CSE 6339 Project

Sentimental Analysis on Yelp data

By:

Pawan Puttaswamy (1001094272)

& Rohit Bhoopalam (1001100534)

Contents

[Objective 2](#_Toc418806306)

[Motivation 2](#_Toc418806307)

[Dataset 2](#_Toc418806308)

[Technologies Used 2](#_Toc418806309)

[MongoDB 2](#_Toc418806310)

[Python NLTK 2](#_Toc418806311)

[Python Flask 2](#_Toc418806312)

[Locu API 3](#_Toc418806313)

[D3.JS 3](#_Toc418806314)

[Implementation 3](#_Toc418806315)

[Task 1 - Generating Review Trends 3](#_Toc418806316)

[Performing Sentimental Analysis 3](#_Toc418806317)

[Task 2- Generating Positive and Negative Food items 4](#_Toc418806318)

[Task 3 – Extracting important business features 4](#_Toc418806319)

[Skyline 5](#_Toc418806320)

[Feature Selection 5](#_Toc418806321)

[Implementation 5](#_Toc418806322)

[Work Flow 6](#_Toc418806323)

[Results 6](#_Toc418806324)

[Dataset 7](#_Toc418806325)

[Yelp Business Data: 7](#_Toc418806326)

[Yelp Check-in data: 8](#_Toc418806327)

[Yelp Review data: 8](#_Toc418806328)

[Experiments: 9](#_Toc418806329)

[Task 1 9](#_Toc418806330)

[Conclusion 9](#_Toc418806331)

[References 9](#_Toc418806332)

# 

# Objective

To perform sentimental analysis on Yelp business data and predict attributes that benefits business management. To give businesses especially restaurants the details of how their business is performing, important features that can affect their business and the measures to be taken to improve their business.

# Motivation

Yelp provides information about businesses such as reviews, ratings, menus, location and contact details, which helps businesses & their customers. Providing more insights to those businesses are very important in order for them to improve their business. These features can be extracted by the customer’s feedback and other data that can be obtained by the businesses activity.

# Dataset

We are using dataset provided by Yelp for the Yelp dataset challenge[1]. Yelp data set contains

Data provided by Yelp was in JSON format

# Technologies Used

## MongoDB

MongoDB is one of many cross-platform document-oriented databases. Classified as a NoSQL database, MongoDB eschews the traditional table-based relational database structure in favor of JSON-like documents with dynamic schemas (MongoDB calls the format BSON), making the integration of data in certain types of applications easier and faster [2]. Updating the data or modifying the structure of the data is very easy and it doesn’t require all the rows to be updated, which makes it easier to use with unpredictable and ever increasing requirements.

## Python NLTK

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, and an active discussion forum [3]. We are mainly using word tokenizers, sentence tokenizers, parts of speech taggers, Naïve Bayes Classifier and Bigram collocation finder.

## Python Flask

Flask is a micro web application framework written in Python and based on the Werkzeug toolkit and a powerful Jinja2 template engine. Many popular applications that make use of the Flask framework are Pinterest, LinkedIn, as well as the community web page for Flask itself. Flask is called a microframework because it does not presume or force a developer to use a particular tool or library. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. [4]

## Locu API

The Locu API gives us access to business data, from opening hours, location to price lists, such as restaurant menus. On sending a web request to the Locu API with search terms like name, location, it returns the requested details that match our search query. There can be multiple results returned for the name search, it is then up to us to select the extract the exact details required. [5]

## D3.JS

D3.js (or just D3 for Data-Driven Documents) is a JavaScript library for producing dynamic, interactive data visualizations in web browsers. It makes use of the widely implemented SVG, HTML5, and CSS standards. It is the successor to the earlier Protovis framework. In contrast to many other libraries, D3 allows great control over the final visual result. [6]

# Implementation

## Task 1 - Generating Review Trends

To provide restaurants an overview of the number of positive and negative review trends over time, we perform sentimental analysis on the reviews provided by the users and give businesses an efficient way to view the cumulative rating over period of time.

### Performing Sentimental Analysis

Reviews data are provided by the Yelp in JSON format, we have put the review data in the MongoDB. We have removed stop words from the reviews and tokenized them using the NLTK’s word tokenizer. We have used these tokens to build a binary Naïve Bayes Classifier with the classes being positive and negative. To generate these positive and negative data we have used the actual rating that has been provided by the user.

We have kept 3.5 stars as a cut-off rating to be considered the review as a positive review. By separating the data like this helps us to split the data into positive and negative samples, using which we can build our desired binary classifier. Learning the actual review data is important for us to perform other operations information like most liked or disliked food items from the reviews. To perform these kinds operations, we assumed learning a model on the same data set should perform better than using a unsupervised method to perform sentimental analysis.

