M2 TAL

Projet Technique Web Individuel
Scraping MPA & SPA
2021

Chinatsu KUROIWA

Table des matières

Présentation

L'objectif de ce projet est de scrapper des informations depuis les sites web https://www.nh-hotels.fr/ et https://ntealan.net et d'afficher des résultats ou des analyses dans une application.

Afin de réaliser ce projet, les tâches de Scraping et de visualisation des données sont indépendantes. C'est-à-dire que on a d'abord extrait des données depuis les sites web et stocké ces données dans les fichiers « **output_natealan.csv** » et « **output_nhHotels.csv** » au niveau du répertoire data.

Ensuite, j'ai créé une application frontale en utilisant ces données.

Lancement

- SUR HEROKU

L'application est déployée sur Heroku à l'adresse suivante :

https://futur-plan.herokuapp.com/

Vous pouvez donc regarder notre présentation en appuyant cet URL directement.

- EN LOCAL

Si vous voulez lancer notre application en local depuis votre ordinateur au lieu de passer par l'URL mentionné ci-dessus, vous pouvez suivre les étapes suivantes :

1. Installez les dépendances : depuis le répertoire du projet, créez un environnement virtuel (avec pipenv https://pypi.org/project/pipenv/), connectez-vous, puis installez les dépendances depuis le fichier requirements.txt :
   $ pip3 install -r requirements.txt

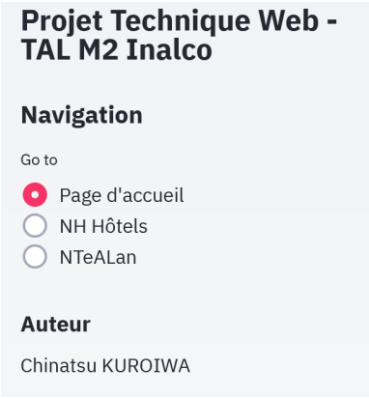
2. Ensuite lancez le fichier app.py qui est dans la racine du projet :
   $ **streamlit run app.py**
   Copiez et collez l'URL qui apparaît dans votre terminal dans le web browser que vous utilisez.

Visualisation

Après avoir lancé l'application, vous allez atterrir sur la page d'accueil comme suit :

Cette application est créée avec **Streamlit** qui est un cadre python pour la création d'applications frontales.

Vous pouvez naviguer entre les pages en appuyant les boutons dans la barre à gauche.



Scraping MAP & SPA

- ACCES AUX DONNEES

Avant faire le scraping de sites, il faut d'abord aller regarder les licences d'utilisation de ces plateformes pour respecter scrupuleusement toutes les règles de sécurités et confirmer ce que nous somme autorisés à faire.

Afin de vérifier cela, nous pouvons accéder au fichier robots.txt qui est placé à la racine de chaque site.

Image 1 : robots.txt du site de NH Hôtels

Image 2 : robots.txt du site de ntealan.net

- SITE NH HOTELS

L'objectif du travail avec la plateforme https://www.nh-hotels.fr est d'extraire les informations nécessaires afin de montrer au client la valeur de la mise en place d'un système de recherche des logements écologiques.

Afin de réaliser ce travail, on a fait d'abord extraire des informations sur ce site en utilisant **Beautiful Soup** et nous avons stocké ces données extraites dans un fichier CSV « output_hôtels.csv ». Enfin les analyses ont présenté avec des graphiques réalisés avec **Plotly Expresse** et présentés avec **Streamlit.**

Le scripte « **hotel_eco.py** » peut être trouvé dans le répertoire « scraping ». Si vous voulez lancer ce scripte, vous pouvez le faire avec la commande suivante :
$ python3 hotel_eco.py

Vous retrouvez le fichier de sortie « **output_nhHotels.csv** » dans le répertoire « data ».

- SITE NTEALAN

L'objectif du travail avec la plateforme https://ntealan.net est d'extraire quelques articles de ce site et de montrer au client la nécessité de soutenir les efforts réalisés par ce site en matière de documentation des langues peu-dotées.

Afin de réaliser ce travail, on a fait d'abord extraire des articles sur ce site en utilisant **Selenium** et nous stockons ces données extraites dans un fichier csv. Ces données sont présentées dans la même application web que celui du site Nh Hotels avec **Streamlit**.

Le scripte « **natealan.py** » peut se trouver dans le répertoire « scraping ». Si vous voulez lancer ce scripte vous avez besoin du browser Chrome et de télécharger le fichier « ChromeDriver » (https://chromedriver.chromium.org/) qui est adapté à votre version de Chrome. Ce fichier doit être mis dans la racine du répertoire.

Vous pouvez ensuite lancer : $ python3 natealan.py

Vous retrouvez le fichier de sortie « **output_natealan.csv** » dans le répertoire « data ».