Yelp Review Contents Ranking

Identifying Review Sub Categories and Details

BY SAM GUTENTAG
JANUARY 16, 2018

Introduction

This project will look to identify important features in review text and extract key terms. Ratings, Reviews and Tips on Yelp are useful to users trying to identify a new restaurant or bar to visit, but users need a way to see a step between the 5 Star Ranking and the possible hundred of reviews. ratings, and tips.

Proposal

The key client of the project is Yelp, with the intent that this would be extended to end users on the platform. This analysis focus on food centric locations and will use trends in reviews, tips, checkins, and ratings from real world users to identify key factors to create sub ratings for users to quickly view and get a better understanding of a business's attributes.

Yelp publishes and Academic Data set that will be used for this analysis consisting of
- 4,736,881 reviews
- 1,027,721 tips
- 1,183,362 users
- 156,639 businesses

This data is provided in well structured JSON formats but is in very large files. First steps will be wrangling the dataset and performing some preprocessing to get into a format desired for this analysis.

All of my analysis will be done using Python packages and delivered in a Jupiter Notebook hosted on my Github repository. The data which will be used in this project is publicly available through The Yelp Open Dataset site [1].