wrangle_report

September 8, 2022

1 WRANGLING REPORT

1.1 INTRODUCTION

L'objectif de ce projet est de mettre en pratique ce que j'ai appris dans la section Data Wrangling du programme Udacity Data Analysis Nanodegree. Les données qui ont été traitées sont une archive de tweets de l'utilisateur Twitter @dog rate, également connu sous le nom de WeRateDogs. WeRateDogs est un compte Twitter qui évalue les chiens des gens en laissant un commentaire humoristique sur le chien. Ces pièces ont presque toujours un dénominateur de dix. Ce projet m'a permis de developper mes competences en colectes de donnes. Le projet est sudivise en 3 parties: Gathering, Assess et le cleaning.

1.2 GATHERING DATA

Dans ce projet il fallait collecter les donnees de 3 sources diferentes et les stocker dans 3 dataset
-Fichier d'archive Twitter : le fichier twitter_archive_enhanced.csv a été fourni par Udacity et téléchargé manuellement.
-Les prédictions d'image de tweet, c'est-à-dire quelle race est présente dans chaque tweet selon un réseau de neurones. Ce fichier (image_predictions.tsv) est hébergé sur les serveurs d'Udacity et a été téléchargé par programme en utilisant le Demande des informations sur la bibliothèque et l'URL
-API Twitter et JSON : J'ai lu ce fichier tweetjson.txt ligne par ligne dans un cadre de données pandas avec identifiant de tweet, nombre de favoris, nombre de retweets, abonnés nombre, nombre d'amis, source, statut retweeté et URL.

1.3 ASSESSING DATA

Une fois les trois tableaux obtenus, j'ai évalué les données comme suit :
Visuellement, j'ai utilisé deux outils. L'une était en imprimant les trois entiers les dataframes se séparent dans Jupyter Notebook et deux en vérifiant le csv fichiers dans Excel.
Par programmation, en utilisant différentes méthodes (par exemple, info, value_counts, exemple, dupliqué, groupé, etc.).
Ensuite, j'ai séparé les problèmes rencontrés en problèmes de qualité et problèmes de propreté. Les points clés à garder à l'esprit pour ce processus étaient que les notes originales avec des images étaient recherchés

1.4 CLEANING DATA

Cette partie de l'analyse des données a été divisée en trois parties : définir, coder et tester le code. Ces trois étapes portaient sur chacune des questions décrites dans l'évaluation section.
La première et très utile étape consistait à créer une copie des trois dataframes d'origine. j'ai écris les codes pour manipuler les copies. S'il y avait une erreur, je pourrais créer un nouvelle copie de l'original. Il y avait quelques étapes de nettoyage qui étaient très difficiles. L'une d'elles était de faire fondre les étapes du chien dans une colonne à la place de quatre colonnes comme original présenté dans l'archive twitter.
Une autre étape de nettoyage très difficile a été lorsque j'ai dû corriger certains numérateurs qui étaient de véritables décimales. Cette question a été portée à mon attention.

1.5 CONCLUSION

La prepation des données est une compétence de base que quiconque manipule des données doit connaître. J'ai utilisé le langage de programmation Python et certains de ses packages. Il y a plusieurs avantages de cet outil (par rapport à par exemple Excel) qui est utilisé par de nombreux des scientifiques des données (y compris les gars de Facebook).

In [ ]:

In [ ]: