Chapitre 4

Projet Python

Les données des projets sont dans un un zip accessible sur le lien suivant : https://1drv.ms/u/s!Am09h0q2OIX0a3BCjIUmAJbvvYE?e=1eyeyQ.

4.1 Projet 1 à 5 : Sample Insurance Portfolio, Real estate transactions, Sales transactions...

Dans le site internet https://support.spatialkey.com/spatialkey-sample-csv-data/ vous avez un ensemble de 5 jeux de données cvs.

- Sample insurance portfolio
- Realestate transactions
- Sales transactions
- Company Funding Records
- Crime Records

Chaque jeux de donnée correspond à un projet. Vous devez :

- choisir l'un des jeux de données
- décrire de manière sommaire de quoi parle le jeux de données choisi
- décrire les variables du jeux de données (leurs types et leurs sens)
- Créer 2 tableaux agrégés : faire un commentaire de ces 2 tableaux
- Créer 2 datavisualisations (diagramme en bâton ou bien histogramme, camembert, nuage de point) : faire un commentaire de ces 2 graphes

Dans le cadre des ces projets vous devez rendre un notebook avec les codes qui vous ont permis de faire les tableaux agrégés et les datavisualisations. Vous devez également faire une présentation power point ou bien latex de 11 slides maximums.

4.2 Projet 6 : Consumer Complaint Database

Les données de ce projet sont accessibles ici :

- https://catalog.data.gov/dataset/consumer-complaint-database
- ou bien sur kaggle https://www.kaggle.com/cfpb/us-consumer-finance-complaints

Vous avez une illustration de l'utilisation des donnée ici : https://towardsdatascience.com/multi-class-text-classification-with-lstm-1590bee1bd17. Bien sur, cette illustration dépasse le cadre de ce cours. En revanche vous pouvez en déduire le sens de certaine variable et comprendre l'utilisation de ces dernières dans le cadre du deep learning.

4.2.1 Travail à faire

Vous devez dans un premier temps :

— décrire de manière sommaire de quoi parle le jeux de données choisi

— décrire les variables du jeux de données (leurs types et leur sens)

Ensuite vous devez garder les lignes de cette base de donnée ayant la colonne consumer_complaint_narrative non vide. Ensuite vous devez :

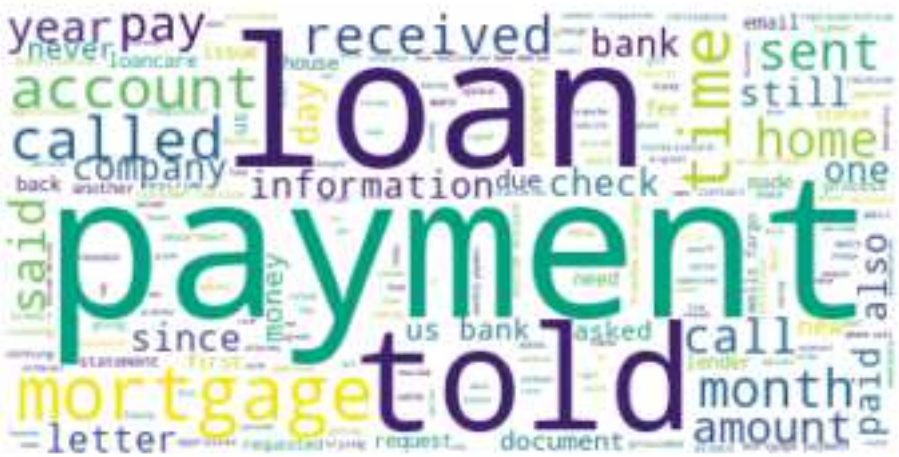
— Créer 2 tableaux agrégés : faire un commentaire de ces 2 tableaux

— Créer 2 datavisualisations (diagramme en bâton ou bien histogramme, camembert, nuage de point) : faire un commentaire de ces 2 graphes

Dans le cadre des ce projet vous devez rendre un notebook avec les codes qui vous ont permis de faire les tableaux agrégés et les datavisualisations. Vous devez également faire une présentation power point ou bien latex de 11 slides maximums.

4.3 Projet 6 bis : Nuage de mots sur les données Consumer Complaint Database

Le but de ce projet est de faire des nuages de mots sur les textes liés aux plaintes de clients d'assurance.



Ce projet est techniquement plus dure que les précédents. Les données de ce projet sont accessibles ici :

— https ://catalog.data.gov/dataset/consumer-complaint-database

— ou bien sur kaggle https ://www.kaggle.com/cfpb/us-consumer-finance-complaints

Vous avez une illustration de l'utilisation des données ici : https://towardsdatascience.com/multi-class-text-classification-with-lstm-1590bee1bd17. Bien sur, cette illustration dépasse le cadre de ce cours. En revanche vous pouvez en déduire le sens de certaine variable et comprendre l'utilisation de ces dernières dans le cadre du deep learning.

4.3.1 Travail à faire

Vous devez dans un premier temps :

— décrire de manière sommaire de quoi parle le jeux de données choisi

— décrire les variables du jeux de données (leurs types et leur sens)

Ensuite vous devez garder les lignes de cette base de donnée ayant la colonne consumer_complaint_narrative non vide. Vous devez transformer la variable Product en utilisant le code ci-dessous.

```
In [0]: df.loc[df['Product'] == 'Credit reporting', 'Product'] = 'Credit reporting,␣
 ↪credit repair services, or other personal consumer reports'
        df.loc[df['Product'] == 'Credit card', 'Product'] = 'Credit card or prepaid card'
        df.loc[df['Product'] == 'Payday loan', 'Product'] = 'Payday loan, title loan, or␣
 ↪personal loan'
        df.loc[df['Product'] == 'Virtual currency', 'Product'] = 'Money transfer,␣
 ↪virtual currency, or money service'
        df = df[df.Product != 'Other financial service']
```

Vous devez ensuite préparer les données relatif aux textes des plaintes des clients. ces textes sont dans la variable Consumer complaint narrative.

— On doit convertir en minuscule la colonne Consumer complaint narrative

— Enlever les chiffres de la colonne Consumer complaint narrative

— Enlever la ponctuation de la colonne Consumer complaint narrative

Vous devez faire des nuages de mots sur les données consumer_complaint_narrative. Il y a un nuage de mots par product. La construction d'un nuage de mot sous python est expliqué sur le lien suivant : https://www.datacamp.com/community/tutorials/wordcloud-python.