Yelp Dataset Analysis Using Hive

Mohsen Alam, Benjamin Cevallos, Oscar Flores, Randall Lunetto, Kotaro Yayoshi

Department of Information Systems, California State University

Los Angeles

e-mail : malam@calstatela.edu, bcevall@calstatela.edu, oflores4@calstatela.edu, rlunetto@gmail.com, kyayosh@calstatela.edu

**Abstract:** Yelp has served and will continue to serve as a data-driven application. Yelp has published a dataset (www.yelp.com/dataset/) containing business information, reviews, user information, and check-in information. This report will examine this dataset to provide descriptive analytics to understand business performance, geo-spatial distribution of businesses, reviewers' rating and other characteristics, and temporal distribution of check-ins in business premises.

**1. Introduction**

Yelp is a common database used by the public that allows the users to view specific data on business. Through the Yelp database you can pull statistical data that can serve us in the future data analysis. In this project we will utilize geo-spatial and temporal distributions to provide data analysis on businesses. This project provides information about tip sentiment analysis which will allow the users to see just how well a business is. Secondly, we have decided to analyze the count per year which gives us a clear understanding of what user, and how many counts of that type of user ratings occur over a span of years. Ratings category count also provides us with information about what the highest rated region is. Lastly, we can also utilize pulling specific data such as what places are the top-rated places for children. With the given different variations of these descriptive analysis, geo-spatial analysis, and temporal analysis, we can use data that is analyzed to determine how well a business is doing as well as anticipate how well it might do in the future.

# 2. Related Work

## 2.1 Sentiment Analysis

Sentiment analysis, also known as opinion mining, has beena popular research method in the natural languageprocessing field. The goal of sentiment analysis is to define automatic tools to be able to extract subjective information such as opinions and sentiments, from text in natural language. In order to generate data that can be utilized by decision support systems.

Sentiment analysis is often used by companies to quantify general social media opinion. One of the simplest and most common sentiment analysis methods is to classify words as “positive” or “negative”, then to average the values of each word to categorize the entire document.

Griffo (2016) has worked on Twitter sentiment analysis in Hive, and we based our work on his project off of a project a Software Engineer Umberto Griffo had done.

His method was to create a counter and name it as polarity then increment or decrement based on the number ofpositive or negative values there were.

When a ‘negative’ was triggered, it would decrement thepolarity counter by 1 while when there was a ‘positive’trigger it would increment the counter by 1.

After which, there would be a small algorithm named tweets\_sentiment where the sum of the polarity would becompared with a 0 value.

If the sum of the polarity was greater than zero, it would be labeled as positive. If the sum of the polarity was less than zero, it would be labeled as negative.

For the purposes of this assignment our analysis counts the number of time the words “positive”, “negative” or “neutral” is used. The main difference between our codes is that Umberto’s utilizes a sentiment dictionary and includes an extra algorithm to delineate between positive and negative words. Ours is a more direct and literal approach for the purposes of displaying the purpose of this type of analysis.

## 2.2 Geo-Temporal Analysis

Geo-Temporal analysis is a process that allows to characterize subsets of items in a geographic database, for example events, with respect to the similarity of their location and timestamp. It is information derived from an analysis of images and data associated with a particular location. It uses imagery to survey and assess human activity and physical geography anywhere on Earth.

Using Lab 4as a basis Woo (2020) we recreated the geo**-**temporal data presentation. The methods and data types were similar. As well as the utilization of Excel with the 3D maps feature to graphically represent the data.

For our presentations, we utilized the techniques presented in Lab 4 and consecutive labs to graphically display geo-temporal data sets. As can be seen with most of our outputs, we also utilized techniques from the Labs to graphically display our geo-temporal data. The only variations were the actual data used but the process was similar to that of the labs.