We tried testing our Naïve Bayes classifier with different kinds of features like using only unigram features which gave 68% accuracy, unigram + bigram features which gave 76% accuracy, unigram + bigram + trigram features which gave close to 76% accuracy. We got a better accuracy with unigram + bigram features, so we decided to use it as our final model.

We built our classifier on 50000 randomly selected reviews and we tested it on 1.5 million reviews. The results are shown in the below figure

|  |  |  |  |
| --- | --- | --- | --- |
|  | Actual class | | |
| Predicted  Class |  | Positive | Negative |
| Positive | 553484 | 492628 |
| Negative | 19918 | 503216 |

The result is shown in using D3.JS with a changeable time on X axis.

## Task 2- Generating Positive and Negative Food items

To generating positive and negative food items we are first fetching menu data from the Locu API. To search the menu we are using restaurant name, city and latitude and longitude. Many results will be returned by the API for a search query, to remove disambiguate, we are restricting the location to be within 10000 meters of the location that we have passed through the search query.

Using the menu items returned by the API, we are doing sentimental analysis on those sentences that contains these food items to find if the user has liked the item or not. We are displaying the cumulated statistics of a restaurant, which gives us the top liked and disliked food items.

## Task 3 – Extracting important business features

For this task we used the yelp check-in and business dataset. The data set provides several attributes such as number of check-ins for the business, number of reviews for a business, useful attributes, etc. With the help of these, we are extracting the top restaurants based on review count, check-in count and rating data. We used the skyline tuples problem algorithm to extract the top restaurants in the city, for the given category.

### Skyline

The skyline problem is to compute the best tuples from a set of ordered *d*-tuples. The name is originated from what the solution represented on 2d plane resembles the scene that urban buildings comprise. Skyline is one of the recommendation queries, and it is considering multi criteria. It is very interesting problem as well as very useful query. This problem has been being intensively studied for recent years [8]. Consider a database table of n tuples and m numeric attributes. The domain of each attribute has an application-specific preference order, with “better” values being preferred over “worse” values. We refer to any subset of k tuples in the table as a k-tuple group. Our objective is to find, for a given k, all k-tuple skyline groups, i.e., k-tuple groups that are not dominated by any other k-tuple groups [7].

The notion of dominance between groups is analogous to the dominance relationship between tuples in skyline analysis. A tuple t1 dominates t2 if and only if every attribute value review count, check-ins count and rating of t1 is either better than or equal to the corresponding value of t2 according to the preference order and t1 has better value on at least one attribute. The set of skyline tuples are those tuples that are not dominated by any other tuples in the dataset. Analogously the dominance relationship between two groups of k tuples each is defined by comparing their aggregates. To be more specific, we calculate for each group a single aggregate tuple, whose attribute values are aggregated over the corresponding attribute values of the tuples in the group [7].

### Feature Selection

Selection features for extracting the top restaurants was a tedious task. The dataset contains several attributes like review count, check-in time and number of check-ins for that hour, rank for each restaurant, business functioning hours, attributes such as parking availability, Accept credit cards, good for kids, Has TV etc. After analyzing, we came up with an idea to use all the available data to find the top businesses and to extract the top features from top 100 skyline businesses (Has parking, Good for kids etc.) for that category in the given city.

### Implementation

Initially we extracted all the required data from the given dataset. The given check-in data contains the check in count for each hour for a business. We calculated the check in count for each business. Once that was ready we linked the check in data with business data with the help of business id. After the data was ready we used the skyline algorithm mentioned above to extract the top skyline businesses for the given category and city. The main focus of this task is to help business and entrepreneurs with the stats they need to know in order to open a business in that location and category.

### Work Flow

Skyline Algorithm

Location (City)

Category of Business

1. Top skyline businesses
2. Top features from top 100 skyline businesses

### Results

The results shown below is for the Category: Active Life and for location (city): Charlotte

|  |  |  |  |
| --- | --- | --- | --- |
| **Restaurant** | **Review Count** | **Checkin Count** | **Rating** |
| Dowd YMCA | 54 | 822 | 4.0 |
| TEN Park Lanes | 93 | 783 | 4.0 |
| U.S. National Whitewater Center | 164 | 717 | 4.5 |
| Time Warner Cable Arena | 86 | 1130 | 3.5 |
| McMullen Creek Greenway | 15 | 116 | 5.0 |
| Ultimate CrossFit Bootcamp | 9 | 366 | 5.0 |

