Ouvrons la page du compte Twitter depuis lequel nous souhaitons extrait de données dans votre navigateur.
- Utilisons les outils de développement du navigateur pour inspecter les éléments de la page.
- Identifions les balises HTML et les classes CSS qui contiennent les données que nous souhaitons extraire, comme le contenu des tweets, les métriques du compte et les métriques des tweets.

Étape 3 : Récupération du contenu de la page

- Utilisons la bibliothèque requests pour envoyer une requête HTTP GET à l'URL du compte Twitter.
- Capturons la réponse et stockons-la dans une variable.

Étape 4 : Analyse du contenu de la page

- Utilisons la bibliothèque BeautifulSoup pour analyser le contenu HTML de la page.
- Utilisons les sélecteurs CSS ou les méthodes de recherche de BeautifulSoup pour extraire les éléments pertinents, tels que les tweets, les métriques du compte et les métriques des tweets.

- Par exemple, nous pouvons utiliser des sélecteurs CSS comme soup.select('.tweet-text') pour récupérer le contenu des tweets.

Étape 5 : Extraction des données

- Parcourons les éléments extraits et extrayons les données souhaitées, telles que le texte des tweets, les médias, les liens, les hashtags, les métriques du compte, etc.
- Organisons les données extraites dans une structure de données appropriée, telle qu'une liste de dictionnaires ou un tableau.

Enregistrement de données

Dès que les données extraites soient disponibles nous allons l'enregistrer dans un dossier spécifique à chaque compte où on va trouver un fichier csv ou Json contenant les informations textuelles ou numériques et pour les documents et les vidéos concernées nous allons l'enregistrer directement dans le même dossier.