المدرسة الوطنية للعلوم التطبيقية بتطوان
ⵜⵉⵎⵙⵙⵓⵔⵜ ⵜⴰⵏⴰⵎⵓⵔⵜ ⵏ ⵜⵙⵏⵙⵉⵡⵉⵏ ⵜⵉⵜⵜⴰⵡⵉⵏⵉⵏ
Ecole Nationale des Sciences Appliquées de Tétouan

Université Abdelmalek Essaâdi

École Nationale des Sciences Appliquées de Tétouan

Filière : Big Data & IA

Rapport du Projet Analyse de Web

Sujet : Crawler web pour extraire des informations sur les clients des restaurants à utiliser pour la publicité

Réalisé par :
Hamza Kholti

Proposé par :
Mr. Imad Sassi

Année Universitaire : 2024-2025

Table des matières

Table des figures

1 Introduction

Le web scraping est une technique permettant d'extraire automatiquement des données depuis des sites web, offrant ainsi un accès structuré à des informations non disponibles sous forme de bases de données. Dans le cadre de ce projet, nous utilisons le web scraping pour collecter des informations sur les clients des restaurants d'une ville spécifique, afin d'analyser leur origine et leur activité en ligne.

L'objectif principal est de développer un crawler web capable d'extraire des données depuis la plateforme d'avis en ligne : TripAdvisor. Ce crawler parcourra les 500 meilleurs restaurants de la ville Los Angeles , identifiera les utilisateurs ayant laissé des avis, et déterminera leur nationalité. L'analyse de ces données permettra de classer les nationalités les plus actives dans la ville sélectionnée, offrant ainsi des insights précieux sur le profil des clients et leur engagement envers la scène gastronomique locale.

Ce projet peut être particulièrement utile pour des applications marketing et publicitaires, en permettant aux restaurateurs et aux professionnels du secteur de mieux comprendre leur clientèle et d'adapter leurs stratégies en conséquence. Le rapport associé détaillera les différentes étapes suivies, les choix techniques effectués, ainsi que les résultats obtenus en termes de classement des nationalités et de performances du crawler.

2 Outils Utilisés

Pour mener à bien ce projet de web scraping et d'analyse des avis sur les restaurants de Los Angeles, plusieurs outils et bibliothèques ont été utilisés. Chacun joue un rôle essentiel dans l'extraction, le traitement et la visualisation des données.

2.1 Python – Langage de Programmation

Python a été choisi pour ce projet en raison de sa richesse en bibliothèques dédiées au **web scraping**, au **traitement des données** et à la **visualisation**. Il offre une syntaxe simple et des outils puissants qui facilitent l'extraction et l'analyse des avis sur TripAdvisor.

2.2 BeautifulSoup – Analyse du Contenu HTML

Utilité :

- Extraire les données des pages HTML obtenues depuis TripAdvisor.
- Parser et structurer les avis, notes et noms des utilisateurs.

Fonctionnement :

- BeautifulSoup permet de naviguer dans la structure du HTML et d'extraire uniquement les informations pertinentes (ex : avis, utilisateurs, notes).
- Il fonctionne bien en complément de **requests**, mais comme TripAdvisor bloque souvent les requêtes directes, **Selenium** est utilisé à la place.

2.3 Selenium – Automatisation du Scraping

Utilité :

- Contourner les protections de TripAdvisor qui empêchent l'accès aux données via `requests`.
- Simuler la navigation humaine pour éviter d'être détecté comme un bot.
- Cliquer sur les boutons, charger dynamiquement le contenu et extraire les avis.

Fonctionnement :

- Selenium ouvre un navigateur (Chrome ou Firefox).
- Il navigue **automatiquement** sur TripAdvisor, charge les pages et récupère les informations affichées.
- Un **délai aléatoire** est inséré entre les actions pour imiter un comportement humain et éviter d'être bloqué.

2.4 Google Colab – Exécution du Code en Cloud

Pourquoi Google Colab ?

