المدرسة الوطنية للعلوم التطبيقية بتطوان
ⵜⵉⵙⴽⵔⵉⵜ ⵜⴰⵏⴰⵎⵓⵔⵜ ⵏ ⵜⵙⵙⵏⵉⵡⵉⵏ ⵜⵉⵜⵡⴰⵏⵉⵏ
Ecole Nationale des Sciences Appliquées de Tétouan

RAPPORT
PROJET WEB SCRAPING TRIPADVISOR

Prepare par :
Ayoub Anhal

Realise par :
Pr. Sassi

SOMMAIRE

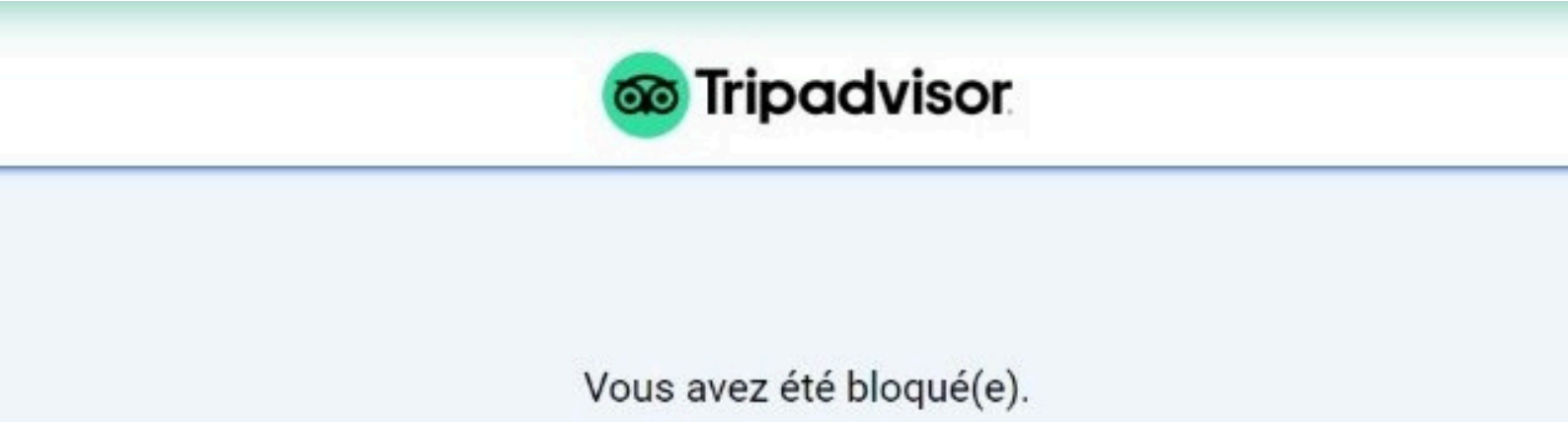
PROBLÉMATIQUE

Lorsque vous tentez d'accéder au site de **TripAdvisor**, un code d'**erreur 403** peut apparaître, indiquant que l'accès vous est **interdit**. Ce message résulte des mesures **anti-bot** mises en place par le site, qui visent à protéger son contenu contre le **scraping et l'automatisation abusive**. **TripAdvisor** analyse notamment la fréquence des requêtes, les **en-têtes HTTP** envoyés et le comportement de navigation pour déterminer si la demande provient d'un utilisateur réel ou d'un script automatisé.

Dès qu'un comportement inhabituel est détecté, par exemple un trop grand nombre de requêtes sur une courte période ou l'absence d'informations standards comme un agent utilisateur valide, le serveur refuse l'accès et renvoie l'erreur 403. Cette étape de contrôle est cruciale pour préserver la sécurité du site et garantir une expérience utilisateur fiable en empêchant l'exploitation excessive de ses ressources.

OUTILS DE SCRAPING

BeautifulSoup est souvent privilégié pour le scraping de données car il offre une **solution légère et rapide** pour extraire le contenu **HTML** d'une page de manière statique. **Contrairement** à des outils comme **Selenium**, qui simulent un navigateur complet et nécessitent de gérer des **interactions dynamiques (clics, exécutions de scripts, etc.)**, **BeautifulSoup** se contente de parser le code source sans déclencher d'actions susceptibles d'attirer l'attention des systèmes anti-bot. Cette approche **réduit le risque** de détection et permet de **gagner un temps** précieux lors de **l'extraction des données**, même si, dans certains cas, des mécanismes de sécurité peuvent finir par détecter l'automatisation après environ une heure d'activité.



