PROJET SERVIER

Dans ce rapport je vous explique ce que j'ai fait sur un notebook. Cependant j'ai en réalité fait les traitement sur Airflow (vous trouverez mon code airflow dans le dossier dags du projet).

1/Dans un premier temps j'ai récupéré les fichier CSV sous le format Dataframe

```
import pandas as pd
import numpy as np

pd.set_option("max_colwidth", None)

df_drugs = pd.read_csv("drugs.csv", header=0, index_col=None, sep=',', encoding="utf-8")

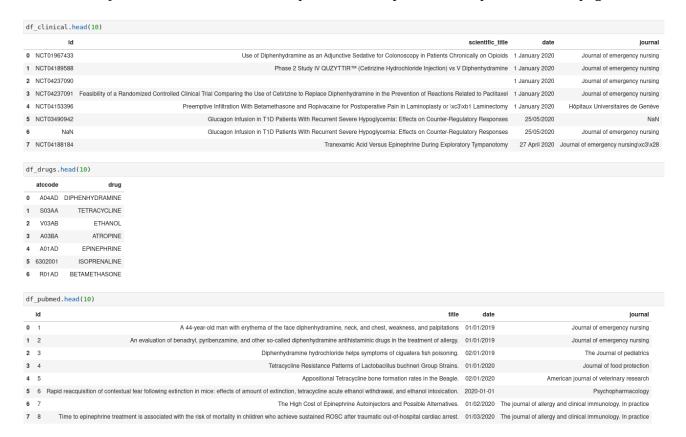
df_clinical = pd.read_csv("clinical_trials.csv", header=0, index_col=None, sep=',', encoding="utf-8")

df_pubmed = pd.read_csv("pubmed.csv", header=0, index_col=None, sep=',')
```

2/J'ai ensuite regardé le type des colonnes de chaque Dataframe

```
df_drugs.dtypes
atccode
           object
drug
           object
dtype: object
df_pubmed.dtypes
id
            int64
title
           object
date
           object
           object
journal
dtype: object
df clinical.dtypes
                     object
                     object
scientific_title
date
                     object
journal
                     object
dtype: object
```

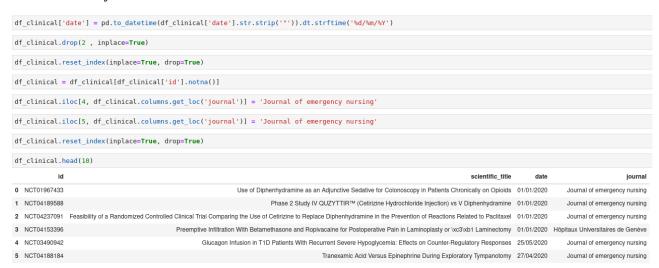
3/Par la suite j'ai affiché le contenu de chaque Dataframe pour voir s'il y avait du nettoyage à faire



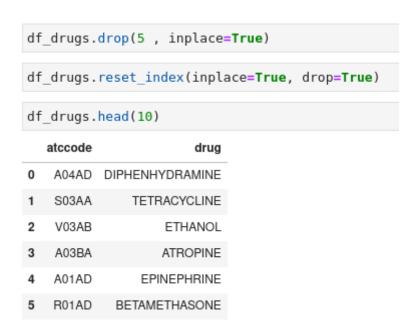
4/En ce qui concerne les données du clinical trials, j'ai constaté que :

- la colonne date était de type object. De plus les dates avaient des formats différents.
- la ligne 6 et 7 ont des valeurs NaN et semble être les mêmes lignes car elles ont le même scientific_title et j'ai supposé qu'un scientific_title ne peut appartenir qu'à un journal
- le journal de la dernière ligne contient des caractères en trop
- une ligne n'avait aucun scientific_title

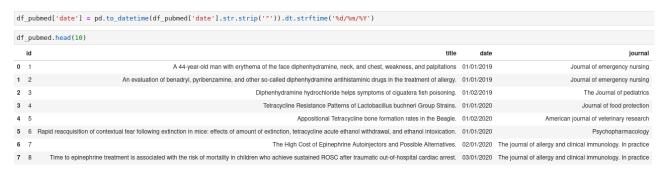
J'ai donc nettoyé ce Dataframe.



5/En ce qui concerne les drugs je me suis contenté de retirer la ligne 6 (indice 5) qui contenait un ID qui n'était pas un ATC code.



6/Ici j'ai juste modifié le format des dates pour qu'elles aient le même format dans chaque Dataframe



J'ai rencontré un problème au niveau de la création de JSON. En effet je ne savais pas comment réaliser le JSON en partant de 3 fichiers CSV.

Airflow

Sur Airflow j'ai réalisé les premières tâches de nettoyage, cependant j'ai rencontré quelques soucis. En effet je n'arrive pas à accéder à mes fichiers CSV déposés sur mon local. Je pense que c'est parce que j'ai lancé Airflow en passant par des containers Docker et, par conséquent, pour relier le container à mon local, il faut faire une manipulation.