- **Éviter le blocage local** : TripAdvisor bloque souvent les adresses IP locales après plusieurs requêtes.
- **Utiliser un environnement distant** : Google Colab exécute le code sur des serveurs distants, rendant l'IP différente et réduisant le risque de blocage.

2.5 Pandas – Traitement des Données Collectées

Utilité :

- Stocker et organiser les données extraites dans des **DataFrames**.
- Nettoyer les données (*suppression des doublons, gestion des valeurs manquantes*).
- Effectuer des jointures entre les datasets des restaurants et des utilisateurs.

2.6 Matplotlib et Seaborn – Visualisation des Données

Utilité :

- Générer des graphiques pour analyser les données collectées.
- Visualiser le **classement des nationalités** les plus actives dans les restaurants de Los Angeles.
- Créer des **histogrammes** et **graphiques en barres** pour mieux interpréter les tendances.

2.7 Groq API avec LLaMA LLM – Détection Automatique des Pays

Utilité :

- Associer automatiquement un **pays** à chaque utilisateur en fonction de sa localisation.
- Utiliser un **modèle d'intelligence artificielle (LLaMA)** pour comprendre et extraire le pays à partir du texte brut.

Fonctionnement :

- Envoi d'une requête à l'API Groq avec une **instruction précise** : *"Retourne uniquement le pays correspondant à cette localisation ou 'unknown' si ce n'est pas un lieu."*
- Stockage du pays détecté dans la colonne `country` du dataset.

Grâce à ces outils, nous avons pu extraire, nettoyer et analyser efficacement les avis des restaurants de Los Angeles sur TripAdvisor, tout en surmontant les défis liés au blocage des requêtes et à l'identification des pays des utilisateurs.

3 Stratégie de Scraping

Le scraping de données sur TripAdvisor présente plusieurs défis, notamment la protection contre l'extraction automatique, le chargement dynamique du contenu et le risque de blocage après plusieurs requêtes successives. Pour surmonter ces obstacles, plusieurs stratégies ont été mises en place afin d'assurer une extraction efficace et continue des données.

3.1 Automatisation et Navigation Dynamique

Le contenu des pages de TripAdvisor est souvent chargé de manière dynamique, ce qui rend impossible l'extraction des informations avec des requêtes classiques. Pour contourner cette difficulté, nous avons utilisé une navigation automatisée permettant de :

- Charger les pages des restaurants en simulant un comportement humain.
- Cliquer sur les boutons "Page suivante" pour extraire les 500 meilleurs restaurants de la ville.
- Attendre dynamiquement le chargement complet du contenu avant de récupérer les données.

3.2 Gestion des Bloqueurs Anti-Scraping

Les sites comme TripAdvisor détectent les comportements suspects liés au scraping en surveillant la fréquence des requêtes et les signatures des navigateurs. Pour éviter d'être bloqué, plusieurs mesures ont été mises en place :

- **Utilisation de Selenium** pour simuler un utilisateur réel naviguant sur le site.
- **Rotation des User-Agents** : Simulation de différents navigateurs pour masquer l'automatisation.
- **Délais aléatoires** entre les requêtes afin d'éviter un comportement suspect (imitation du temps de lecture humain).
- **Exécution sur Google Colab** pour contourner le blocage des adresses IP locales en utilisant des serveurs distants.

3.3 Stockage Progressif et Reprise en Cas d'Interruption

Un scraping de grande envergure peut être interrompu à tout moment par des erreurs réseau, des blocages ou des coupures de session. Pour éviter de perdre les données déjà collectées, nous avons mis en place une stratégie de stockage progressif :

- Sauvegarde des restaurants extraits dans un fichier CSV après chaque page traitée.
- Enregistrement du dernier restaurant traité dans un fichier de suivi, permettant de reprendre l'extraction en cas d'interruption.
- Écriture progressive des avis des utilisateurs pour garantir qu'aucune donnée ne soit perdue.

