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

# Project Proposal

## 1. Background

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 ([1]) 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 metropolitan areas, the scope of this project will focus on Boston and it's neighborhoods. The dataset is given as a set of JSON files and SQL that are related to business information, check-ins, photos, reviews, tips, and user information, but we will mostly focus on the restaurant reviews.

## 2. Goal and Outline

The goal of our project is to apply (mostly) unsupervised machine learning algorithms to answer some interesting questions about the dataset. We look to address and analyze the following questions:
(1) What topics are discovered frequently in reviews and do they correlate to a positive or negative review? 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 word association with respect to positive and negative reviews that are based on the number of stars. Overall we'd like to see if we can find through word association what restaurants may be doing wrong in order to provide feedback. 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. To go further in-depth, we will also look more closely at neighborhoods to see if we can determine a more touristy area in comparison to a more authentic area. This could help business owners determine who competitors are along with potential restaurant locations.

## 3. Preparation and Implementation

Before we begin the implementation of the algorithms, some work must be completed around the data preparation and preprocessing. The MySQL workbench will be used to import the yelp.sql data and then select all of the candidate observations according to the category (Restaurants) and location (Boston). To verify that we have a good sample of the data, we will pick restaurants with more than 200 reviews.

[1] `https://www.yelp.com/dataset`

### 3.1 Data Preprocessing

In order to answer question one, a bit of language processing will be needed. After we have the candidate dataset and observations, the following methods will be used:

- Tokenization
- Filtering
- Lemmatization
- Stemming

### 3.2 Information Extraction and Summarization

At this point, we are going to extract the desired information from preprocessed data. In order to answer question 1, 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:

- HMM (Hidden Markov Models)
- CRF(Conditional Random Fields)

(2) Sentimental Analysis:
Analysis the sentiment of each sentence so that we can know if a customer has positive or negative feedback about a restaurant. Even though the review is associating with rating, there may be some interesting differences between reviews and the rating.
Candidate Algorithms:

- KNN (k Nearest Neighbor)
- MaxEnt(Maximum Entropy Modeling)
- SVM (Support Vector Machine Learning)

Note: According to some papers, corpus data might be more important than algorithms.

### 3.3 Clustering

Since the Yelp data set has already identified the type of cuisine of a business, we will look to use the following algorithms and visuals for our data representations:

- K-Means Clustering
- Hierarchical Clustering

Note: in order to answer the question related to subgroups of neighborhoods to see if we can tell authentic from trendy, we may use some of our initial language processing work to associate reviews with this.

### 3.4 Topic model

To give business advice to restaurants, it will be interesting to explore what topics are most likely to be covered in customers review. We can explore the differences of topics of reviews for "good" restaurants and "bad" restaurants so that we can advise what negative areas need to be addressed.
Candidate Algorithm:

- LDA (Latent Dirichlet Allocation)

## 4. Team Work

Our first task will be related to preprocessing and cleaning the reviews. The entire group will participate in this task so that everyone has a clear understanding in the data. Peer to peer coding will take place during this process so that we can continuously discuss and verify our questions and approaches. We will look to work in two week sprints, with the scope of the first sprint being the language processing and data cleaning. Working in pairs, one pair of two will focus on the implementation and analysis of question one while the other pair focuses on question two.

Figure 1: Project Schedule