Capstone Project Summary

Team Member's Name, Email and Contribution:

1. Lakshmi Keerthana

Email - keerthana826@gmail.com

- Data Cleaning
- Exploratory Data Analysis
- Data Preprocessing
- Textual Data Preprocessing
- K Means Clustering
- Agglomerative Clustering
- Topic Modeling

Please paste the GitHub Repo link.

GitHub Profile Link: - https://github.com/Keerthana826

GitHub Repository Link: - https://github.com/Keerthana826/Zomato-Data-Analysis

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

Zomato Restaurant Clustering and Sentiment Analysis Summary

This project focuses on Customers and Company, we have to analyze the sentiments of the reviews given by the customer in the data and make some useful conclusions in the form of visualizations. Also, cluster the zomato restaurants into different segments. The data is visualized as it becomes easy to analyze data at instant. The analysis also solves 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. This could help in clustering the restaurants into segments. The data has valuable information around cuisine and costing which can be used in cost vs. benefit analysis Data could be used for sentiment analysis. The metadata of reviewers can be used for identifying the critics in the industry.

In this unsupervised machine learning project, we looked into multiple unsupervised models and sentiment analysis. We started with loading the data, inspecting the data and Data Cleaning

Through Exploratory Data Analysis, we visualized the data through bar plot, pie chart, wordcloud and distplots.We have seen a correlation matrix is a table showing correlation coefficients between variables through heatmap.We have the capped outliers with the highest and lowest limit using IQR method.

During data pre-processing, we have transformed raw data into an understandable format for ML algorithms.For textual data pre-processing, NLP pipeline include steps such as sentence segmentation, word tokenization, lowercasing, stemming or lemmatization, stop word removal and spelling correction.

Principal Component Analysis was used to reduce the dimensionality of large datasets.

Models like K-means were used for grouping the data into clusters and visualizing them in clusters. Elbow method and Silhouette score was used to find the optimum number of clusters.Agglomerative Clustering is a hierarchical clustering algorithm that was used to cluster the data points. Dendrogram plot was used to find the optimal number of clusters.

For Sentiment Analysis, topic modeling was a technique to extract the hidden topics from large volumes of text. Latent Dirichlet Allocation(LDA) is the algorithm that was used for topic modeling and extracted good quality topics that are clear, segregated and meaningful.