Figure 3.1

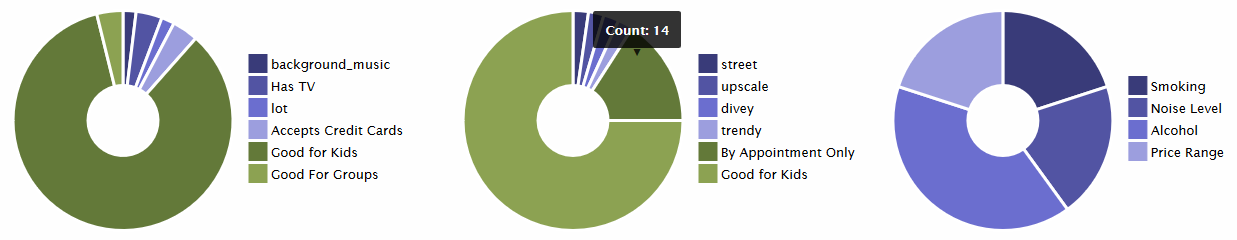


Figure 3.2

The above results shows that for category “Active Life” and for City “Charlotte” the top skyline businesses and top features in “Active Life” category. Figure 3.1 shows the top skyline businesses and the attributes associated with that business such as review count, Check-In count and rating. With the help of these data one can know the attributes that are influencing on the rating of a business. Figure 3.2 shows the top features from top 100 businesses in the given category and city. The first chart in the figure shows the stats for true attributes, for example the number of business with background music = “true”, similarly the second chart shows the stats for the number of businesses with street=”false” and the third chart shows count of other top attributes which are influencing the business excel. With the help of these charts and attributes one can know what attributes are important for the healthy functioning of a business, also what attributes contribute less for the business to excel. This would also help the entrepreneurs to decide what features include and where to open a business.

# Dataset

The dataset we used was provided by Yelp for yelp dataset challenge [1]. The Yelp dataset contains datasets for business, reviews provided by customers, business check-in information, user information and tip data. In the provided dataset, we are utilizing Business, Check-in and Review datasets for our project. The dataset is in json format [9].

## Yelp Business Data:

{"business\_id": "mVHrayjG3uZ\_RLHkLj-AMg", "full\_address": "414 Hawkins Ave\nBraddock, PA 15104", "hours": {"Tuesday": {"close": "19:00", "open": "10:00"}, "Friday": {"close": "20:00", "open": "10:00"}, "Saturday": {"close": "16:00", "open": "10:00"}, "Thursday": {"close": "19:00", "open": "10:00"}, "Wednesday": {"close": "19:00", "open": "10:00"}}, "open": true, "categories": ["Bars", "American (New)", "Nightlife", "Lounges", "Restaurants"], "city": "Braddock", "review\_count": 11, "name": "Emil's Lounge", "neighborhoods": [], "longitude": -79.866350699999998, "state": "PA", "stars": 4.5, "latitude": 40.408735, "attributes": {"Alcohol": "full\_bar", "Noise Level": "average", "Has TV": true, "Attire": "casual", "Ambience": {"romantic": false, "intimate": false, "classy": false, "hipster": false, "divey": false, "touristy": false, "trendy": false, "upscale": false, "casual": false}, "Good for Kids": true, "Price Range": 1, "Good For Dancing": false, "Delivery": false, "Coat Check": false, "Smoking": "no", "Accepts Credit Cards": true, "Take-out": true, "Happy Hour": false, "Outdoor Seating": false, "Takes Reservations": false, "Waiter Service": true, "Wi-Fi": "no", "Caters": true, "Good For": {"dessert": false, "latenight": false, "lunch": false, "dinner": false, "breakfast": false, "brunch": false}, "Parking": {"garage": false, "street": false, "validated": false, "lot": false, "valet": false}, "Music": {"dj": false}, "Good For Groups": true}, "type": "business"}

The business dataset contains information such as unique businessid, business name, business location (city, state, address, latitude, longitude), functioning hours and days, business category, attributes (Good for Kids, Price Range, Good for Dancing, Delivery etc.), review count, rank etc. We are using most of the data from business data.