3.4 Identification Automatique des Pays

TripAdvisor ne fournit pas directement la nationalité des utilisateurs, ce qui complique l'analyse de leur origine. Pour résoudre ce problème, nous avons adopté une approche basée sur l'intelligence artificielle :

- Extraction des localisations des utilisateurs à partir des avis collectés.
- Utilisation d'un modèle LLaMA via l'API Groq pour identifier le pays correspondant à chaque localisation.
- Stockage direct des pays détectés dans le dataset afin de faciliter l'analyse.

3.5 Fusion et Visualisation des Données

Une fois les données extraites et nettoyées, elles sont fusionnées pour permettre des analyses approfondies :

- Association des avis aux restaurants correspondants.
- Suppression des doublons et correction des incohérences.
- Visualisation des nationalités les plus actives sous forme d'histogrammes et de graphiques interactifs.

Grâce à ces stratégies, nous avons pu extraire et structurer les avis des restaurants de Los Angeles tout en minimisant les risques de blocage et en garantissant une collecte fiable et continue des données.

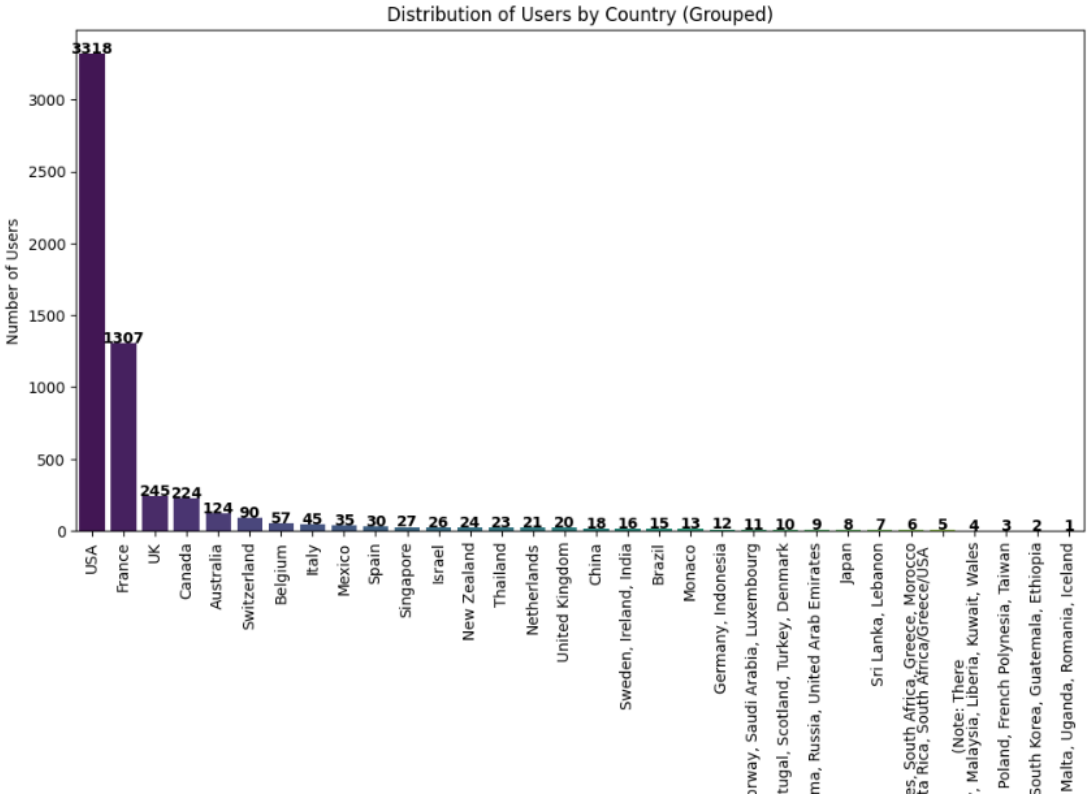

FIGURE 1 – Distribution des utilisateurs par pays

4 Analyse des Resultats

4.1 Analyse du graphique

La figure 1 illustre la répartition des utilisateurs par pays dans les avis recueillis. Voici les principaux constats que l'on peut tirer de cette visualisation :

