Digging In: Text Mining of Yelp Reviews for Subtopic Identification and Sentiment

BY SAM GUTENTAG
MARCH 2, 2018

Introduction

This project will look to identify important features and extract key terms from Yelp Reviews posted to Arizona Restaurants. Ratings, Reviews, and Tips created by Yelpers provide insight into what factors are most important to patrons and as such identifying key terms in Negative and Positive reviews will help uses better identify differentiating features between restaurants as well as provide business owners and managers a means to improve customer experiences and thus drive more sales.

Proposal

The key clients of this project will be both Yelp Users as well as Restaurant Owners. To narrow our context, this analysis will look at Reviews and Tips posted to restaurants in Arizona only. This will allow us to remove ambiguity of terminology and be more confident in the speech/text patterns of Yelp reviewers.

A final deliverable product will be a collection of important Sub Categories for ranking Restaurants such as ambiance, food quality, cleanliness, value, and possibly more. Additionally, rankings against other restaurants serving the same cuisine, a ranking against other restaurants in the zip code, and a ranking against other restaurants in the chain (if available) will be generated.

Data

Yelp publishes and Academic Data set that will be used for this analysis consisting of:

- 5,261,668 reviews
- 1,098,324 tips
- 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 things 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 a Github repository. The data which will be used in this project is publicly available through The Yelp Open Dataset site [1].