Rapport décrivant nos efforts de Traitement de données

Dans le cadre de ce projet, nous allons travailler sur les archives de Twitter de WeRateDogs contenant des données de base sur 5000 tweets, les données supplémentaires collectées à partir de l'API Twitter et les fichiers de prédictions d'images. Après rassemblage et évaluation (visuelle et programmatique), nous avions recensé les problèmes de structures (d'ordre) et de qualités suivants :

1. Problèmes de structures

- Les noms des espéces de chiens (doggo, floofer, pupper, puppo) sont séparés en quatre (04) colonnes différentes.
- Toutes ces bases de données sont liées mais séparées en trois (03) bases de données distinctes.

2. Problèmes de qualités

A. Base des Archives Améliorées de Twitter

- Il y a 181 retweets comme indique retweeted_status_id.
- Le type de tweet_id doit être une chaîne au lieu d'un entier.
- Timestamp devrait être sous format date (datetime).
- Certains identifiants des chiens ont des photos manquantes.
- Certaines valeurs (440) dans la colonne rating_denominator ne sont pas 10.
- Quelques valeurs dans la colonne rating_numerator sont inférieures à 10.
- Quelques valeurs dans la colonne rating_numerator sont égales à zéro
- retweeted_status_timestamp doit être supprimé car nous nous intéressons au tweet.
- Les Valeurs nulles sont représentées par (None) dans la colonne name.
- certaines des lignes expand_urls ont 2 URL et nous avons juste besoin d'un lien tweeter.
- 59 valeurs manquantes dans la colonne expanded_urls.

B. Base des Tweets de Prédictions d'Image

- Certaines lettres en p commencent par une lettre majuscule tandis que d'autres sont en minuscule.
- tweet_id doit être une chaine de caractère et non un entier.

C. Base des Tweets via l'API de Tweeter

- Le nom des identifiants des colonnes doit être tweet_id au lieu de id.
- tweet_id doit être une chaîne et non un entier.