|  |
| --- |
| **Restaurant Menu Recommender: Yelp Dataset Review Sentiment Analysis** |

**Aditya Bhardwaj Karan Chaudhri Shubham Goyal**

[*abhardw2@ncsu.edu*](mailto:abhardw2@ncsu.edu) *krchaudh@ncsu.edu gsshubha@ncsu.edu*

**Abstract**

Thinking about which food item to order from the menu is a cumbersome task for any customer at a restaurant. In this project, we are performing sentiment analysis over the restaurant reviews given by users in the Yelp Dataset scoring every food item in the review with a value and creating list of food items in ascending order of score, top being the first food item. On basis of reviews and score given to that food item, we suggest 5 food items to try at restaurant.

**1 Introduction**

Restaurant reviews plays a vital role in selection of restaurant for any customer, the more important aspect comes when customer want to try some food at the restaurant. Such a task can be cumbersome and usually put a customer in frenzy on what to choose and what not to. To simplify such a task, we address the business problem where customer will be given food recommendation based on the restaurant reviews by other customers.

At current stage, recommendations are based on the

**1.1 Data set**

1. Business Data

Rows = 11, Columns = 8

2. Review Data:

Rows = 11, Columns = 9

3. Tips Data:

Rows = 24158, Columns = 6

4. Food Items List:

Food Items count = 4000+

**1.2 Tools used**

1. SQLite: For Cleaning and preprocessing of data.

2. Python and R: For running the algorithm and creating a wrapper app for Food recommendation.

**2 Research Papers related to Project**

**3 Business Questions**

**4 Methods**

**4.1 Data Cleaning**

The original data set contained many business, their information and review from various customers which was an overhead in our project which is focusing on the Restaurants. During data cleaning, we needed to remove the business overhead by removing all the business which does not fall under the category of restaurants.

As the data was preprocessed, so we didn’t have to worry about null values in the data but we ran the query to eliminate any null values from the data. After removing all other business categories, the data size was still significant. Extra fields in the tables were removed from each dataset in order to perform accurate analysis over the data.

After removal of unnecessary data from the tables, the tables contain following information in them:

Business Dataset: Business ID, Stars, City, State

Review Dataset: Business ID, Stars, Text

Tips Dataset: Business ID, Tips

As we are focusing only on Business, the dataset only contains information pertaining to evaluate reviews and tips for a particular business. The next steps before analysis includes sampling of data to make it manageable.

**4.2 Sampling**

As the data provided by the Yelp dataset was huge even after cleaning. Therefore, there was need to sample the data to make it more manageable for sentiment analysis. The reviews and business tables were sampled on state values and we extracted various tables for states including Illinois, South Carolina, North Carolina etc. to perform Sentiment Analysis more robustly over a smaller dataset.

Following queries were implemented in SQLite Browser to clean and sample the data:

|  |
| --- |
| CREATE TABLE businessClean AS  SELECT b.\*, yelp.city, yelp.state, yelp.latitude, yelp.longitude  FROM business AS b INNER JOIN yelp\_academic\_dataset\_business AS yelp ON b.business\_id = yelp.business\_id; |
| CREATE TABLE tipsCleaned AS  SELECT tip.business\_id, tip.user\_id, tip.text, tip.date  FROM tip INNER JOIN business  ON tip.business\_id = business.business\_id; |
| CREATE TABLE business\_SC AS  SELECT \* FROM  businessClean WHERE businessClean.state = 'SC'; |
| CREATE TABLE review\_SC AS  SELECT \* FROM  reviewCleaned INNER JOIN business\_SC ON reviewCleaned.business\_id = business\_SC.business\_id; |
| CREATE TABLE tip\_SC AS  SELECT \* FROM  tipsCleaned INNER JOIN business\_SC ON tipsCleaned.business\_id = business\_SC.business\_id |
| CREATE TABLE cleaned\_SC  SELECT f2.field1, r.text FROM foodlist AS f2, review\_SC AS r  WHERE r.text LIKE '%'||f2.field1||'%'; |

**4.3 Sentiment Analysis**

The sentiment analysis discussed under the Ruchi paper was based on the Stanford NLP toolkit which we are not using in the project. Our approach is based on the Sentiment Analysis as taught in the class using the NLP approach over Doc2Vec. We don’t have the set of positive and negative review for Doc2Vec to be implemented efficiently, so we adapted NLP approach based on the list of positive and negative words as provided in the Sentiment Analysis project.

After Sentiment Analysis, we are organizing positive and negative reviews for a particular food item in the list in dictionary based format, more the positive review for that particular food item, more would be its positive count. The organized data is saved in a JSON format for the convenience to be implemented in the wrapper application for faster recommendation to a user.

**4.4 Data Obtained after Sentiment Analysis**

Below is the structure of the data which is generated after performing Sentiment Analysis on the data. This structure would be utilized in a wrapper application to support our deductions on suggesting food items to the user.

|  |
| --- |
| {‘business\_id1’: {  “business\_id”: ‘business\_id1 value’,  “food\_items”: {  “food1”: {  pos\_review: [review text],  pos\_count: positive count value,  neg\_review: [review text],  name: ‘food name’,  neg\_count: negative count value  },  “food2”: …  }  },  ‘business\_id2’: …  } |

Figure Structure of the data created after sentiment analysis

**5 Team Roles**

**References**

[1] Buon Appetito - Recommending Personalized Menus: Michele Trevisiol, Luca Chiarandini, Ricardo Baeza-Yates, Yahoo Labs, Universitat Pompeu Fabra

[2] Ruchi: Rating Individual Food Items in Restaurant Reviews: Burusothman Ahiladas, Paraneetharan Saravanaperumal, Sanjith Balachandran, Thamayanthy Sripalan and Surangika Ranathunga

[3] Dataset: <https://www.yelp.com/dataset_challenge>

[4] Converting Yelp Data from JSON to CSV: <https://github.com/Yelp/dataset-examples>