Réalisé par : Essagar Ismail

Traitement de données

Etape 1 : Collection des données

Premièrement j'ai téléchargé le fichier twitter_archive_enhanced.csv à partir Udacity et je l'ai lu par **pd.read_csv** après j'ai obtenu le contenu fichier image_predictions.tsv en utilisant **request** et le lien https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv , j'ai sauvegardé ce contenu dans un fichier que j'ai appelé aussi image_predictions.tsv avec la fonction write et que j'ai lu par **pd.read_csv** et le séparateur était \t . Pour posséder la troisième dataset j'ai ouvert un compte développeur tweeter, j'ai suivi les instructions mais malheureusement je n'ai rien reçu de tweeter alors j'étais obligé de télécharger le fichier **tweet_json.txt** d'Udacity, j'ai mis le contenu de ce fichier dans un dataframe (pd.dataFrame) pour l'exploiter et savoir travailler avec.

Maintenant que j'ai les trois dataset, l'évaluation des données commença.

Etape 2 : Evaluation des données

Cette étape mène à bien connaître nos datasets. D'abord j'ai vu des échantillons de chaque dataset (**Évaluation visuelle)**, j'ai vu les colonnes, les types, le nombre de valeurs non nulles de chacune, j'ai compté le nombre de valeurs de quelques colonnes (**Évaluation programmatique**) afin de distinguer des problèmes de qualités et classer d'autre de rangement.

Je me suis sorti avec 10 problèmes de qualité et 3 problèmes de rangement.

Etape 3 : données de nettoyage

Cette étape consiste à résoudre les problèmes de qualités et de rangement que j'ai documenté lors de l'évaluation des données.

Pour cette étape j'ai essayé de résoudre les problèmes qui se ressemblent d'un seul coup comme le changement des types qui était simple grâce à la bibliothèque Pandas.

J'ai remplacé les noms faux des chiens que j'ai collecté sachant que tous commencent par des minuscules par np.nan ainsi que None qui n'est pas un nom valide par un chien.

J’ai retiré les retweets et les réponses des tweets qui ne nous intéressent pas, c’est pour ça j’ai supprimer de mêmes les colonnes dont on a pas besoin.

J’ai rendu le contenu de la colonne texte visible en faisant agrandir la colonne.

J’ai rassemblé les 4 colonnes des stades des chiens par une seule. Ce problème m’a pris beaucoup de temps et d’effort pour le résoudre afin de ne pas rater aucune information et garder l’exactitude des valeurs.

J’ai réduit les prédictions des chiens à une seule en gardant la première vraie prédiction.

J’ai fusionné les trois datasets dans une seul avec pd.merge.

Etape 4 : stockage des données

J’ai sauvegardé le résultat du nettoyage des données dans un fichier que j’ai nommé twitter_archive_master.csv grâce à to_csv .