Geospatial Analysis for Yelp User Behaviour

Exploring User Behavior, Business Clusters, and Sentiment Patterns

Team 010: Mengyuan Li, Siying Liu, Xianyi Nie, Hao Hu, Yaxin Xue, Xiaocheng Chen

Motivation

This project is motivated by the growing importance of online review platforms like Yelp, which influence customer decisions and impact businesses. We want to analyze Yelp data to understand customer behavior and business trends across four US cities, taking into account regional differences.

The goal is to help businesses tailor their strategies to meet customer needs in different locations and assist policymakers in promoting economic growth. Additionally, consumers will benefit from better understanding regional differences in user behavior, helping them make informed choices.

Data

In this project, we primarily use data downloaded from Yelp's official website. It contains nearly 7 million records of reviews, 150k businesses, and 11 metropolitan areas.

Experiments

Our testbed consists of Yelp datasets containing reviews and business information for four selected US cities: Carpinteria, Philadelphia, Santa Barbara, and Tucson. Our experiments revealed distinct sentiment patterns and correlations between user behavior and business characteristics in the four cities. We successfully identified clusters of similar businesses and user behaviors, highlighting areas with similar patterns of sentiment and user activity. Moreover, the geospatial analysis provided valuable insights into the geographic distribution of user behavior patterns.

Our Approach

1. Sentiment Analysis
We applied TextBlob, a Python library for natural language processing (NLP), to perform sentiment analysis. TextBlob calculates the sentiment polarity of a given text, which ranges from -1 (negative sentiment) to 1 (positive sentiment). We applied TextBlob to the review texts to obtain sentiment scores for each review.

2. Geospatial Analysis
To visualize the geospatial patterns in the sentiment of reviews and user behavior, we used folium, a Python library for creating interactive maps. We plotted the businesses' locations and assigned colors based on their sentiment scores. This helped us identify areas with high and low sentiment scores in the city of interest.

3. Clustering Analysis
We applied the K-Means clustering algorithm to group businesses based on their latitude, longitude, and sentiment scores. K-Means is an unsupervised learning algorithm that partitions the data into K clusters, where each observation belongs to the cluster with the nearest mean.

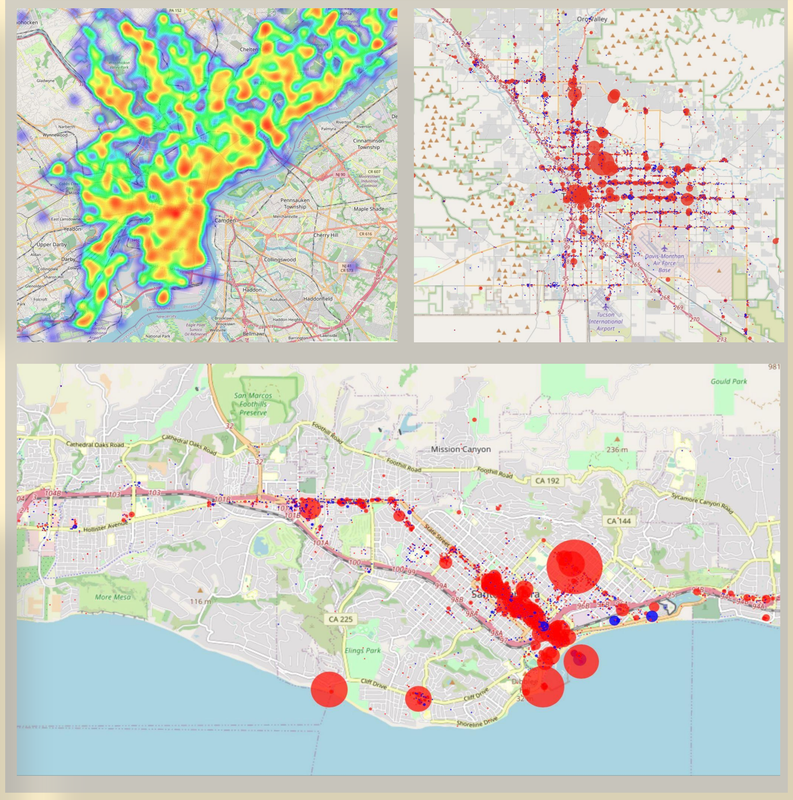
4. Keyword Extraction
To extract the top keywords from user reviews, we first preprocessed the text by tokenizing, removing stop words, and stemming. Then, we used the Term Frequency-Inverse Document Frequency (TF-IDF) approach to identify the most relevant words in the reviews. We obtained a list of top keywords that represent common themes and trends in user behavior.

Compared With Other Methods

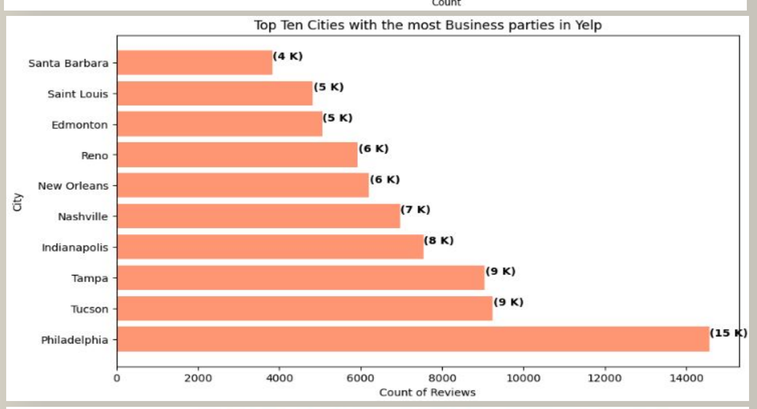
Our approach uniquely combines sentiment analysis with geospatial techniques to uncover sentiment patterns and their spatial distribution. This innovation enables stakeholders to effectively visualize regional differences and trends, leading to better-informed decisions.

