RAPPORT DU PROJET

A) Introduction

Le jeu de données, que nous allons traiter et analyser, est l'archive de tweets connue sous le nom de WeRateDogs. Ce compte twitter évalue les chiens des gens avec un commentaire humoristique sur le chien. L'objectif de ce projet consiste à traiter les données : rassembler les données à partir de diverses sources et dans divers formats, évaluer leur qualité et leur propreté, puis les nettoyer.

B) Traitement des données (collecte, évaluation, nettoyage)

1) La collecte

La collecte est la première étape du processus d'extraction des données. Dans cette partie, nous téléchargeons les fichiers .csv, .tsv et json de la page web d'Udacity. Nous avons au total trois sources :

1.1) Archive Twitter améliorée

Les archives Twitter de WeRateDogs ont été téléchargées manuellement depuis la page web d'Udacity : twitter_archive_enhanced.csv. Cette archive Twitter contient les données de base des plus de 5000 tweets

1.2) Fichier de prédiction d'image

Nous téléchargeons le fichier à partir de l'URL : https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv , en utilisant la bibliothèque Requests.Le fichier de prédiction d'image (image_predictions.tsv) nous donne des informations sur le brend de chien présent dans chaque tweet. Il contient en particulier les trois meilleures prédictions de brend pour chaque ID de tweet, l'URL de l'image et le numéro de l'image correspondant à la prédiction la plus sûre.

1.3) Données JSON du Tweet

Nous téléchargeons egalement sur le site de Udacity. Les données JSON de chaque tweet sont écrites sur leur propre ligne dans un fichier appelé tweet_json.txt. Le fichier contient des informations supplémentaires : le nombre de retweets et le nombre de favoris.

2) Evaluation

Dans cette section, j'évalue les données recueillies précédemment. L'évaluation des données est la deuxième étape du traitement des données. Le but de ce processus est d'identifier les problèmes de qualité des données (problèmes de contenu) et le manque de rangement (problèmes structurels). Ce projet demande que seuls les tweets avec des évaluations originales avec des images, sans retweets, soient considérés. Ces deux informations se trouvent dans le fichier d'archive des tweets. Ensuite, les tweets originaux seront fusionnés avec la prédiction d'image et les données Json.

3) Problèmes de qualité :

Les colonnes contenant des informations sur les retweets - retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp contiennent au total 181 valeurs non nulles.

- Les colonnes contenant des informations sur les réponses - reply_to_status_id , in_reply_to_user_id contiennent au total 78 valeurs non nulles.

- La colonne avec les informations sur les urls contient 2297 expanded_urls non nulles, et 59 tweets sans urls (1 est un retweet, et 55 sont des réponses) ; et le s trois autres sont présents dans les données JSON mais pas dans le fichier de prédiction Imagine

- Nom de colonne : les chiens sans noms sont 'None' au lieu de NaN (ils sont 745) et certains noms (109 au total) ne sont pas corrects.

- Le format de la colonne timestamp n'est pas correct (data, time, +000) et il de vrait être de type datetime, et non de type string.

- les colonnes 'dog stages' - doggo, floofer, pupper, et puppo ont des valeurs q ui sont la chaîne "None" au lieu de NaN

- Colonne 'rating_denominator' : il y a 17 tweets originaux avec rating_denomi nator différent de 10 (13 d'entre eux concernent plusieurs chiens/chiens, et p euvent être supprimés).

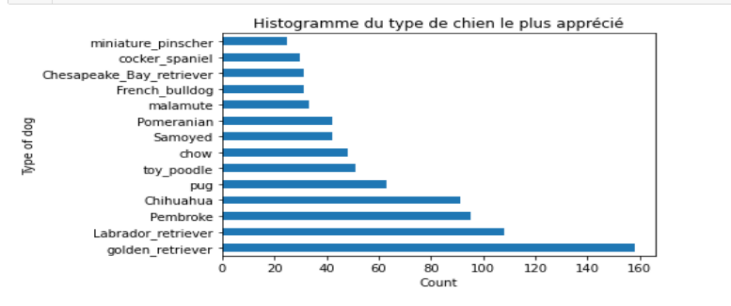
- Colonne 'rating_numerator' : elle contient 28 tweets avec rating_numerator su périeur ou égal à 15. La valeur maximale est de 1776, ce qui n'a pas de sens. Pour les tweets avec rating_denominator égal à 10, il y a 12 tweets avec ratin g_numerator >= 15 (7 d'entre eux sont des retweets et des réponses, et 5 des tweets originaux

- Colonne tweet_id : c'est un type numérique (int64). Il devrait être de type cha îne de caractères

4) Visualisations

1) Quelles sont les races les plus populaires ?

La race de chien la plus courante prédite est le Golden Retriever avec 154 tweets.



2) Quel est le stade le plus courant ?

La plupart des chiens sont classés dans le stade "Pupper".