M2 TAL

Projet Technique Web Individuelle

Chinatsu KUROIWA

Table des matières

[Présentation 2](#_Toc70338006)

[Lancement 2](#_Toc70338007)

[- en local 2](#_Toc70338008)

[Visualisation 2](#_Toc70338009)

[Scriping MAP & SPA 3](#_Toc70338010)

[- Accès aux données 3](#_Toc70338011)

[- Site NH Hotels 4](#_Toc70338012)

[- Site Ntealan 5](#_Toc70338013)

# Présentation

L’objectif de ce projet et de scrapper des informations nécessaires depuis des sites web (<https://www.nh-hotels.fr/> et [https://ntealan.net](https://ntealan.net/dictionaries/content/fr-af/yb_fr_3031)) et afficher des résultats ou des analyses dans un application.

Afin de réaliser ce projet, les tâche du scrapping et de la visualisation des données sont indépendants.

C’est-à-dire que j’ai d’abord extrait des données depuis les sites web et stockez des données dans les fichiers csv (**output\_natealan.csv** et **output\_nhHotels.csv** ) dans le répertoire data.

Ensuite, j’ai créé une application en utilisant ces données extrait.

# Lancement

## - en local

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

**$ pip3 install -r requirements.txt**

1. Ensuite lancez le fichier app.py qui est dans le répertoire « visualisation » :

$ **streamlit run app.py**

Copiez et collez l’URL qui apparaît dans votre terminal dans un browser que vous utilisez.

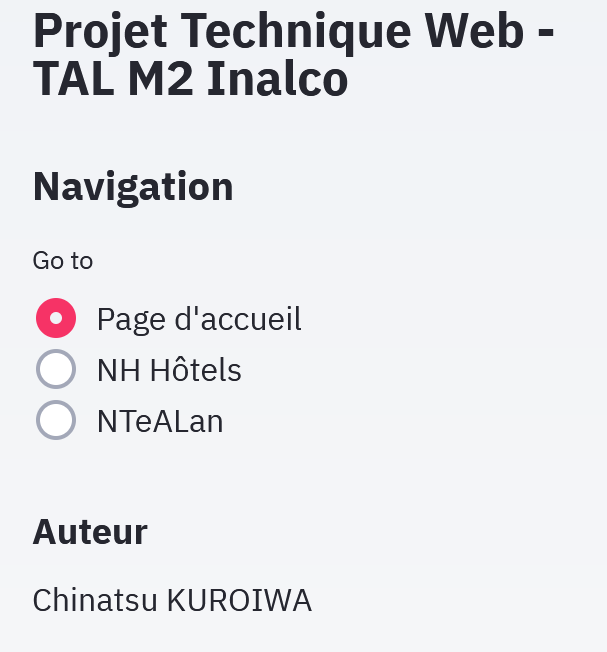
# Visualisation

Après avoir lancé l’application comme ci-dessous, vous avez trouver la page d’accueil comme suivante. :



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

Vous pouvez déplacer les pages en appuyant les boutons dans un bar à gauche.



# Scriping MAP & SPA

## - Accès aux données

Avant faire des scriping des sites, il faut d’abord aller regarder les licences d’utilisation de ces plateformes pour respecter scrupuleusement toutes les règles de sécurités pour regarder ceux 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 de travail avec la plateforme <https://www.nh-hotels.fr> est d’extraire des 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 un outil **Beautiful Soup** et stocker ces données extraites dans un fichier CSV (output\_hôtels.csv). Enfin les analyses ont présenté avec des graphiques traités avec un outil **Plotly Expresse** et les présenté avec un outil **Streamlit.**

Le scripte (**hotel\_eco.py**) peut se trouve dans un répertoire « scraping ». Vous pouvez lancer ce scripte comme suivante : $ python3 hotel\_eco.py

Vous retrouvez le fichier sorti (**output\_nhHotels.csv**) dans le répertoire « data ».

## - Site Ntealan

L’objectif de travail avec la plateforme  <https://ntealan.net> est d’extraire des quelques articles dans ce site et de montrer au client la nécessite de soutien le travail tel que ce site réalise.

Afin de réaliser ce travail, on a fait d’abord extrait des articles sur ce site en utilisant un outil **Selenium** et stocker ces données extraites dans un fichier csv.   
Ces données sont présentées dans un même page web que celui de site d’hôtel avec un outil **Streamlit**.

Le scripte (**natealan.py**) peut se trouve dans un répertoire « scraping ». Si vous voulez lancer ce scripte vous avez besoin de browser Chrome et télécharger le fichier driver qui est adapté à la version de votre chrome. Ce fichier driver doit être mettre dans la racine du répertoire.

Vous pouvez ensuite lancer : $ python3 natealan.py

Vous retrouvez le fichier sorti (**output\_natealan.csv**) dans le répertoire « data ».