TOP 500 RESTAURANTS

Pour commencer, l'extraction des données des **Top 500 restaurants** est une étape **simple et rapide**. En utilisant **BeautifulSoup**, on peut parcourir **les pages web** et récupérer les informations essentielles **sans nécessiter** d'interactions **dynamiques**, comme les **clics**, ce qui **minimise les risques** d'être détecté par les systèmes **anti-bot** et permet d'économiser du temps.

CODE : TOP 500 RESTAURANTS

Les pages de restaurants se forment selon le modèle **URL :**

1. **"**https://www.tripadvisor.com/Restaurants-g293916-oa{}-zft20693-Bangkok.html**"**, où l'offset est **"0" pour la première page** et **"30" pour la deuxième.**
2. Chaque **page** affiche **30 restaurants**, ce qui signifie qu'il faut environ **17 pages** pour extraire **500 entrées**.
3. Le **code envoie** une **requête HTTP** avec des en-têtes personnalisés pour simuler un navigateur et éviter la détection anti-bot , si le reponse **requête HTTP** egale **200** alors **bien connextion** sinon **erreur 403** false de connextion.
4. **BeautifulSoup** est utilisé pour **parser le HTML** et rechercher les balises contenant les noms et URL des restaurants.
5. Pour c**haque restaurant** détecté, un dictionnaire avec **son nom** et **son URL** est créé et ajouté à une liste.
6. Enfin, la liste est tronquée à **500 restaurants** et sauvegardée dans un **fichier JSON (restaurants.json)** pour une exploitation ultérieure.

```
Scraping 1...
Scraping 2...
Scraping 3...
Scraping 4...
Scraping 5...
Scraping 6...
Scraping 7...
Scraping 8...
Scraping 9...
Scraping 10...
Scraping 11...
Scraping 12...
Scraping 13...
Scraping 14...
Scraping 15...
Scraping 16...
Scraping 17...
Fin resturants.
Les informations des restaurants ont été enregistrées dans le fichier TP.json.
```


```
{
    "nom Restaurant": "1. SEEN Restaurant & Bar Bangkok",
    "url Restaurant": "https://www.tripadvisor.com/Restaurant_Review-g293916-d10326104-Reviews-SEEN_Restaurant_Bar_Bangkok-Bangkok.html"
},
{
    "nom Restaurant": "2. Spectrum Lounge & Bar",
    "url Restaurant": "https://www.tripadvisor.com/Restaurant_Review-g293916-d16726460-Reviews-Spectrum_Lounge_Bar-Bangkok.html"
},
{
    "nom Restaurant": "3. Blue Sky Rooftop Bar",
    "url Restaurant": "https://www.tripadvisor.com/Restaurant_Review-g293916-d3468507-Reviews-Blue_Sky_Rooftop_Bar-Bangkok.html"
},
```

UTILISATEURS & NATIONALITE

Dans la partie d'extraction, nous ne récupérons pas les commentaires eux-mêmes mais uniquement **les noms des utilisateurs** ayant commenté chaque restaurant, ainsi que leur **nationalité**. Cette étape est particulièrement **chronophage** et, malgré plusieurs ajustements comme **la modification des en-têtes HTTP** pour contourner les mesures **anti-bot**, nous rencontrons régulièrement **des erreurs 403** qui ralentissent l'extraction des données.

