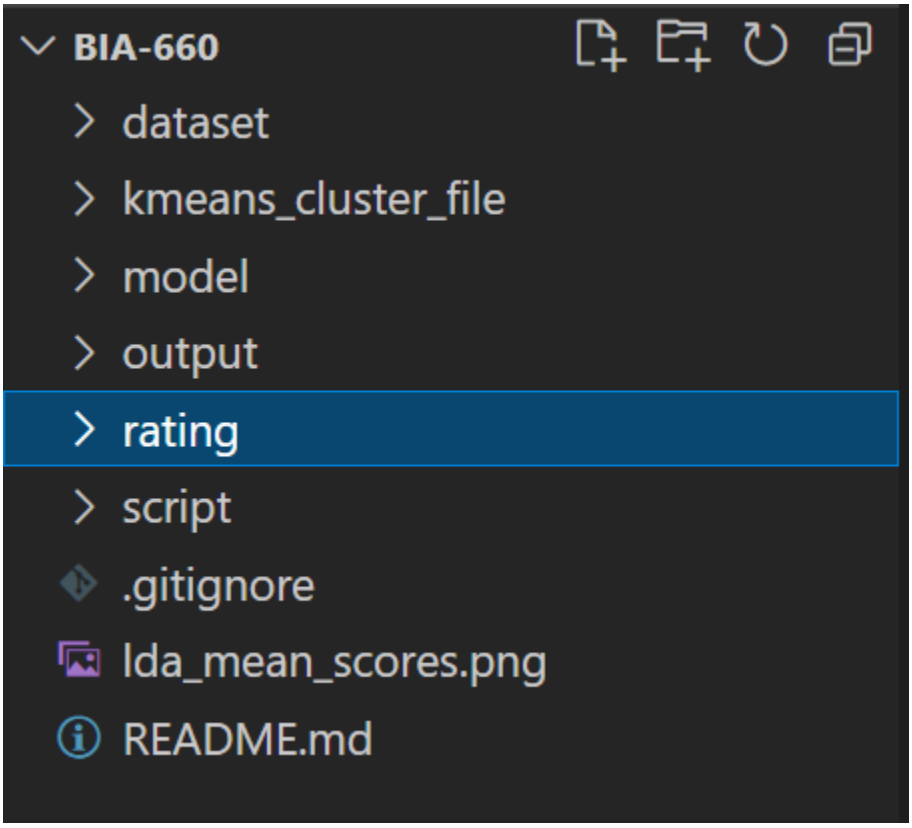
ReadMe

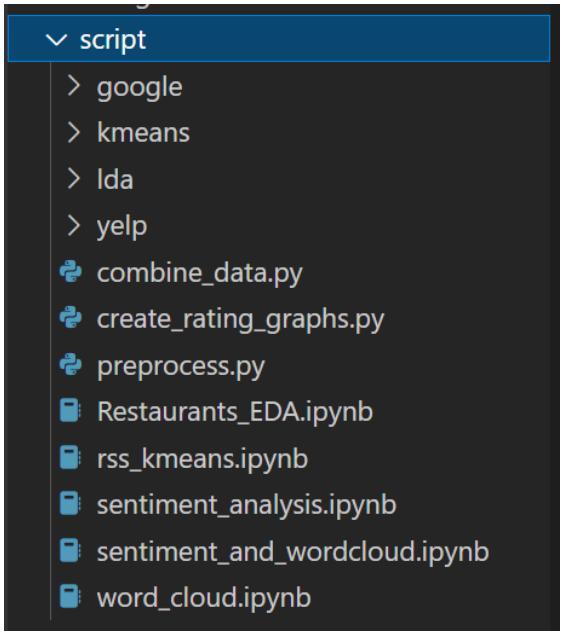
*PLEASE NOTE OUR EDA HAS BEEN REDONE
PLEASE CHECK THE FINAL REPORT FOR MORE INFO*



Folder structure of the zip file

The lda_mean_scores.png contains the graph generated by lda.ipynb

Scripts Folder



Scraping:
We've performed scraping of restaurant reviews from two sources, i.e. one from Google Reviews and another one from Yelp.

