Capstone Project Submission

Team Member's Name, Email and Contribution:

Name: Sethupathy M
Email: sethuqr@gmail.com
Contribution:

- NULL Value Treatment
- Top 10 Cuisines Visualization
- Restaurants and Cost bar plot
- Text Pre-Processing
- Agglomerative Clustering
- DBSCAN Clustering
- Top terms in each cluster from DBSCAN Clustering
- Restaurants clustered together in each cluster from DBSCAN Clustering
- Word cloud for Positive and Negative reviews
- Bar plot for Top rated restaurants
- Vader Model for Sentiment Analysis
- Bar plot for most common words in positive and negative sentiments respectively
- Bar plot for Restaurants with more no of positive and negative reviews
- Data frame with Top 5 expensive and best rated restaurants
- Word cloud for most nominated words from 'Cuisine' feature
- TFIDF vectorizer
- Elbow Visualizer to find appropriate 'K' value
- KMeans Clustering and Model Validation
- Top terms in each cluster from KMeans Clustering
- Restaurants clustered together in each cluster from KMeans Clustering
- Hierarchical Clustering
- Scatter plot between Rating and Review length
- Bar plot for number of reviews by reviewers
- Bar plot between rating and positive, negative, compound polarity
- Distribution plot for polarity
- Data frame with Top 5 affordable and best rated restaurants
- TextBlob Model for Sentiment Analysis
- Model comparison

GitHub Link: - https://github.com/SethupathyM/Unsupervised_ML_Clustering_and_Sentiment_Analysis_Zomato_Restaurant

Unsupervised – ML – Zomato Restaurant Clustering and Sentiment Analysis

Summary

This project contains two datasets, one of which contains the information about Zomato restaurants names and Meta data and the other contains Zomato restaurants names, rating, reviews. There are a total of 6 variables describing 106 observations in ‘Zomato restaurants names and Meta data’ and total of 7 variables describing 10000 observations in ‘Zomato restaurants reviews’.

The purpose of this project is to try a Clustering algorithm on ‘Zomato restaurants names and Meta data’ to cluster the restaurants based on cuisines.
Then perform Sentiment analysis on dataset which is a merged version of above both.

We began by loading the ‘Zomato restaurants names and Meta data’ data set from the google drive into our colab notebook, which was in.csv format, and performing basic operations such as shape, describe and info, is null, and so on to gain a basic understanding of the dataset's contents. Null value treatment was done.

This dataset contains ‘Cost’ numerical feature, EDA was performed for better understanding the dataset. Then ‘Cuisine’ is supposed to feed into the Clustering algorithm but it contains non-ascii words, punctuations, uppercase letters, unlemmatized and unstemmed words so text pre-processing is done to eliminate unwanted characters.

KElbow visualizer was used to find appropriate K value. Then by KMeans Clustering restaurants are clustered on cuisines with the silhouette score of 0.195. Model validation is done with different k values. Hierarchical and Agglomerative Clustering is also applied for clustering the restaurants.

DBSCAN Clustering is also one of the finest clustering algorithms and it is robust to outliers. It labels the outlier observation as -1. This algorithm is applied for restaurant clustering. Then restaurants in each cluster are visualized.

We began by loading the ‘Zomato restaurants reviews’ data set from the google drive into our colab notebook, which was in.csv format, and performing basic operations such as shape, describe and info, is null, and so on to gain a basic understanding of the dataset's contents. And basic EDA is done on numeric feature for better understanding of the dataset.

Then two datasets are merged for Sentiment analysis. EDA is performed on merged dataset.

Vader model helps us to assign sentiment for given text. The output of the Vader model is a dictionary with keys ‘pos’, ’neg’, ’com’, ’neu’ and values for each key is between -1 to 1 for ‘com’ and 0 to 1 for others

Based on the compound score particular observation is assigned as ‘Positive’, ’Negative’,’ Neutral’ sentiment. Then sentiment analysis like most common words from each sentiment is done. Distribution plot for polarity and top best rated and affordable, expensive restaurants are analyzed.

TextBlob model gives output as float value between -1 to 1. Distribution plot for polarity and top best rated and affordable, expensive restaurants are analyzed from TextBlob model.

Then both Vader and TextBlob model is compared using scatter plot and Visualization.