SOLUTION : WEB UNBLOCKER D'OXYLABS

Le **Web Unblocker d'Oxylabs** est une solution avancée conçue pour faciliter l'extraction de données à grande échelle en contournant **les systèmes anti-bots sophistiqués**. Cette technologie permet aux entreprises d'accéder aux données publiques de sites web complexes en se faisant passer pour de véritables utilisateurs, assurant ainsi une collecte de données fluide et efficace.
En pratique, le Web Unblocker d'Oxylabs fonctionne en gérant **automatiquement les paramètres tels que les cookies**, **les en-têtes HTTP** et **le rendu JavaScript**. Cela permet aux utilisateurs d'accéder aux données publiques sans se soucier des blocages ou des captchas, même sur des sites web complexes.
De plus, Oxylabs offre un vaste pool de plus de **100 millions d'adresses IP** résidentielles réparties dans **195 pays**, ce qui garantit une couverture mondiale et une rotation efficace des proxys.
Oxylabs Web Unblocker en utilisant une authentification basée sur un nom d'utilisateur et un mot de passe. Voici comment cela fonctionne :

- **Définition des Identifiants d'Authentification :**

```
USERNAME = '.....'
PASSWORD = '.....'
```

Ces variables contiennent les informations de connexion fournies par Oxylabs. Elles sont nécessaires pour s'authentifier auprès du service proxy.

- **Construction de l'URL du Proxy :**

```
PROXY_URL = F'HTTP://{USERNAME}:{PASSWORD}@UNBLOCK.OXYLABS.IO:60000'
```

unblock.oxylabs.io est l'adresse du serveur proxy d'Oxylabs.
60000 est le port utilisé pour accéder au Web Unblocker.
Le format http://USERNAME:PASSWORD@ est une méthode standard pour authentifier un utilisateur lors de la connexion à un proxy.

- **Définition des Proxys pour les Requêtes :**

```
PROXIES = {'HTTP': PROXY_URL, 'HTTPS': PROXY_URL}
```

dictionnaire de proxies qui sera utilisé pour rediriger tout le trafic HTTP et HTTPS via le proxy Web Unblocker.

SOLUTION : WEB UNBLOCKER D'OXYLABS

```
USERNAME = 'ayoub_a9Gyd'
PASSWORD = 'Ayoub1234555_'
PROXY_URL = f'http://{USERNAME}:{PASSWORD}@unblock.oxylabs.io:60000'
proxies = {'http': PROXY_URL, 'https': PROXY_URL}
```

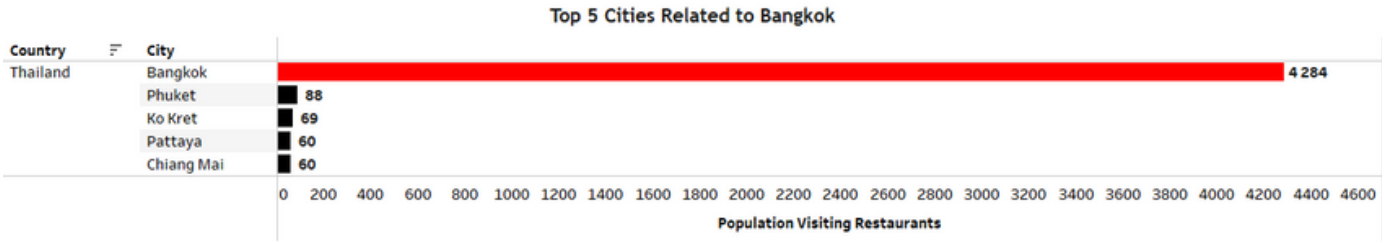
CODE : USERS & NATIONALITES

- **Chargement des restaurants** ,Le script ouvre le fichier **restaurants.json(link restaurants)** et charge les restaurants à scraper. Si un restaurant est déjà traité, il passe au suivant pour éviter les doublons.
- Construction de l'URL ,Pour chaque restaurant, il **génère une URL** de base à partir de url Restaurant, puis ajoute un paramètre pour la pagination **des avis.**
- **Utilisation du proxy** , Les requêtes HTTP passent par **le proxy Oxylabs**, permettant **d'éviter les blocages** et d'accéder aux données même si le site applique des restrictions.
- **Extraction des avis** ,BeautifulSoup **analyse le HTML** et cherche les balises contenant les avis des utilisateurs. Chaque avis est ensuite extrait et structuré sous forme de dictionnaire.
- **Identification du pays** ,La fonction **get_country** utilise Nominatim pour trouver le pays d'un utilisateur en fonction de la ville indiquée dans son profil.
- Gestion des erreurs , Si une requête échoue (ex. site bloqué, problème réseau), le script affiche un message d'erreur et passe au restaurant suivant sans interrompre l'exécution.
- Sauvegarde des données, Chaque nouvel avis est ajouté à **all_reviews.json** pour éviter de perdre les résultats en cas d'arrêt du script.
- Limite de pages et avis ,Le scraping s'arrête après un nombre défini d'avis (max_reviews) ou de **pages (max_pages = 5)** pour **éviter d'envoyer trop de requêtes.**
- Fin du script ,Une fois tous les restaurants traités, le script affiche un message confirmant la fin du scraping et termine l'exécution.

