Analysis of Yelp Dataset

Data Cleaning and Preparation Project

Akshay Karki (avk1063@rit.edu)
Gaurav Gawde (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 to be used

In this section, we will consolidate the data collected from various sources for Yelp [1]. Given the dataset and how it could be analyzed, we feel using Python to code and develop will be the right way to proceed. We also plan to use Python libraries like SciPy, Matplotlib, and Pandas for data visualization. A specific tool, Tableau, just for data visualization might be used to further enhance the project.

2.2 Data Cleaning and Analysis

The data collected from various sources almost never come clean. Hence we will diagnose data for problems and perform cleaning to prepare it for analysis. This is an important part of the project and so we plan to spend the most time here. We will primarily look for problems in data such as inconsistency in column names, missing data, outliers, duplicate rows, untidiness (special characters that need to eliminated etc.) and removing unwanted columns. Along with this, we will also be using some of the techniques mentioned in the papers [2] and [3]. The cleaned data will then be piped to the machine learning model for analysis.

2.3 Data Mining and Visualization

In this stage, we will build a machine learning model and provide visualization of predictions of the raw data and compare those to predictions of cleaned data. The model will perform an analysis which would help us predict several aspects of the dataset such as what time a particular restaurant is most busy based on reviews by the user or recommend the user a new set of restaurants based on his/her review and location etc. The visualized data might even be helpful to associated businesses so as to plan better and increase their revenues.

3. REFERENCES

[1] Yelp Dataset, 2013.

[2] M. M. Hamad and A. A. Jihad. An enhanced technique to clean data in the data warehouse. In *2011 Developments in E-systems Engineering*, pages 306–311, Dec 2011.

[3] D. A. K. Sapna Devi. Study of data cleaning comparison of data cleaning tools. *International Journal of Computer Science and Mobile Computing*, 4(3):360–370, March 2015.