Data Analysis – Yelp Dataset

Sruti Jain (saj160430@utdallas.edu)
Spring Semester, 2017

Motivation

Yelp reviews have been a great source of reviews to customers, which help them choose the best businesses in their region. Currently, if a prospective entrepreneur wants to understand the market scenario, he has to read all the reviews in the region to get an idea about the demand in that region. This strategy is neither feasible nor accurate. However, we are looking to provide entrepreneurs trying to set up new business in a particular region with relevant suggestions for establishing a successful business. We have used topic modeling (LDA) on various yelp reviews to get users' preferences and taste in a given geographic location. Our model is able to predict the best matched region for a prospective entrepreneur.

Why It is Important?

In this project, we have found the inherent features and deduced their ratings which will help businessmen understand the requirement/demand of a region and find regions where chances of their business being a success are maximum. Thus, we have provided a holistic analysis of the region to them before starting their business.

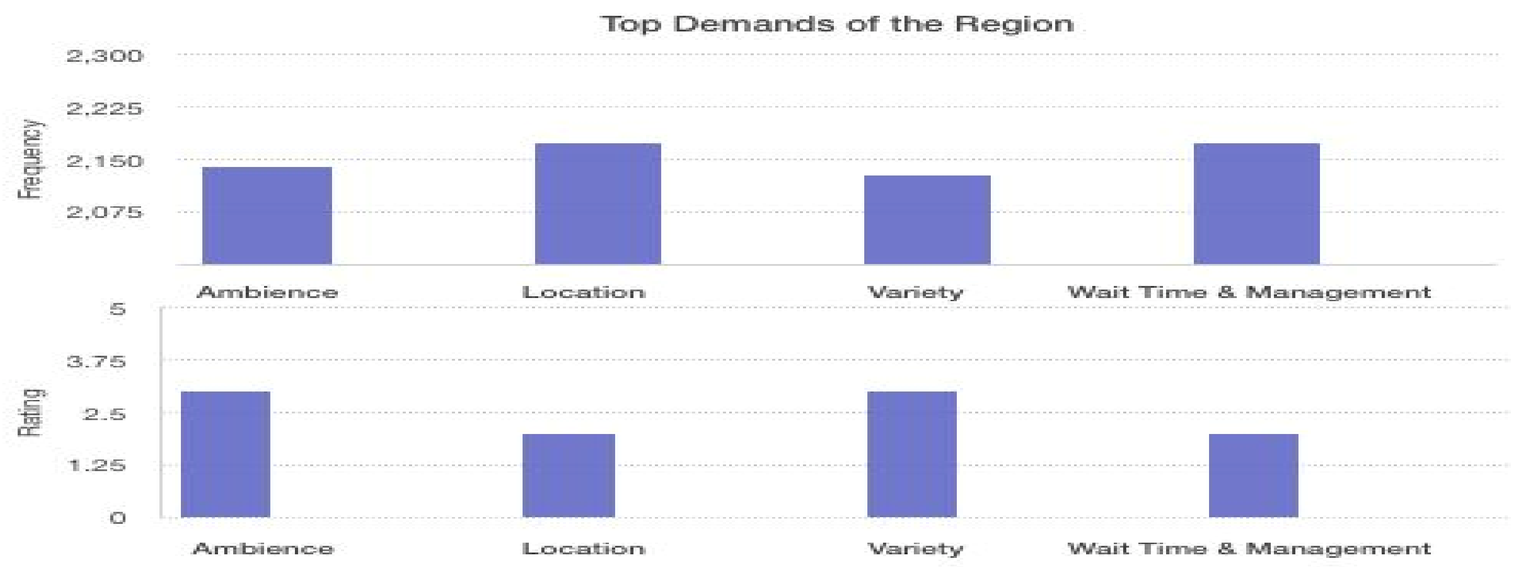
Data-set: Yelp

The Challenge Dataset includes data from **Phoenix**, **Las Vegas**, **Madison**, **Waterloo** and **Edinburgh**:

- **42,153** businesses
- **320,002** business attributes
- **31,617** check-in sets
- **252,898** users
- **955,999** edge social graph
- **403,210** tips
- **1,125,458** reviews

Results

Word Cloud of Latent Topics Obtained



Evaluation

1. The most important topics found after performing LDA are Ambience, Location, Variety and Wait Time & Management.

2. It can determined from the figure above that the central part of Phoenix lacks in providing Ambience and Service.

3. The project is aimed at extracting ratings for the individual sub topics & helping entrepreneurs with reviews rather than the customer. The project can be a recommendation system that matches the services offered by a restaurant with a locality and predicts the optimal locality for the restaurant.

Approach and Method

Data Collection and Normalization

Extracted reviews from json and imported to MongoDB collection - Reviews. Split each review into sentences, remove stopwords, extract parts-of-speech tags for all remaining tokens, filters out all words which are not nouns, use Lemmatizer/Stemmer to lookup lemma of each noun. Finally, store each review i.e. reviewId, business name, review text together with nouns' lemmas to new MongoDB collection called Corpus.

Topic Modeling (LDA)

Used Gensim LDA Model to perform LDA on filtered collections. This is a training process which takes number of topics to be generated as input and output for this process is Dictionary file (represents dictionary structure for words), and Corpus (saved in Market Matrix format).

Latent Topic Rating

Aggregated Ratings-Ratings for a given topic in a Review will be obtained by averages over all of these review ratings to get the hidden topic rating.