CODE : USERS & NATIONALITES

```
Restaurant 1. SEEN Restaurant & Bar Bangkok déjà traité, passage au suivant.
Restaurant 2. Spectrum Lounge & Bar déjà traité, passage au suivant.
Restaurant 3. Blue Sky Rooftop Bar déjà traité, passage au suivant.
Restaurant 4. Thiptara Thai Restaurant déjà traité, passage au suivant.
Restaurant 5. Sra Bua By Kiin Kiin déjà traité, passage au suivant.
Restaurant 6. Riverside Terrace déjà traité, passage au suivant.
Restaurant 7. RedSquare Rooftop Bar déjà traité, passage au suivant.
Restaurant 8. La Scala déjà traité, passage au suivant.
Restaurant 9. Cocotte Farm Roast & Winery déjà traité, passage au suivant.
Restaurant 10. Vertigo Rooftop Restaurant déjà traité, passage au suivant.
Restaurant 11. Salathip Thai Restaurant déjà traité, passage au suivant.
Restaurant 12. Celadon déjà traité, passage au suivant.
Restaurant 13. Ministry Of Crab - Bangkok déjà traité, passage au suivant.
Restaurant 14. Amaya Food Gallery at Amari Bangkok déjà traité, passage au suivant.
Restaurant 15. ABar Rooftop déjà traité, passage au suivant.
Restaurant 16. Punjab Grill Bangkok déjà traité, passage au suivant.
Restaurant 17. Penthouse Bar + Grill déjà traité, passage au suivant.
Restaurant 18. Mexicano Restaurante Autentico déjà traité, passage au suivant.
Restaurant 19. Goji Kitchen + Bar déjà traité, passage au suivant.
Restaurant 20. Akira Back Bangkok déjà traité, passage au suivant.
Restaurant 21. The Kitchen Table déjà traité, passage au suivant.
Restaurant 22. Scarlett Wine Bar & Restaurant déjà traité, passage au suivant.
Restaurant 23. JAM JAM Eatery & Bar déjà traité, passage au suivant.
Restaurant 24. Pastel Rooftop Bar & Mediterranean Dining déjà traité, passage au suivant.