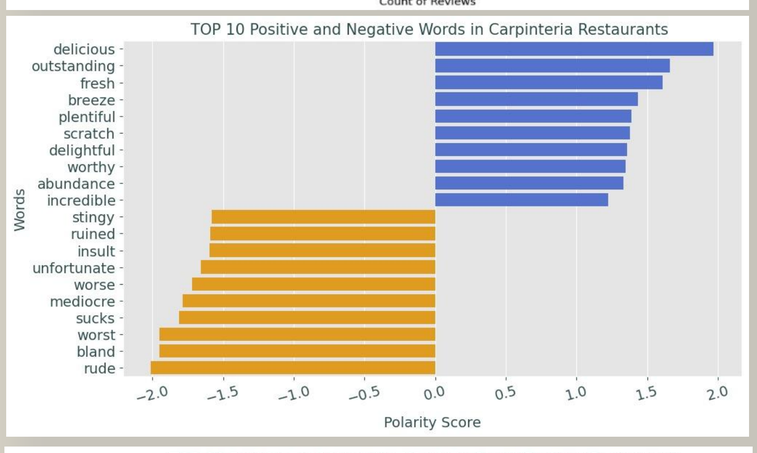
The novelty lies in the focus on location-based sentiment analysis and geospatial visualization, allowing stakeholders to gain insights into regional sentiment variations and preferences. Additionally, spatial clustering helps identify hotspots of similar sentiment and keyword usage, revealing distinct patterns across regions.

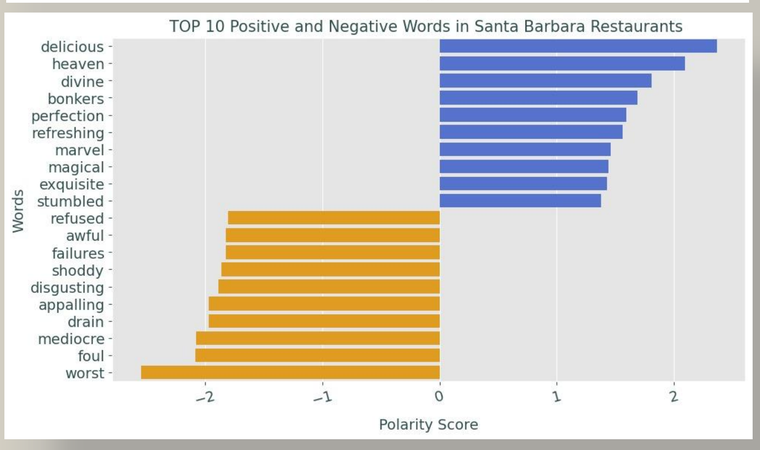
Results - Heatmap



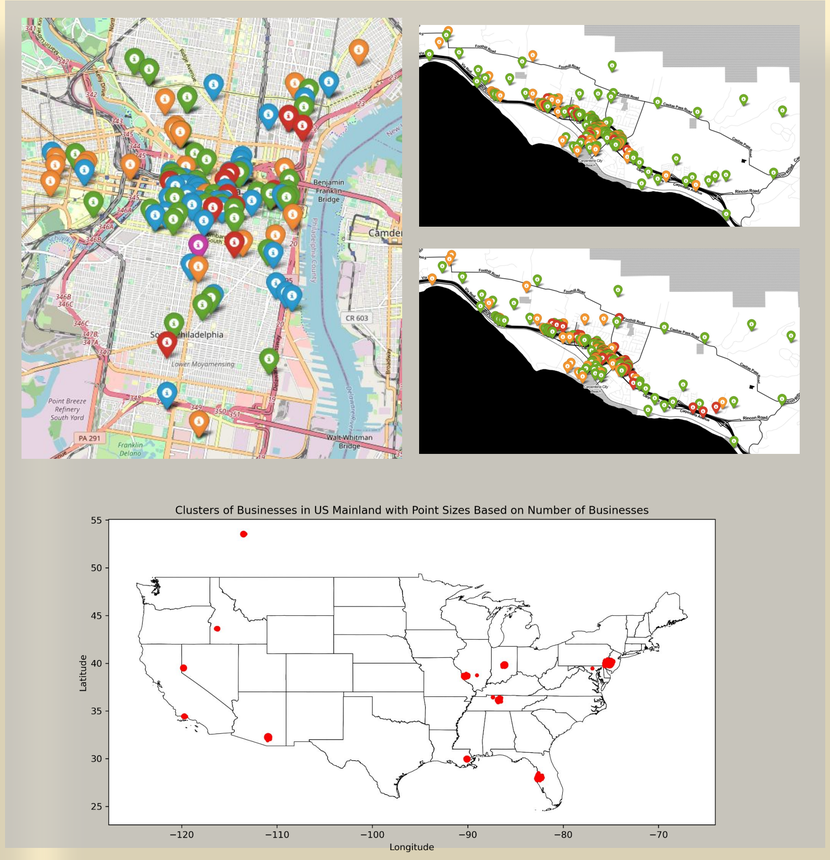
Results - Sentiment Distribution







Results - Clustering Analysis



Results - Word Cloud