Northeastern University
CS6220 – Data Mining Techniques
October 2017, Emily Dutile, Yu Tian, Xiwen Song, Vyshaal Narayanam

Project Proposal

1. The Dataset

Today we live in a society where businesses are greatly impacted by customer reviews. For some businesses, this can contribute to their success and for others it can ruin them. Looking to improve the experience for restaurant owners and customers, we will be using the Yelp Open Dataset (https://www.yelp.com/dataset) in order to implement different data mining techniques to better understand these reviews. Although this dataset consists of 4.7 million reviews, 156,000 business, and 12 metropolitant areas, the scope of this project will focus on Boston and it's neighborhoods. The dataset is given as a set of JSON files related to bussiness information, checkins, photos, reviews, tips, and user information, but we will mostly focus on the restaurant reviews.

2. Questions

We look to address and analyze the following three questions:
(1) What should a restaurant focus on to make their rating/reviews better?
(2) What neighborhoods in Boston have the best cuisine selection?

To answer question one, we want to focus on what restaurants are doing wrong in order to provide feedback on how to improve. First we will distinguish the difference between good and bad reviews, and then find frequently found words in a negative review in hopes of giving us insight as to what's going wrong. In question two we will discover if there are a dense amount of cuisines in a particular Boston neighborhood and compare it to the ratings. This could give good recommendations to users as to where they can try their favorite cuisine at a few different restaurants in the neighborhood.

3.Work to Do

Before answer the question, we will do some work on data preparation and preprocessing.

3.1 Data preparation

We are going to use MySQL workbench to import yelp.sql dataset. Select all candidate observations according to the category (Restaurants) and location (Boston). Pick up restaurants with more than 200 reviews to make sure of the sample data size.

3.2 Data Preprocessing

After we have candidate dataset and observations, some preprocessing methods are going to be employed at this point.Following candidate methods will included:

- Tokenization

- Filtering
- Lemmatization
- Stemming

3.3 Information Extraction and Summarization

At this point, we are going to extraction the information from preprocessed data. Basically, in order to answer Question1, we need to figure out two things. (1)Keywords Extraction Picking up keywords of each sentence so that we can get known of what the customer is talking about in each sentence.
Candidate Algorithms:

- NER (Named Entity Recognition)
- HMM (Hidden Markov Models)
- CRF(Conditional Random Fields)

(2) Sentimental Analysis:
Analysis the sentiment of each sentence so that we can know if customer feel good or bad for some aspects of the restaurant.
Candidate Algorithms:

- KNN (k Nearest Neighbor)
- MaxEnt(Maximum Entropy Modeling)
- SVM (support vector machine learning)

noteAccording to some papers, corpus data might be more important than algorithms.

3.4 Topic model

To give business advice to restaurant, it should will be interesting to explore what topics are most likely to be covered in customers review. What's more, we can explore differences of topics of reviews for "good" restaurants" and "bad restaurants" so that we can focus on what should "bad" ones make improvements
Candidate Algorithm:

- LDA (Latent Dirichlet Allocation)

4. Team Work

Our first task will be related to preprocessing and cleaning the reviews in order to better understand a textual relationship between positive and negative reviews, and why they are such. We will also need to filter out the dataset we are looking to focus on, which is Boston neighborhoods.