Restaurant 25. Prego Bangkok déjà traité, passage au suivant.
...
Restaurant 451. Ebisensei déjà traité, passage au suivant.
Restaurant 452. Vecho déjà traité, passage au suivant.
Restaurant 453. House Of Tango déjà traité, passage au suivant.
Scraping terminé pour tous les restaurants !
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

DASH

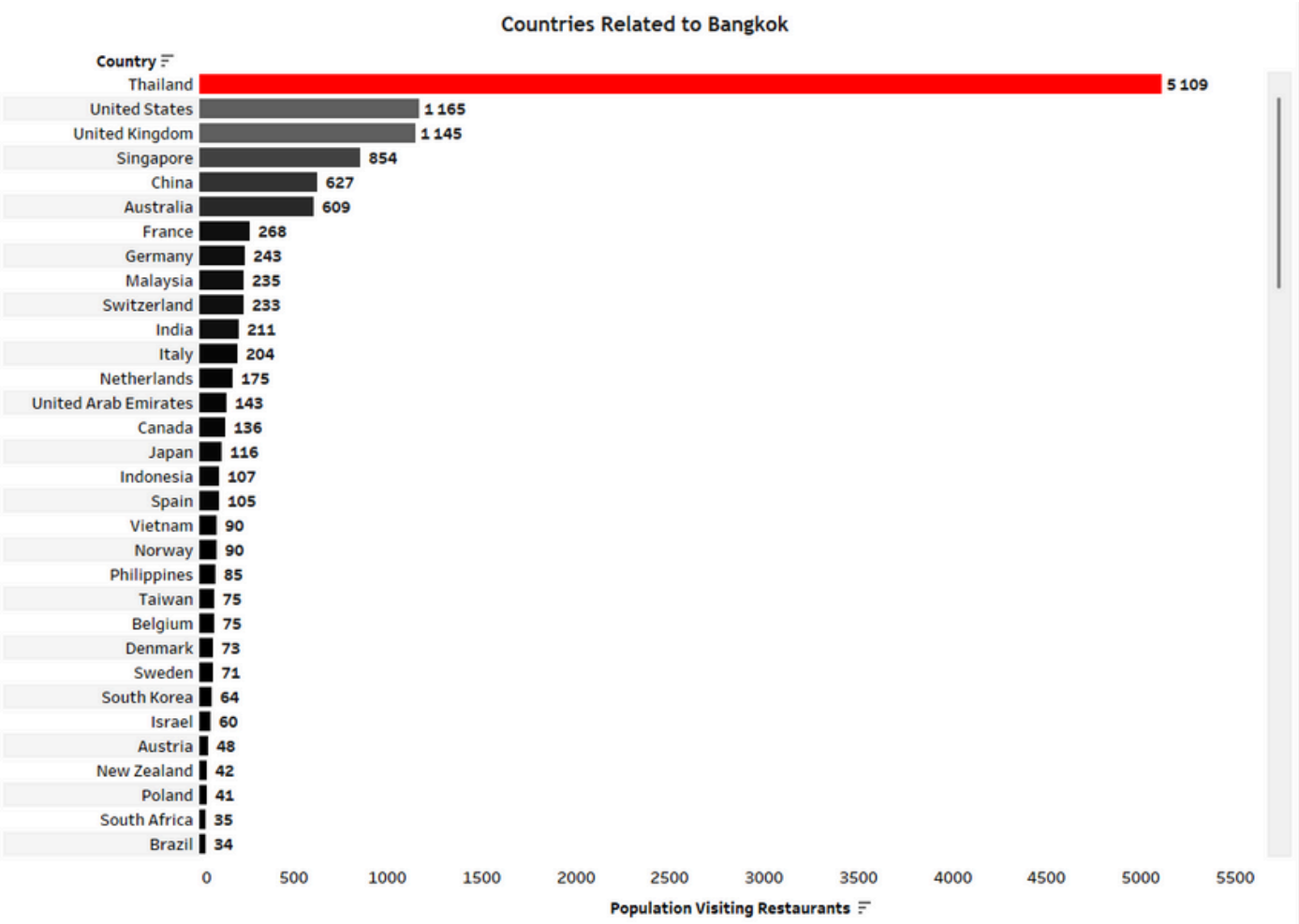


- **Thaïlande** était **le pays** ayant **le plus de visiteurs** dans les restaurants de **Bangkok.** Il détaille maintenant les **5 principales villes** de Thaïlande liées à Bangkok en termes de fréquentation des restaurants.

Analyse du graphique :

- **Bangkok** domine largement avec **4 284 visiteurs**, ce qui confirme son statut de centre gastronomique et touristique principal.
- Les autres villes, comme **Phuket (88), Ko Kret (69), Pattaya (60) et Chiang Mai (60)**, ont des chiffres bien plus faibles, montrant une concentration de la population visitant les restaurants à Bangkok.

DASH



- La **Thaïlande** domine largement avec **5 109 visiteurs**, surpassant tous les autres pays. Les **États-Unis (1 165)** et le **Royaume-Uni (1 145)** suivent, indiquant une forte présence de touristes anglophones. **Singapour, la Chine et l'Australie** se distinguent également, mais avec des chiffres bien inférieurs. **Les pays européens** comme la **France, l'Allemagne et la Suisse** affichent des valeurs modérées. Enfin, plusieurs pays comme **le Brésil, la Pologne et la Nouvelle-Zélande** comptent très peu de visiteurs, suggérant une influence touristique limitée.

DASH

Lien Dash : https://public.tableau.com/views/DASHBangkok/Dashboard1?:language=en-US&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link