CSE6339

# Sentimental Analysis on Yelp Data

By: Rohit Bhoopalam (1001100534)
and
Pawan Puttaswamy (1001094272)

## Objective:

To perform sentimental analysis on Yelp business data and predict attributes that benefits business management.

1. On searching for a restaurant, to display a chart showing the positive and negative review trend for that restaurant over a period of time.
2. On selecting a city and category of the business, a set of charts will be displayed showing relative statistics for that city and that business category for that particular location, which can be helpful to decide the business location for an entrepreneur.
3. To display most liked and disliked food items of a restaurant based on the reviews. Only the food items that are listed in the menu of a particular restaurant are used to extract the food sentiments. This can be further extended to suggest the best restaurant around a user's location and the food item one is willing to eat (using the menu of the near-by restaurants).

## Overview:

Performing sentimental analysis based on reviews to predict whether a review is positive or negative. Location analysis for any business, based on category. Extracting most liked food items of a restaurant based on the reviews.

## Challenges:

1. To come up with an approach to categorize a review into a positive or a negative review.
2. Gathering relevant features to decide whether the location is suitable or not. We may have to cluster the location of the businesses to gather the statistics.
3. Extracting the food items from the reviews and predicting the reviewer's sentiment towards those food items into a positive or a negative food review. Then aggregating the results to show most liked and disliked food of a restaurant.

## Dataset:

Yelp Dataset (https://www.yelp.com/dataset_challenge/dataset) and we want to get restaurant menu using the Yelp API.

1. Review data:

{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "jQQfOa1QgIjz-ukcvzkmJg", "review_id": "G2xHQmHIx4krbxQEC857fw", "stars": 1, "date": "2009-07-15", "text": "This is an overpriced fast food restaurant that is 24 hours. So overpriced for what you get.", "type": "review", "business_id": "6gm0pYy-YQRJwtavzlXUJw"}

2. Business data:

{"business_id": "vcNAWiLM4dR7D2nwwJ7nCA", "full_address": "4840 E Indian School Rd\nSte 101\nPhoenix, AZ 85018", "hours": {"Tuesday": {"close": "17:00", "open": "08:00"}, "Friday": {"close": "17:00", "open": "08:00"}, "Monday": {"close": "17:00", "open": "08:00"}, "Wednesday": {"close": "17:00", "open": "08:00"}, "Thursday": {"close": "17:00", "open": "08:00"}}, "open": true, "categories": ["Doctors", "Health & Medical"], "city": "Phoenix", "review_count": 9, "name": "Eric Goldberg, MD", "neighborhoods": [], "longitude": -111.98375799999999, "state": "AZ", "stars": 3.5, "latitude": 33.499313000000001, "attributes": {"By Appointment Only": true}, "type": "business"}

## Methods:

1. a) Preprocessing of the text from the reviews like tokenizing and removal of stop words.
   b) Performing lexical analysis on the preprocessed reviews.
   c) Performing sentimental analysis on the tokenized reviews by taking features like unigrams, bigrams etc.
   d) Classifying the assessed review into a positive or a negative review.
2. a) Selecting features to decide on the location's suitableness.
   b) To summarize and display the findings for that location.
3. a) Getting food items from the restaurant's menu using Yelp API based on the restaurant's name.
   b) Performing sentimental analysis on the reviews for those food items.
   c) Aggregating the results to show most liked and disliked food items.

## Initial implementation:

1) Preprocessing of text from the yelp reviews
   a. Tokenizing the data
   b. Stop words removal

2) Dividing reviews into positive reviews and negative reviews based on yelp user ratings for building the model.
    a. 3.5 rating or more is considered as positive review
    b. Less than 3.5 is considered as negative review
3) Feature selection
    a. Unigram features
    b. Unigram + Bigram features
4) Built Naïve Bayes classifier on the both kinds of features and the accuracy for the Unigram + Bigram features was the best so far.

## Evaluation Plan:

1) Using the reviewer's rating, which is provided in the Yelp dataset we can evaluate whether the review is positive or negative.
2) We can manually evaluate the number of positive and negative responses for food items for few restaurants

## Initial stage Result:

1) Using unigram feature and Naïve Bayes classifier we achieved 68.7% accuracy.

```
train on 7540 instances, test on 2514 instances
accuracy: 0.687748607796
Most Informative Features
                refused = True              neg : pos    =     23.0 : 1.0
                 ripped = True              neg : pos    =     17.7 : 1.0
                 poorly = True              neg : pos    =     16.3 : 1.0
           unacceptable = True              neg : pos    =     15.0 : 1.0
                  worst = True              neg : pos    =     14.9 : 1.0
                   zero = True              neg : pos    =     14.4 : 1.0
                   a-ok = True              neg : pos    =     14.3 : 1.0
                 divine = True              pos : neg    =     13.7 : 1.0
             disgusting = True              neg : pos    =     13.3 : 1.0
                apology = True              neg : pos    =     13.0 : 1.0

real    3m22.890s
user    2m58.995s
sys     0m23.375s
```

2) Using Unigram and Bigram features we achieved 76.21% accuracy.

```
train on 7540 instances, test on 2514 instances
accuracy: 0.762132060461
Most Informative Features
                   worst = True              neg : pos    =     27.1 : 1.0
          ('two', 'stars') = True              neg : pos    =     26.3 : 1.0
         ('take', 'order') = True              neg : pos    =     22.3 : 1.0
       ('would', 'highly') = True              pos : neg    =     21.0 : 1.0
                flavorless = True              neg : pos    =     17.0 : 1.0
              unacceptable = True              neg : pos    =     16.3 : 1.0
          ('call', 'back') = True              neg : pos    =     16.3 : 1.0
   ('service', 'horrible') = True              neg : pos    =     15.7 : 1.0
       ('wanted', 'like') = True              neg : pos    =     15.0 : 1.0
      ('speak', 'manager') = True              neg : pos    =     15.0 : 1.0

real    3m40.740s
user    3m18.560s
sys     0m21.603s
```

## Tasks to be completed:

1. Improvising the sentimental analysis classifier.
2. Performing entity recognition to identify food items in the review and extracting the sentiment towards the item.
3. Extracting menu for each restaurant.
4. Feature selection to decide whether the location is suitable for business or not.

## Expected challenges ahead:

1. Finding features that improves the classification task, like giving higher weights to adjectives etc.
2. Finding the best split rating value to divide data into Positive review and Negative review. Example split rating value is 3.5, rating greater than 3.5 is considered as positive, otherwise negative.
3. To get menu for all the restaurant through Yelp API and merging the current data set with the data collected using the API.