1. For Google Reviews, we have two scripts, one fetches the restaurants, other one fetches the reviews - stored inside the Google directory
    a. Run "restaurants.py" file
    b. Run "reviews.py" file
2. For Yelp reviews, we have a scrape_yelp file that does the same thing, - stored inside the Yelp directory
    a. To scrape the data from Yelp, run the "scrape_yelp.py" file

Preprocessing:
Run the preprocess.py file for doing the normalization, lemmatization, removal of stop words and removal of empty lines. Each restaurant file scraped from above will be created an individual file with _processed appended to filenames.

Combining all Processed Datasets:
Run the combine_data.py file for combining all the reviews together. The purpose of this script is to combine all processed files created from preprocessing. It will result in a file called all_reviews_processed.csv. It is located in the dataset folder.

Rating Graphs:
Run the create_graphs.py file for generating the count of ratings by restaurant. The graphs will be located inside the ratings folder.

Restaurant Sentiment Analysis KMeans
Sentiment analysis using compound scores of KMeans topic outputs. This is implemented in script: rss_kmeans.ipynb

sentiment_analysis.ipynb
Sentiment analysis using the VADER is present in this file. Run snippents in the "Script/sentiment_analysis.ipynb' file to get sentiment scores of the topics from LDA and Kmeans algorithm along with the word cloud for positive and neutral topics.

Sentiment and WordCloud
Exploratory data analysis and visualizations including the WordClouds and bar plots. This implementation is done in sentiment_and_wordcloud.ipynb

LDA Folder

LDA

- Train and find the best n_components
- Save LDA and CountVectorizer models
    - Located under models folder
- Script: lda.ipynb

LDA Get Topics

- Retrieve model and predict aspects of each restaurant using reviews
- Save topics to CSV files located under dataset
    - Restaurant_with_topics_lda.csv
- Script: lda_get_topics.ipynb

