4201C - DATA ENGINEERING USER GUIDE

Table des matières :

1. Installation des packages nécessaires

Tout d'abord, pour lancer l'installation, il faut s'assurer que plusieurs packages soient installés, notamment :

- Google Chrome, pour le scraping.

- Sélénium : Dans un terminal, taper la commande « pip install selenium ».

- WebDriver Manager : Dans un terminal, taper la commande « pip install web-driver manager ».

Nous avons utilisé une commande permettant d'auto-générer un chrome-driver afin d'exécuter notre scraping. Néanmoins, une autre manière de faire est de mettre dans le dossier « Data-Traitement » un ChromeDriver associé afin d'exécuter la récupération des données.

- Mongo, PyMongo : En local, suivez ce lien-ci permettant d'installer et lié python - Mongo et Pymongo :

https://docs.mongodb.com/manual/tutorial/install-mongodb-on-windows/

- Pandas : Dans un terminal, taper la commande « pip install pandas ».

- Numpy : Dans un terminal, taper la commande « pip install numpy ».

- Flask : Dans un terminal, taper la commande « pip install Flask ».

- Plotly : Dans un terminal, taper la commande « pip install plotly ». L'extention « plotly-orca » doit également être nécessairement installé.

Toutes ces installations sont effectuées en local, car nous avons décidé de travailler avec Sélénium. Néanmoins, si la première partie de l'exécution, c'est-à-dire le scraping, semble difficilement faisable avec Docker, les autres parties, c'est-à-dire la mise vers mongo et l'exécution de l'application pourront se faire, à condition d'avoir dans le même dossier les images associés de Docker, notamment Mongo et Plotly-Orca.

2. Récupérer les données d'internet

La première partie de l'exécution consiste à récupérer les données d'internet. Comme mentionné précédemment, il faudra être en local pour cette partie, sauf si vous avez trouvé un moyen de faire fonctionner le WebDriver Manager et sélénium sur Docker.
Pour effectuer cette étape, ouvrez le dossier « Data_Traitement ». Ce dossier présente quatre notebooks. Il faut d'abord ouvrir les deux se nommant « understat_scraping » et « wincomparator_scraping ». À l'intérieur de ceux-ci, se trouvent les commandes afin de récupérer les données, ainsi qu'un certain traitement et la mise en forme de celles-ci. Ainsi, effectuer un « run_all » afin d'exécuter les deux notebooks. À l'exécution de chacun, une page chrome a dû s'ouvrir puis se fermer. À la fin, de l'exécution de ces deux fichiers, deux dataframes sont apparus dans le même dossier, ils sont indispensables à obtenir pour la prochaine partie, il faudra également veiller à ne pas les déplacer.

3. Vers Mongo

La prochaine étape consiste à mettre nos données nettoyées dans une base de données Mongo. Pour cela, on reste dans le dossier « Data_traitement ». Il faudra ouvrir les deux autres fichier notebooks nommés « understat_vers_Mongo » et « wincomparator_vers_Mongo ». Il est important de noter, que ce soit pour la partie précédente ou celle-ci, l'ordre des deux fichiers générés n'a pas d'importance. Vous devez néanmoins par contre générer d'abord exécuter les deux fichiers liés au scrapping avant de générer les deux fichiers liés à la mise sous base de données sous Mongo.
Comme précédemment, l'exécution d'un « run_all » pour chacun des deux notebooks permet normalement la mise des dataframes crées sous une base de données Mongo.

4. L'exécution de l'application

Après les deux opérations précédentes, soit la récupération des données sur des sites internet et la mise en forme de ses données sous Mongo, nous pouvons exécuter l'application. Pour cela, on sort du dossier « Data_traitement » et on ouvre le fichier notebook « contenu_app ». Celui-ci contient le contenu de notre application en Python. L'exécution permet notamment de générer les quatre graphiques. Ceux-ci seront placés dans le dossier « template » et n'ont pas le besoin d'être déplacer ou modifier. Le fichier notebook créer, ou modifie le fichier « contenu_app.py ». On pourra le lancer notamment à l'aide de la dernière cellule (encore une fois, faites « run_all » et le fichier s'exécutera normalement.
À l'exécution, un lien du type `Running on http://127.0.0.1:5000/`
Il faudra copier-coller ce lien dans une barre de recherche internet afin de lancer le site.

5. Naviguer sur le site

Après avoir copié-collé le lien sur le site, et appuyer sur la touche « Entrée », vous apparaissez sur la page Home. Cette page vous explique en quelques lignes les objectifs et les informations présentes sur le site. En dessous, un menu d'accès aux autres pages vous permettra de naviguer sur ces pages en fonctions de ce que vous voulez pour prendre des décisions :

- Pour voir les prochaines confrontations et les meilleures cotes sur les matchs, cliquez sur le lien « Accéder au Calendrier et Meilleures cotes en cliquant ici ».

- Pour voir les données générales sur la saison des équipes, cliquez sur le lien « Accéder au Classement actuel en cliquant ici ».

- Pour voir les graphiques de comparaisons des équipes, permettant une représentation visuelle de la réussite et du niveau des équipes pour chaque match, cliquez sur le lien « Accéder au Graphique de comparaison en cliquant ici ».

Chacune de ces trois pages ont également un menu d'accès aux autres pages. De plus, les données et les observations présentes sur le site permettent simplement de prendre en compte des informations supplémentaires sur la réussite des équipes ou même leur performance globale pour les gens qui souhaitent avoir plus d'informations en main, notamment au moment de placer un pari sportif par exemple.

Exemple de menu de sélection de page :