- **Fort volume d'utilisateurs des États-Unis (USA) :** Avec plus de 3000 utilisateurs recensés, les États-Unis se distinguent très nettement comme la nationalité la plus active dans les avis sur les restaurants de Los Angeles. Cela peut s'expliquer par le fait que Los Angeles se trouve aux États-Unis, attirant naturellement une majorité de clients locaux.
- **Présence significative de la France :** Plus de 1000 utilisateurs identifiés comme Français se classent en deuxième position. Cette forte représentation suggère que Los Angeles est une destination populaire pour les voyageurs français, ou que la communauté francophone y est bien implantée.
- **Canada, Suisse, Espagne et Belgique :** Ces pays apparaissent avec quelques centaines d'utilisateurs. Bien que leur nombre soit moins élevé que celui des États-Unis et de la France, ils restent néanmoins des nationalités notables dans les avis.
- **Diversité des autres pays :** Au-delà du top 5, on observe une multitude de nationalités (Singapour, Chine, Brésil, Malaisie, Allemagne, etc.), témoignant d'une clientèle internationale variée à Los Angeles.
- **Long tail** : La présence de pays moins représentés, avec seulement quelques dizaines d'utilisateurs, indique que Los Angeles attire également des touristes ou résidents d'origines très diverses, même si ces groupes restent minoritaires.

En conclusion, ce graphique met en évidence la dimension internationale de la clientèle

des restaurants de Los Angeles, avec une nette prédominance des utilisateurs américains et une forte présence de touristes ou expatriés français. Les données collectées montrent ainsi la nécessité de stratégies marketing adaptées à plusieurs nationalités, notamment américaine et européenne, afin de mieux répondre aux attentes de cette clientèle internationale.

5 Structure du Répertoire

```
Project_TripAdvisor_scraping/
    Project_TripAdvisor_scraping/
        .env
        cache.json
        ColabNotbook.ipynb
        country_histogram.png
        country_histogram_grouped.png
        Execute.py
        last_page.txt
        last_processed.txt
        last_restaurant.txt
        NationalityClassement.ipynb
        Restaurants.csv
        Scraper_Functions.py
        user.csv
        WebAnalysisReport.pdf
```

la description de chaque fichier :

- **Scraper_Functions.py** : Fichier contenant les fonctions Python utilisées pour le scraping.
- **Execute.py** : Script principal pour lancer le scraping et gérer l'extraction des données.
- **ColabNotbook.ipynb** : Notebook Jupyter utilisé pour exécuter scraping sur Google Colab en cas de block pour changer l'adress IP.
- **NationalityClassement.ipynb** : Notebook Jupyter analysant la répartition des nationalités des utilisateurs.
- **last_page.txt** : Sauvegarde le numéro de la dernière page traitée pour reprendre en cas d'interruption.
- **last_processed.txt** : Stocke la dernière ligne de données traitée pour éviter les doublons.
- **last_restaurant.txt** : Mémorise le dernier restaurant extrait pour reprendre le scraping.
- **Restaurants.csv** : Fichier contenant les données extraites des restaurants.
- **user.csv** : Fichier contenant les informations des utilisateurs extraits.
- **WebAnalysisReport.pdf** : Rapport d'analyse sur la structure du site TripAdvisor et les techniques de scraping utilisées.
- **.env** : Contient les variables d'environnement (ex. clés API, configurations).
- **cache.json** : Stocke temporairement certaines données pour éviter de refaire les mêmes requêtes.

— **country_histogram.png** : Histogramme des nationalités des utilisateurs extraits.
— **country_histogram_grouped.png** : Version groupée de l'histogramme des nationalités.