## Yelp Check-in data:

{"checkin\_info": {"11-2": 1, "15-1": 3, "15-0": 1, "15-3": 1, "15-2": 1, "15-5": 1, "15-4": 3, "18-2": 1, "18-3": 1, "18-4": 1, "18-5": 1, "18-6": 1, "16-6": 1, "14-6": 1, "17-5": 1, "14-2": 2, "16-4": 3, "19-5": 1, "16-5": 2, "13-2": 1, "11-6": 1, "11-1": 1, "13-6": 2, "14-5": 1, "17-4": 4, "12-2": 2, "9-1": 1, "9-0": 1, "9-3": 2, "14-1": 2, "16-2": 2, "16-0": 2, "16-1": 2, "17-3": 2, "17-2": 1, "17-1": 3, "17-0": 2, "8-4": 1, "10-2": 1, "10-6": 2}, "type": "checkin", "business\_id": "P1fJb2WQ1mXoiudj8UE44w"}

The check-in data contains the check-in count for a business in hourly basis. We are using check-in data to compute the skyline businesses.

## Yelp Review data:

{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user\_id": "MwgMlBTQwf9MGNJrvq-pbw", "review\_id": "-TYJ7-\_xCgRdmuoQM0azgw", "stars": 2, "date": "2014-06-15", "text": "The front drive is beautiful and well kept, however the front view reminded me of an apartment complex. Once I entered I felt I had arrived. the inside lobby is fantastic. The flooring and other decorations are very nice. Inside is a grand stair going done to a bar and leading out to the pool/garden area. The pool area is wonderful and very well tended. They had live music outside and overall it was a fantastic place. I went inside and sat done (outside was full) and it took some time to get drinks but they were busy. We order a Sangria and when it arrived we had no idea what it was it tasted like a gin and tonic, but they assured us it was a Sangria. They took it off the bill. Once in our room I found that the toilet seat must be held up or it falls. Near the roof over the toilet is an intake vent that has never been cleaned it is covered in dust and what looks like maybe hair. I had one of the worst nights sleep because the pillow are either so flimsy that it is like sleeping without one or so overstuffed it is like sleeping with a large rock as a pillow. The furnishing in the room were plastic covered, I think it may have been meant to be like leather but they failed. They were also not comfortable. Overall it was not worth the price and I will not be coming back unless it is paid for by work or someone else. Stay away, go to the Biltmore if you want luxury and comforT.", "type": "review", "business\_id": "jPhmaY35qGP2qNc68viXPg"}

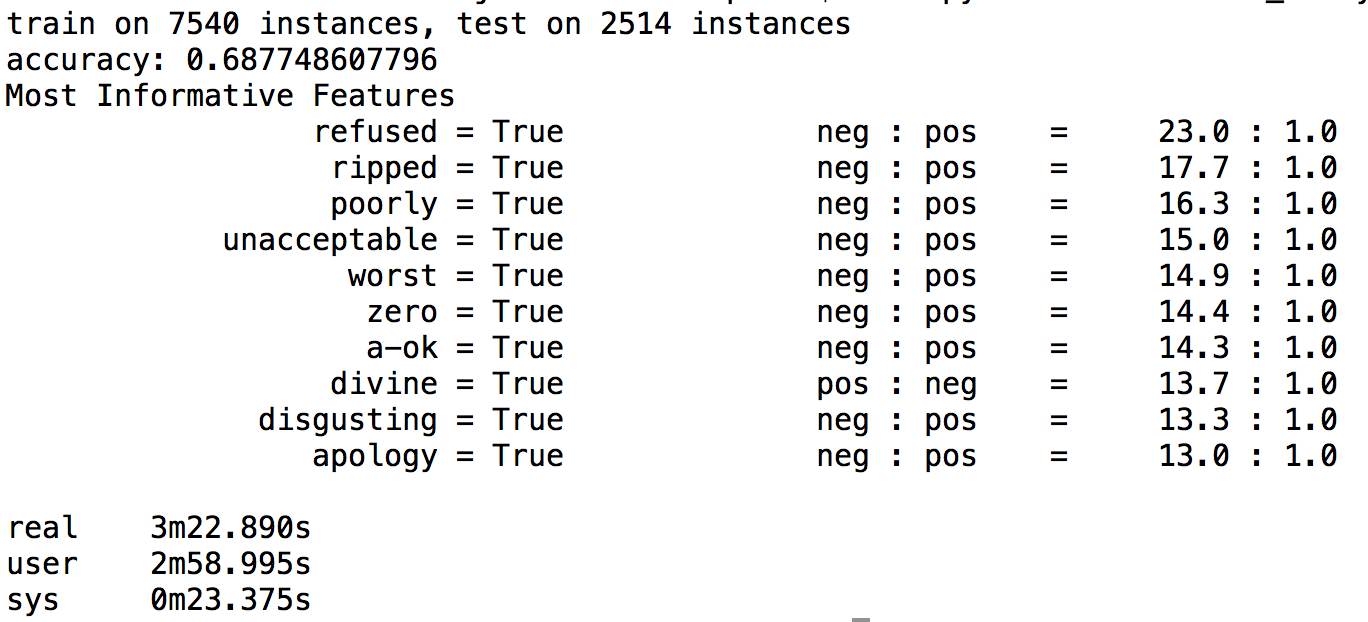
The review data contains reviews with respect to each business, date one which the review was written, review id, user id, votes etc. We are considering the review text for our sentimental analysis task and also for our top food extraction tasks.

