IAE Paris Sorbonne Business School

UE9 Scrapping de données

Groupe 2

Guide de Scraping des Données

1. Problèmes Rencontrés

a. Difficultés Techniques

- Blocages fréquents (CAPTCHAs, restrictions API, détection de scraping par les sites médias).
- Limites d'outils standards comme Octoparse sur les contenus complexes ou non structurés.
- Impossibilité d'accéder directement à certaines API de sites médias.

b. Obstacles Légaux et Éthiques

- Conformité aux conditions d'utilisation et au RGPD
- Sensibilité accrue des données collectées, notamment pour des sujets comme le complotisme

2. Organisation et Méthodes

a. Division par Sources

- Organisation des tâches en fonction des types de plateformes : YouTube, Reddit, médias en ligne
- Focus sur les sources critiques et spécifiques (contenus vidéo, discussions).

b. Optimisation des Données

- Utilisation de **Keywords de filtrage** pour prioriser les données pertinentes
- Nettoyage en amont pour éviter les doublons.

c. Approches Outils

- **Octoparse :** Développement de templates personnalisés (pagination avancée, boucles imbriquées) pour les plateformes comme YouTube et Reddit, où il est efficace
- **Python :** Développement de scripts avancés pour les sites médias nécessitant des solutions plus flexibles et robustes.

3. Bonnes Pratiques

- Respecter les règles des plateformes (conditions d'utilisation, RGPD).
- Déployer des **techniques avancées en Python** :
    a. Rotation des proxies et des User-Agents pour éviter les blocages

    b. Utilisation d'API spécialisées comme **News API** ou **ScraperAPI** pour des extractions fiables
    c. Automatisation via des interfaces utilisateurs (IHM) pour simplifier l'expérience
- Former les équipes à:
    a. Reconnaître les limites légales et techniques
    b. Adapter les outils à chaque type de source

4. Limites d'Octoparse

- **Efficace** pour les plateformes comme YouTube grâce à des templates optimisés
- **Limité** pour les sites medias:
    a. Détection facile des processus automatisés.
    b. Incapacité à gérer les interactions dynamiques ou les API.
- **Solution Alternative :** Utilisation de scripts Python sur mesure, plus flexibles, permettant d'intégrer des stratégies avancées.

5. Recommandations

a. **Automatisation Interne :** Développer une solution Python robuste pour le scraping complexe
b. **Surveillance Continue :** Mettre en place un processus d'analyse automatisée pour détecter les discours critiques
c. **Équipes Formées :** Sensibiliser les data scientists aux contraintes légales et aux techniques modernes de scraping.