KMeans Folder

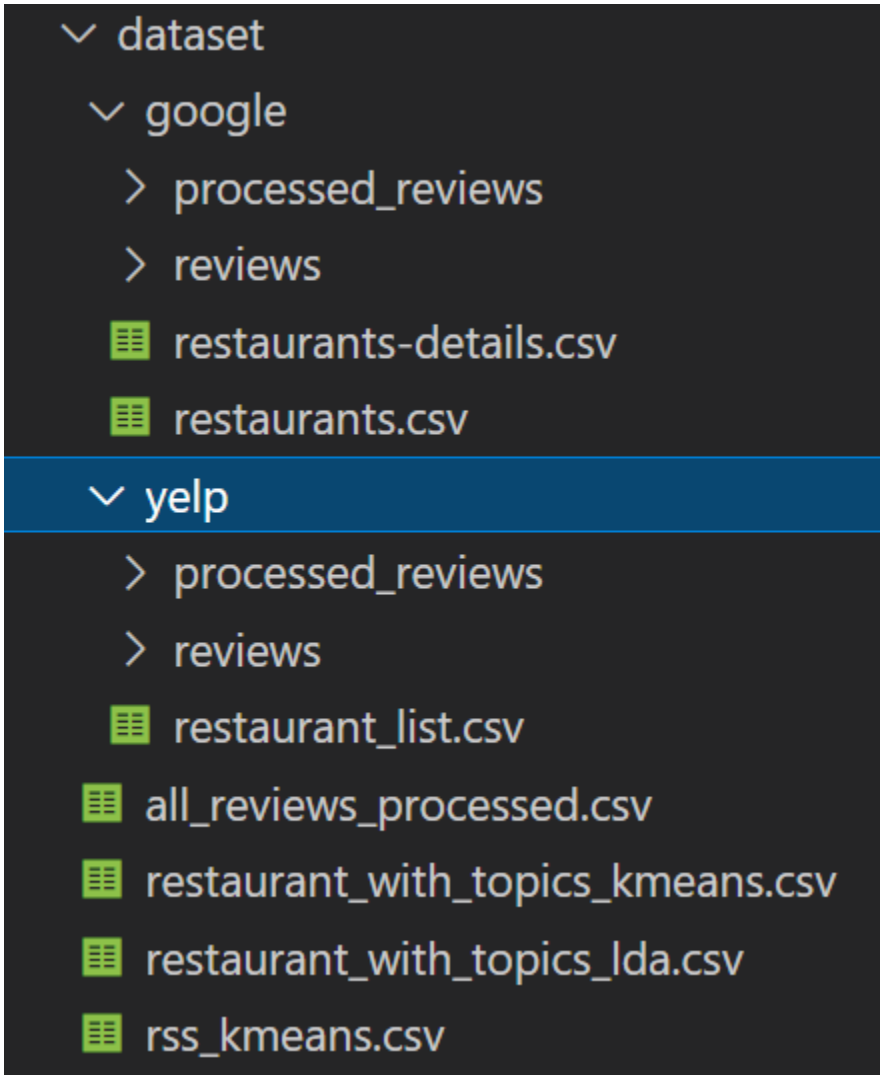
KMeans Get Topics

- Retrieve model and predict aspects of each restaurant using reviews
- Save topics to CSV files located under dataset
    - Restaurant_with_topics_kmeans.csv

Script :kmeans_get_topics.ipynb

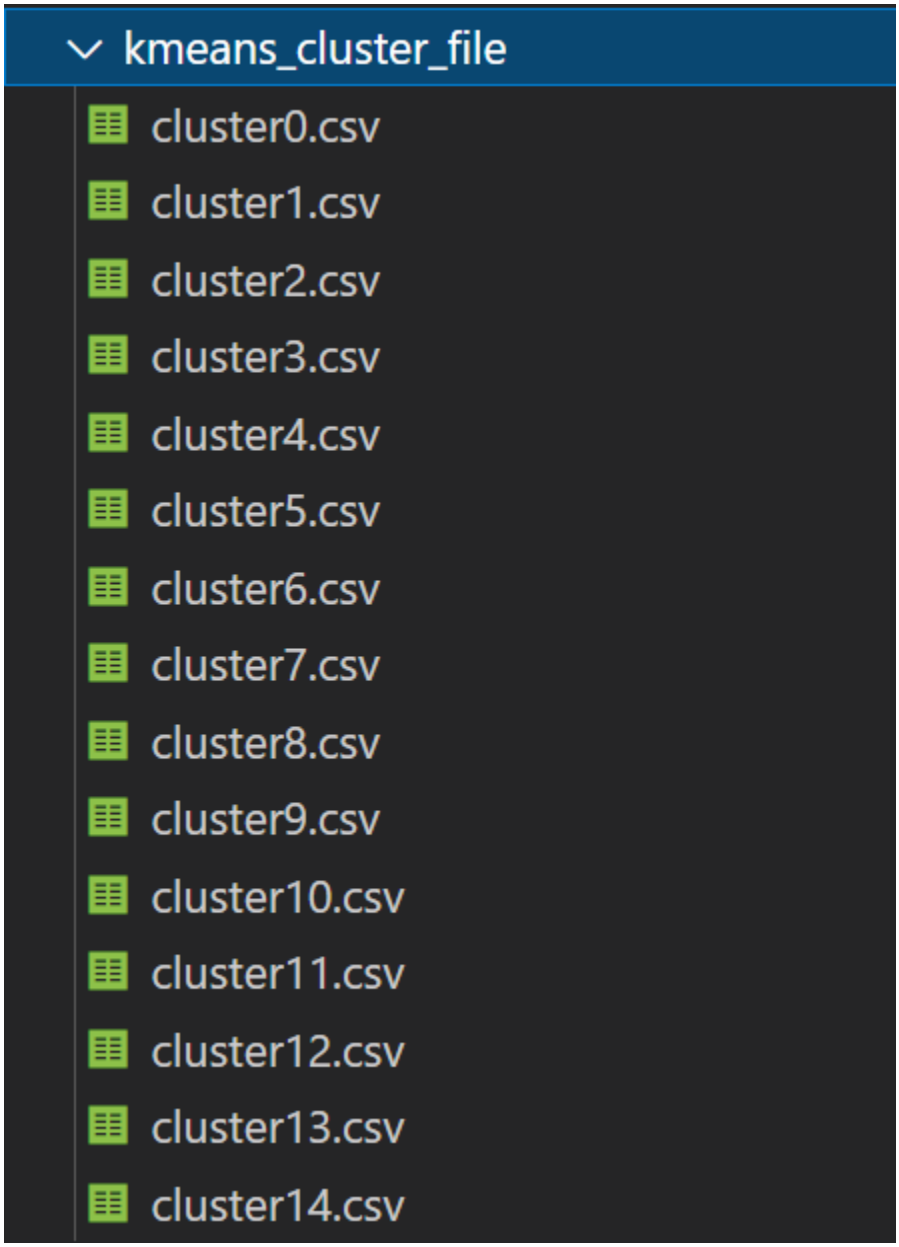
KMeans Report
External evaluation by using test set with manually labelled ground truth Labels to arrive at the classification report
Script: kmeans_report.ipynbsc

