Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

1) **Rahul Chauhan**
   E-mail: rahulchauhan161298@gmail.com

   - Data Cleaning: NaN/duplicates/outliers
   - Data Preparation
   - Data Analysis
   - Data Preprocessing
   - Data Manipulation
   - Data Visualization: Seaborn/Matplotlib
   - Line Plot, Bar plot, Histogram
   - Silhouetee Score Method
   - K means clustering: Elbow method
   - Dendrogram
   - Hierarchical clustering
   - Nltk
   - Pipeline setup: vectorizer/Tfidvectorizer/classifier/logistic regression

Please paste the GitHub Repo link.

Github Link:- https://github.com/Rahulchauhan1612/Zomato-Restaurant-Clustering-Sentiment-Analysis.git

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

Problem Statement:

The Project focuses on Customers and Company, you have to analyze the sentiments of the reviews given by the customer in the data and made some useful conclusion in the form of Visualizations. Also, cluster the zomato restaurants into different segments. The data is vizualized as it becomes easy to analyse data at instant. The Analysis also solve some of the business cases that can directly help the customers finding the Best restaurant in their locality and for the company to grow up and work on the fields they are currently lagging in.

Conclusion:

Top 5 cuisines:

North Indian, Chinese ii) North Indian iii) Continental iv) Ice Cream, Desserts v) Fast Food

Bottom 5 cuisines:

i) American, Fast Food, Salad, Burger ii) Continental, Italian, North Indian, Chinese iii) North Indian, Italian, Continental, Asian iv) Mexican, Italian, North Indian, Chinese, Salad v) Momos

Top 5 collections:

i) Unknown ii) Food Hygiene Rated Restaurants in Hyderabad iii) Great Buffets iv) Trending This Week v) Hyderabad's Hottest

Bottom 5 collections:

i)Sneak Peek Hyderabad ii) Best Milkshakes iii) Happy Hours, Top-Rated, Gold Curated iv) Best Bakeries
v) Great Breakfasts, Late Night Restaurants

In collection top 3 vocab used is week, visit and veggie.
In cuisines top 3 vocab used is wrap, Thai and sushi.

For n_clusters = 4, we get highest silhouette score is 0.4722959202437076
From elbow method we get 4 number of clusters is best among all.
Used dendrogram to find optimal number of clusters
Applied agglomerative hierarchical clustering from this we find 4 number of cluster good fit our model.
By applying different clustering algorithm to our dataset. we get the optimal number of clusters is equal to 4.
we have categorized rating in 3 types i.e., good, bad and average. 4500+ good, 1700+ bad and 900+ average ratings given by customer.
We have applied logistic regression on reviews dataset. getting 82% Accuracy on model.

References :

- GeekforGeeks
- Kaggle
- Analytics Vidya
- W3 school
- Sci-kit learn