DATA Wrangling

WeRateDogs TWITTER

INTRODUCTION

Les données du monde réel sont rarement claires. L'ensemble de données utiliser dans le cadre de ce projet est l'archive de tweets de l'utilisateur de Twitter **@dog_rates**, également connu sous le nom de **WeRateDogs**. **WeRateDogs** est un compte Twitter qui évalue les chiens des gens avec un commentaire humoristique sur le chien.

À l'aide de Python et de ses bibliothèques, nous collecterons des données provenant de diverses sources et dans divers formats, nous en évaluerons leur qualité et leur ordre, nettoierons, puis créerons des analyses et visualisations.

Les objectives et motivation du projet sont :

- Préparation des données qui consiste à :
    1. Collecte des données
    2. Évaluation des données
    3. Nettoyage des données
- Stockage, analyse et visualisation des données
- Rapports sur les efforts fourni dans la préparation les données, les analyses et les visualisations des données.

1. Collecte des données

Ce projet utilisera des données provenant de trois sources différentes

Les archives Twitter de WeRateDogs : le fichier **twitter_archive_enhanced.csv** a été fourni par Udacity, disponible en téléchargement. Cette archive contient des données sur plus de 5000 tweets, qui ont été filtré avec des notes à 2356 tweets.

Les prédictions de l'image tweet : ce fichier **image_predictions.tsv** est présent dans chaque tweet selon un réseau neuronal. Il est hébergé sur les serveurs d'Udacity et doit être téléchargé par programme en utilisant la bibliothèque Requests et l'URL suivante: **https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv**

Twitter API : en utilisant les ID de tweet dans l'archive Twitter de WeRateDogs, interrogeons l'API Twitter pour obtenir les données JSON de chaque tweet à l'aide de la bibliothèque Tweepy de Python et stockez l'ensemble des données JSON de chaque tweet dans un fichier appelé **tweet_json.txt.** Les données JSON de chaque tweet doivent être écrites sur leur propre ligne. Ensuite, lisez ce fichier .txt ligne par ligne dans un tableau de données pandas avec l'ID du tweet, le nombre de retweets et le nombre de favoris.

2. Évaluation des données

Après avoir recueilli les trois éléments de données, évaluons de manière visuelle et par programmation pour les problèmes de qualité et de propreté. Voici le résumé de toutes les évaluations effectuées.

Quality issues

df_archive_data

1. On a beaucoup des valeurs manquantes dans les colonnes : **in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp**
2. Les colonnes **name, doggo, floofer, pupper et puppo** ont des valeurs manquantes dénotées **None** au lieu de **NaN**
3. Supprimer +0000 de la colonne **timestamp**
4. Le type de données des colonnes **timestamp et retweeted_status_timestamp** doivent être de type **datetime** au lieu du type **object**

5. La colonne name à des noms de chien erronées comme : **None, a etc**
6. Certaines valeurs de numerateur ont des valeurs incorrectes (**9,75/10 au lieu de 75/10 ;11,26/10 au lieu de 26/10**)

df_images

7. Les prédictions **p1, p2, p3** des chiens ont **la première lettre** parfois en **majuscule** ou en **minuscule**. On doit les uniformiser enfin d'avoir une cohérence dans le format des prédictions des noms
8. Les valeurs manquantes (2075 entries au lieu de 2356 entries)

df_api_tweets

9. Les valeurs manquantes (2327 entries au lieu de 2356 entries)

Tidiness issues

1. Les colonnes doggo, floofer, pupper et puppo doivent être remplacer par **un seul variable dog_stades**
2. df_api_tweets doit être une partie de df_archive_data

3. Nettoyage des données

Le nettoyage des données est la troisième étape du processus de préparation des données. Nous corrigerons les problèmes de qualité et de mise en ordre que nous avons identifiés lors de l'étape d'évaluation.

Cela nécessite trois étapes :

- **Définir** : définissez comment nous allons nettoyer la question en mots ;
- **Coder** : convertissez nos définitions en code exécutable ;
- **Tester** : testez nos données pour vérifier que notre code a été correctement implémenté.

En utilisant les étapes ci-dessus, nous avons nettoyé les données de la manière suivante :

- Supprimer les colonnes **in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp** qui ont beaucoup des valeurs manquantes ;
- Fusionner le **df_archive_data** avec **df_api_tweets** en tenant compte des entites manquantes dans le dataframe **df_api_tweets ;**

- Concaténer les colonnes **doggo, floofer, pupper et puppo** en une seule colonne **dog_stades**. Ensuite extraire **doggo, floofer, pupper et puppo** dans la colonne **dog_stades** en remplaçant **None** par le **NaN**. Enfin supprimer les colonnes **doggo, floofer, pupper et puppo ;**
- Supprimer +0000 de la colonne **timestamp ;**
- Convertir la colonne **timestamp** en **datetime ;**
- Remplacer les noms de chien erronés avec **np.nan ;**
- Remplacer les 3 valeurs de numérateur incorrectes par les valeurs arrondies des vraies valeurs (**9,75/10 par 10/10 et 11,26/10 par 11/10**) ;
- Convertir les prédictions **p1, p2, p3** des chiens en utilisant la méthode **str.capitalize()** enfin d'uniformiser le format des prédictions ;
- Supprimer les lignes de **df_archive_data** qui n'existent pas dans **df_images** , en se basant sur **tweet_id.**

4. Stockage, analyse et visualisation des données

Après avoir fini de nettoyer correctement les données, nous allons les stockées dans les fichiers csv nommé **twitter_archive_master.csv et image_predictions_clean.csv**

Enfin nous allons répondre aux quatre (4) questions suivantes pour analyses et visualisations les données :

- Quels sont les noms de chien les plus populaires ?
- Quelle est la corrélation entre le nombre de retweets et le nombre de favoris ?
- Quelle sont les stades de chien les plus courant ?
- Quels sont les stades de chien les plus Likes

5. Rapports

Nous allons rédiger des rapports sur les efforts fournis dans la préparation les données et sur l’analyse et la visualisation des données.

CONCLUSIONS

Ce projet a été location de mettre en pratique les étapes du processus de préparation des données qui sont : la collecte, l'évaluation et le nettoyage des données. Par ailleurs nous avons aussi exploré quelques questions enfin de faire l'analyse et la visualisation des données.

Comme les étapes du processus de préparation des données sont itératifs, il est fort possible qu'il existe encore beaucoup de problèmes dans les ensembles de données nettoyées et beaucoup d'autre questions à explore pour analyse et la visualisation cet ensemble de données.