Dataset folder



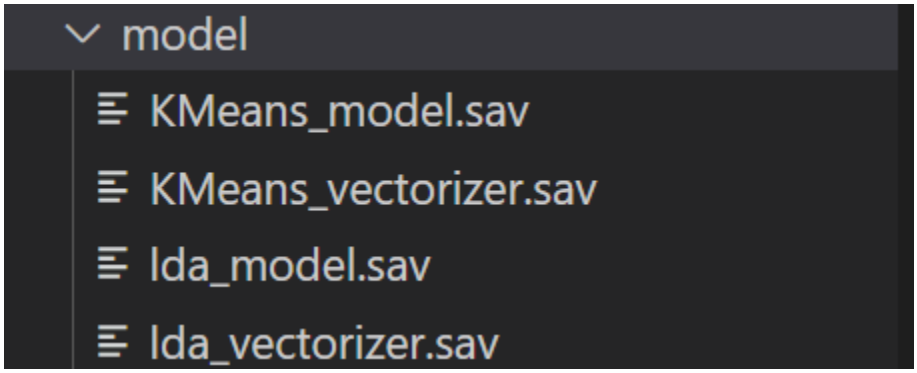
Contains all the dataset used for creating graphs, visuals and training

KMeans Cluster Files



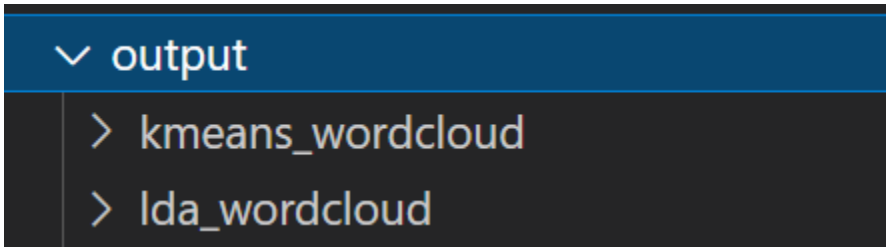
Contains the files of all the clusters generated by KMeans

Models



Generated by training of the LDA and KMeans algorithms. It is generated by lda.ipynb and kmeans_get_topics.ipynb.

Output



Contains the WordClouds for all the restaurants. It is generated by sentiment_analysis.ipynb

Rating

Contains all the Google and Yelp rating graphs. It is generated by create_rating_graphs.py