Rapport sur la réparation des données (we rate dogs)

Cette préparation s'est déroulée suivant plusieurs phases, que j'expose de manière suivante :

Phase 1 : Collection des données

Dans cette phase je procède d'abord au téléchargement, par la méthode requests, du fichier 'image-predictions.tsv'

Ensuite j'utilise pandas pour effectuer le chargement des trois datasets 'twitter-archive-enhanced.csv' dans le dataframe tweets, 'image-predictions.tsv' dans le dataframe predict_images et 'tweet-json.txt' retweet_and_favorite.

Phase 2 : Evaluation des données

Dans cette partie j'ai effectué une lecture visuelle et programmation pour comprendre mes données et détecter des potentielles insuffisant ou défauts de qualité au de rangement (ordre) dans les jeux des données. Dans cette procedure nous avons énumérer les problème de qualité et d'ordre suivants (voir tableau 1 les défauts de qualité et tableau 2 pour les défauts de rangement):

Tableau 1: Défauts de qualité

N°	Remarque
tweets	
1	Eliminer les retweets
2	Eliminer les replies
3	La valeur 'a' dans la colonne name n'est présice
4	La colonne timestamp est une est string au lieu de datatime
5	Il existe des colonnes non necessaire pour une analyse
6	None pour definir pour designer NAN dans les colonnes doggo, floofer, pupper, puppo et name.
La table predict_images	
7	Des doublons sur la colonne jpg_url
8	Les colonne p1,p2,p3 ne sont intuitevement parlantes
Table retweet_and_favorite	
9	L'identifiant (id) du tweet diffère des deux autres tables (tweet_id)

Tableau 2: Défauts de rangement

N°	Remarques
1	La table **retweet_and_favorite** doit faire partie la table **tweets**
2	la table **predict_images** doit faire partie de la table **twwets**
3	Quatre valeurs en une colonne (doggo floofer pupper puppo) sur la table tweets.

Phase 3 : Nettoyage des données

Dans cette phase j'ai essayé de résoudre les différents problèmes des données remarqués plus haut. Mes propositions peuvent être comme suit :

Tableau 3: Résolution des problèmes de qualité

N°	Remarque
tweets	
1	Récupérer les retweets et les supprimer
2	Récupérer les tweet réponses et les supprimer
3	Supprimer les enregistrements avec nom='a'
4	Convertir la colonne timestamp au type datatime
5	Eliminer les colonnes qui n'ajoute pas de la valeur à l'analyse
6	Remplacer les colonnes contenant None par du vide (pas par NaN pour faciliter la concaténation des colonnes doggo, floofer, pupper, puppo)
La table predict_images	
7	Supprimer les doublons sur la colonne jpg_url
8	Renommer les colonnes p1, p2, p3 en prediction (_1, _2, _3) pour plus de clarté.
Table retweet_and_favorite	
9	Renommer l'identifiant, préalablement nommer id, en tweet_id.

Tableau 4:Résolution des problèmes d'ordre

N°	Remarques
1	Faire un merge sur le tweet_id pour joindre la table **retweet_and_favorite** à la table **tweets**
2	Faire un merge sur le tweet_id pour joindre la table **predict_images** à la table **tweets**.
3	Regrouper les colonnes doggo floofer pupper puppo en une colonne etape_chien et supprimer les colonnes originales (doggo floofer pupper puppo). Remplacer les valeurs vides de la colonne etape_chien par des NaN.

Phase 4 : Sauvegarde

Cette phase consiste à enregistrer dans un fichier csv 'twitter_archive_master.csv' le datframe contenant les données nettoyées.