STA 9760

Juna Iafelice

Midterm Project / Spring 2020

Intro Project phase

YELP REVIEWS

This project consists on a dataset of Yelp’s businesses, reviews, and user data. The datasets will have information about business across 11 metropolitan areas in four countries. The datasets is saved in JSON format. In total there are 8,000,000 user reviews, information on 174,000 business. All the data sets combined exceeds 10GB of data. There are five originals datasets. I have worked only on two of them Business and Review datasets (yelp\_academic\_dataset\_business.json, yelp\_academic\_dataset\_review.json). The main link where you can download the original datasets is on Kaggle site

<https://www.kaggle.com/yelp-dataset/yelp-dataset>

For both datasets (Business and Reviews) I will provide a shortened sample of them attached on the zip file. In the zip file you will find the python code, the summary document and a file with the most used words in English.

The main question I want to answer in this project is:

**Question 1: What are the top ten most used words grouped by user star rating reviews?**