CSV File Information

CSV and excel files stored in the **data-for-reference folder** serve as both references and checkpoint for the purpose of validating if the exporting file will be in the correct format. It is mainly used for data cleaning and preparation process.

Files in Data Reference Folder (files are listed in order of production):

1) JokeText.csv: File by Aakaash Jois, he compiled the jokes into excel format. https://www.kaggle.com/aakaashjois/jester-collaborative-filtering-dataset

2) Cleaned_Jokes.csv: Using JokeText.csv, an additional column was created named "Cleaned Reviews". It contain the cleaned joke text whereby python and natural language processing toolkit (NLTK) was used to remove all the punctuation and stop words like he, she and etc.

3) JokeText_Category.xlsx: Using Cleaned_Jokes.csv another column name "Type_of_Jokes" was inserted whereby the team categorized the jokes into category namely, Ironic, Dark, Reference, Political Humor, Wordplay and Juvenile.

4) Word_Counter.csv: Generated using NLTK to tokenize the cleaned joke text from the above Clean_Joked.csv. It contain the data of how many time a particular word appear throughout all 100 jokes given.

5) Jester-data-1-with-header.xls: A modification of the original dataset found on NTULearn mini project link. Headers was manually written for all the column to make the file more readable in Jupyter Notebook

6) DataPreparation-1.csv: Using Jester-data-1-with-header.xls, this file reduce the number of user to 10000 and is extracted to validate the accuracy of the data cleaning process.

7) DataPreparation-2.csv: The final dataset which contain all the information needed into one single file. It contain joke text, type of joke, mean of the joke rating and the 10,000 user rating on 100 jokes with all the rating normalized.

CSV and excel files stored in the **data-for-analysis** is used mainly for analysis of the project.

1) Cleaned Dataset_Joker.csv: It is the same file as DataPreparation-2.csv which is named and stored differently so that the index is reset and keeping DataPreparation-2.csv as a backup copy.

2) Cluster_Ready.csv: It contain the user rating grouped into its cluster. This is generated after the clustering is done.

3) ForClustering Model.csv: This file can contain 1 million rows. It contain the rating of the user rating in row. It is used as a form of tagging whereby one column contain the joke id and the other contain the user id. The last row contain the rating which is why it is 1 million rows. This is used in the attempt to visualize the data. This file can be ignored as it is not used in the project.