Get Yelp

A Comprehensive Analysis of Review Trends

Michael Gigiolio, Dante Pasionek and Dominic Deckys

Our project intends to locate relationships between different factors in Yelp reviews. We seek to identify patterns in presumably unrelated information such as ratings and location or hours of business. Additionally, we are looking to find what factors are generally indicative of an unsuccessful or poorly rated business.

Prior Work

Performed by Yelp

- Photo Classification
- Natural Language Processing
- Sentiment Analysis

Data and Procedure

Data Set

https://www.kaggle.com/yelp-dataset/yelp-dataset/data

Data set provided by Yelp. Set contains 5,200,000 data points with 174,000 different businesses reviewed across 11 different metropolitan areas. Data set currently downloaded onto all group members’ computers.

Proposed Work

Cleaning and Preprocessing

- Removing quotation marks
- Standardizing times and postal codes
- Handling missing neighborhoods
- Removing unneeded attributes
- Managing foreign syntax

Integration and Analysis

- Visualizing the data
- Handling location information
- Trend analysis
- Typing analysis

Tools and Methods

Programming Tools

- Python 3
- Pandas
- SciPy
- NumPy

Mathematical Tools

- Linear Regression
- OLS Regression
- P-Value Calculations
- Confidence Intervals
- Hypothesis Testing

Evaluation

Desired Results

We intend to evaluate our data set for various implicit trends which are not immediately evident from the raw data itself. By ordering review scores, we will be able to determine which attributes are more strongly associated with higher or lower ratings. In addition, we can find and compare the average ratings of specific types of businesses to one another. By determining these factors, we will be able to provide concrete information by which establishments can improve their consumer appeal.