# Experiments:

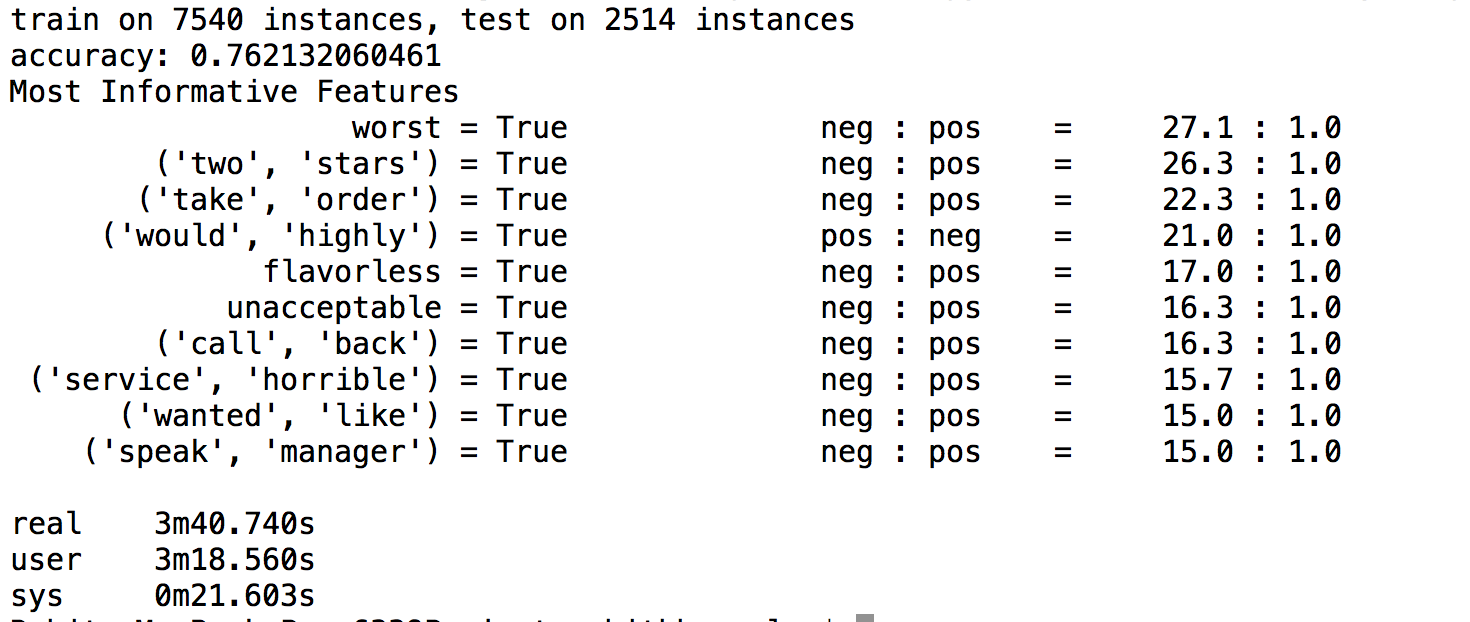
## Task 1

1. We are using ratings of the reviews to evaluate each method.
2. Reviews with rating of 3.5 and above are considered as Positive reviews
3. Less than 3.5 are considered as Negative reviews

#### Unigram Model Accuracy:



#### Unigram + Bigram accuracy



# Conclusion

The 3 tasks that we have come up with will help the business to predict the ratings of the customers based on reviews and also gives them the best food items liked by their customers expressed via reviews. And also gives them the idea of the most important attributes in order to excel their business.

# 

# References

[1] <https://www.yelp.com/dataset_challenge/dataset>

[2] <http://en.wikipedia.org/wiki/MongoDB>

[3] <http://www.nltk.org/>

[4] <http://en.wikipedia.org/wiki/Flask_(web_framework>)

[5] <https://dev.locu.com/>

[6] <http://en.wikipedia.org/wiki/D3.js>

[7] <https://cse.uta.edu/research/Publications/CSE-2012-2.pdf>

[8] <http://diveintodata.org/2009/09/06/a-brief-introduction-to-skyline-problem-pareto-optimal-tuples-1/>

[9] <http://www.json.org>