Analysis of Yelp dataset

Team "The Data Analytics Group"

Akshay Karki (avk1063@rit.edu)
Gaurav Gawade (gdg6776@rit.edu)
Shristika Yadav (sy2109@rit.edu)

ABSTRACT

Yelp ratings are often seen as a respectable metric for businesses. But, how can we better something which is already good? Wouldn't it be cool if we could predict the number of stars a business would get just by looking at the number of stars it gets on each review? Or we could predict the busiest hours a business should expect based on the frequency of people? These are some of the questions that we hope to tackle with this project.

1. INTRODUCTION

The dataset contains about 1.6M records. But given the time constraints associated with the project, we plan to reduce the total number of records to analyze without skewing the actual value of the dataset. The dataset is spread over multiple JSON files and the data in all the files is interlinked. The task involved in this project will be to parse the JSON files using Python and Pandas library. The parsed data will then be stored in a dictionary and then further used for visualizing trends and analyzing patterns.

There is a review file where multiple users have reviewed a business by giving it some ratings as well as comments. A business file consists of all the information regarding that business while a user file gives information about the reviewer. A check-in file tells which time the business is usually the busiest at.

With the given dataset, some of the important questions which we will be able to answer are, how well a business is doing, is the business good for kids, depending on user reviews what is the next business he will review or depending on the content of the review what can be the rating for that business if the rating is missing.

2. IMPLEMENTATION

The implementation part of the project will mainly consist of three stages such as dataset collection/assembly, clean data and analyze, and data mining and visualization. Our main focus will mainly be on cleaning the data as much as possible in order to have optimum analysis result.

2.1 Tools used

In this section, we will consolidate the data collected from various sources for Yelp [3]. Given the dataset and how it could be analyzed, we used Python to code. We also used Python libraries like SciPy, Matplotlib, and Pandas for data visualization. We used Weka for clustering purposes, but given the size of the dataset, it took a long time to load the dataset and perform clustering on the same. So, we switched over to Python and wrote the code to visualize our data. We hope to use Weka again to gain a different perspective on the same dataset or a section of the dataset. Tableau might be used for data visualization to further enhance the project.

2.2 Data Cleaning and Analysis

The data collected from [3] had to be cleaned a lot since it had missing data and outliers. We first had to convert the JSON file from the dataset into a CSV file, since we felt that we would be more comfortable working with a CSV file in Python. We wrote a Python code for the same and used the 'json'library to convert the JSON files into CSV files. The files generated were 'checkins'file, which had the count of checkins for each day and 'time'file, which had the count of checkins during the entire day. Each record in the generated files has a unique business ID.

Having domain knowledge regarding the dataset that we were dealing with, we could better handle the missing data present. We had lots of 0 values. But, that is not a missing value since a day can have 0 checkins for a given day. So, we decided to see if we had any records with 0 values over a period of five days. About 28K records out of the given 140K records had such values. So, we got rid of these records since we felt that those records could not sufficiently contribute to the overall analysis.

We plan on further exploring some of the techniques given in [2] and [1]. We did not face the problem of incorrect column names or useless attributes since the CSV file was generated by us and we made sure to keep these points in mind. We do not remove duplicate records if the business ID is different but all the days have exactly same checkins. This is because it is probable for two different business IDs to have the same order of checkins.

.

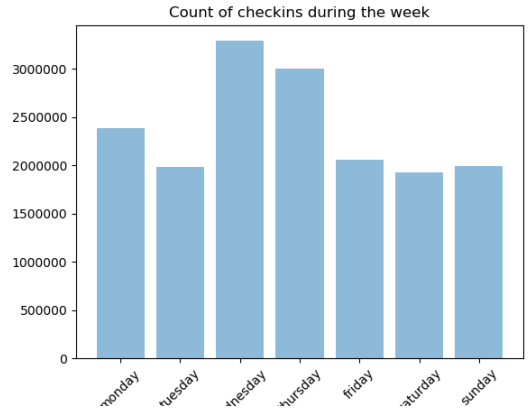

Figure 1: Count of checkins during the week

2.3 Data Mining and Visualization

Given that we are still coming to grips with the huge dataset at hand, we decided to initially see which days might be attracting a lot of customers. We plotted a bar graph which helped us to see which day might have the highest number of checkins in the given dataset. The figure above shows the bar graph plotted using the Matplotlib library.

We see that surprisingly Wednesday has the highest count of checkins. Further analysis on the dataset would reveal the time of day when we would have the highest checkins. We then hope to predict the customer checkin for a given restaurant. We may also pursue trying to predict the stars given by a customer, if time constraints are in check.

We could also use KMeans algorithm to cluster the days of the week when we had most checkins using the 'checkins'file. Given there are seven days of the week, we hope to have seven clusters each representing the day of the week and the time when it was the busiest. For the time of the day using the 'time'file, since we have built six time ranges covering the entire day, we hope to have six clusters.

References

[1] M. M. Hamad and A. A. Jihad. "An Enhanced Technique to Clean Data in the Data Warehouse". In: *2011 Developments in E-systems Engineering*. Dec. 2011, pp. 306–311. DOI: 10.1109/DeSE.2011.32.

[2] Dr. Arvind Kalia Sapna Devi. "Study of Data Cleaning and Comparison of Data Cleaning Tools". In: *International Journal of Computer Science and Mobile Computing* 4.3 (Mar. 2015), pp. 360–370.

[3] *Yelp Dataset*. 2013. URL: https://www.yelp.com/dataset%5C_challenge/.