TP Thème 3 - Apache Logs – Analyse d'erreurs 404

Objectif : Extraire, traiter et visualiser les IPs responsables d'erreurs 404.

Contexte :

Un fichier **access.log** vous a été remis par l'administrateur système. Il contient les requêtes HTTP d'un serveur Apache. Vous devez l'analyser pour identifier les IPs générant le plus d'erreurs 404.

Étapes du TP :

1. **Chargement et parsing du fichier**
    - Charger **access.log** dans un **DataFrame**.
    - Extraire les colonnes suivantes : **ip**, **datetime**, **method**, **url**, **status**, **user_agent**.
    - Nettoyer ou ignorer les lignes malformées.
2. **Filtrage des erreurs 404**
    - Isoler les lignes où le status est **404**.
3. **Top 5 des IPs fautives**
    - Grouper par IP.
    - Compter et trier pour afficher les 5 IPs générant le plus d'erreurs.
4. **Visualisation**
    - Créer un histogramme (bar chart) avec **matplotlib**.
        - **Axe X** : IPs
        - **Axe Y** : Nombre d'erreurs 404
        - Personnaliser le graphique (titre, couleurs, labels).

Bonus : Détection de bots

- Filtrer les lignes dont le **user_agent** contient "bot", "crawler" ou "spider".
- Identifier les IPs suspectes.
- Calculer le pourcentage d'erreurs 404 provenant de bots.

Résultat attendu

- Un script Python contenant :
    - Le code structuré (fonctions recommandées).

    - Des aperçus intermédiaires (print, head(), etc.).
    - Un graphique généré.
    - Une discussion des résultats (actions possibles ?).

Discussion finale

- Quelles conclusions peut-on tirer ?
- Ces IPs doivent-elles être bannies ?
- Peut-on automatiser ce type de détection ?