SAÉ3.VCOD.01 – COLLECTE AUTOMATISÉE DE DONNÉES WEB

Boubou Thiam NIANG

Pour l'année 2023-24

COLLECTES AUTOMATISEE DE DONNEE WEB – SCRAPING

- Nom couramment utilisé
  - Scraping web (grattage web)

SCRAPING WEB

- Définition:
  - technique automatisée permettant d'extraire des informations à partir de sites web

- En d’autres termes
  - Simule la recherche d’info qu’on effectuerait manuellement

SCRAPING WEB

- Définition:
  - technique automatisée permettant d'extraire des informations à partir de sites web

- En d'autres termes
  - Simule la recherche d'info qu'on effectuerait manuellement

- Quelques cas d'usages:
  - **Scraper des actualités**

  Collecter des titres, des articles et des informations à partir de sites d'actualités populaires.

  - **Analyse de prix en ligne :**

  Créer un scraper pour collecter les prix de produits sur plusieurs sites de commerce électronique

  - **Suivi des médias sociaux :**

  Récupérer des données à partir de profils sociaux comme Twitter, Instagram ou Facebook pour analyser les tendances, la fréquence des publications ou les interactions.

Le choix de l'alternance
IUT Lumière Lyon 2

SCRAPING WEB

- Définition:
  - technique automatisée permettant d'extraire des informations à partir de sites web

- En d’autres termes
  - Simule la recherche d’info qu’on effectuerait manuellement

- Quelques cas d’usages:
  - **Scraping de données météorologiques**

  Collecter des informations sur les conditions météorologiques à partir de différents sites pour fournir des prévisions ou des analyses.

  - **Suivi des offres d'emploi :**

  Scraping des sites d'emploi pour collecter des offres correspondant à certains critères, les trier ou les analyser.

  - **Statistique des joueurs de foot**

  - **Tendance jeux vidéo**

SCRAPING WEB – LES ÉTAPES

- **Identification des Sources de Données**
  - Déterminez les Sites Cibles
- **Collecte des Informations**
  - Accès aux Pages Web
  - Récupération du Contenu HTML
- **Analyse HTML**
  - Analyse Structurée à l'aide de librairie pour identifier les balises contenant les données à extraire.
- **Extraction des Données**
  - Sélection des Éléments:

  Identifiez les éléments HTML (balises, classes, ID) qui contiennent les données recherchées.
  - Parsing pour extraire les éléments
- **Nettoyage et Prétraitement**
  - Élimination des Données Redondantes : Supprimez les balises HTML, les caractères spéciaux ou autres éléments indésirables.
  - Normalisation des Données : Mettez en forme les données extraites pour qu'elles soient cohérentes et utilisables.
- **Stockage des Données**
  - Choix de la Méthode de Stockage :

  Enregistrez les données extraites dans une base de données, un fichier CSV, un JSON, ou tout autre format adapté à vos besoins.

PRÉREQUIS – DOM ((DOCUMENT OBJECT MODEL)

- **Quoi:**

représentation hiérarchique et structurée d'une page web

- **Principe**
  - Hiérarchie Structurée :

    Le DOM représente une page web comme une structure arborescente où chaque élément HTML est un nœud (node).

  - **Relation Parent-Enfant :**

    Les éléments HTML sont organisés en une structure parent-enfant.

- **Accès aux Éléments :**

  Le DOM permet d'accéder aux éléments individuels de la page web via des méthodes et des propriétés spécifiques.

- **Modifications Dynamiques :** En utilisant le DOM, on peut ajouter, supprimer ou modifier des éléments, du contenu, des styles et des attributs d'une page web en temps réel.

PRÉREQUIS – DOM ((DOCUMENT OBJECT MODEL)

- **Exemple DOM**

```
<!DOCTYPE html>
<html>
<head>
  <title>Exemple DOM</title>
</head>
<body>
  <div id="conteneur">
    <h1>Titre</h1>
    <p>Un paragraphe</p>
  </div>
</body>
</html>
```

LES LIBRAIRIES JAVA – POUR LE SCRPING

- Quelques librairies Java
  - HtmlUnit (utilisé dans ce TD)
    - https://htmlunit.sourceforge.io/
  - Jsoup
    - https://jsoup.org/
  - WebMagic
    - https://webmagic.io/en/
- Vous pouvez explorer d'autre
  - Exemple de comparatif (soyez curieux)
    - https://www.zenrows.com/blog/java-web-scraping-library#htmleasy

AVERTISSEMENT – LÉGALITÉ - ATTENTION

- Légalité
  - Considérer page avec les données en accès libre (sans login)
  - Vérifier le fichier systématiquement les fichier **robot.txt** s'il y en a
    - Exemple en live:

    https://www.google.com/robots.txt

    https://www.facebook.com/robots.txt

    https://twitter.com/robots.txt
  - Lire les conditions générales s'il y 'en a

EXEMPLE PRATIQUE DE SCRPING AVEC HTMLUNIT

- En live
  - Site https://new.uschess.org/player-search

PROJET

- Formation de binôme/trinôme
- Choix d’une thématique
- Etude: qu’est-ce qu’on cherche
- Identification des sites
- Développement
- Stockage
- Analyse – visualisation

PROJET - RENSU

- Document word (2 pages environ)
  - Introduction
    - Motivation de la thématique choisie
  - Liste des sites web choisit
  - Un paragraphe sur la légalité des sites choisit
  - Résultats
  - Conclusion suite à l'analyse
- Choix de la librairie
- Code source
- Fichier readme avec les étapes pour reproduire