**Yelp Review Critiques**

*Objective:*

To determine the count of positive and negative yelp reviews for a particular business.

*Overview:*

Using the Business data and review data provided in <http://www.yelp.com/dataset_challenge> we will perform text mining determine the sentimental value of businesses and categorize the reviews for the specified business name as positive or negative and will also provide the count for the same.

A GUI will be created through which the user can enter the business name and the business name will be sent to the model which in turn will fetch the reviews from the review data based on the business id determined by the business name which is common in both business and review data. The reviews will then be categorized as positive and negative based on the positive/negative words used in the reviews. A counter will be maintained for the positive and negative words. The count and the categorization will be displayed as result.

*Challenges of the project:*

**Data cleaning** – The Business data and the Review data is in JSON format. The data needs to be cleaned from all stop words, punctuations and digits. The input data must precise to determine satisfactory results.

**Categorizing the reviews –** The reviews for every business needs to be labelled as either positive or negative and the respective count has to be displayed on the GUI

**Listing positive and negative words in a review** – The reviews are categorized based on the sentimental value of the review. The value is obtained from the affinity file which is a collection of all positive and negative words which have predefined sentimental value.

*Architecture/components:*

The project is coded in python using Pycharm IDE which uses Anaconda. The file with all the reviews for various businesses is considered as the input data. All the reviews for the respective BusinessIDs are extracted and stored in Python dictionary as key-value pairs. Using NLTK, all the stop words, punctuations and the digits are removed from the input data. The values of the key “BusinessID” are compared with the affinity file to determine the sentimental value for that particular business. This helps us to categorize the reviews for any business as positive or negative. A counter for positive and negative reviews is initialized and incremented accordingly every time the categorization takes place.

*Methods:*

Data cleaning – Review file is read and converted to a dictionary consisting of key value pairs of each business id with all the other corresponding fields.The dictionary created is further converted to a dictionary having business id as key and all the reviews as its value. Punctuations and digits are removed. Stop word are removed using nltk stop word corpus and manually created stop word list.Any word that does not start with an alphabet is removed. Single letter words are removed. Multiple spaces are replaced with single space/

*Initial implementation:*

Once the data is cleaned, the reviews can be analyzed as positive or negative by determining the positive and negative words in each review. The categorization without the sentimental value for words is less effective.

*Clear evaluation plan*

* Data cleaning evaluated with small dataset of 3500 reviews.
* The result to be evaluated based on calculating the sentimental value of all the words in the reviews matching them with the affinity file.
* The reviews are then bifurcated into negative and positive reviews. It is evaluated by comparing the sum of positive and negative reviews count with the total reviews count.

*Challenges faced:*

The input data is a huge file of 1.32GB. The memory needs to be managed and the program needs to be written efficiently.

*Tasks to be accomplished:*

* Data integration : To map the business name from the business data file to every business id in the review data file.
* Sentimental Analysis : To compare the words in the review with an affinity list that has sentimental value assigned to every word in order to determine positive and negative count.
* GUI creation : To create GUI which will ask the user to select the business name.

*Sequence of steps to be implemented:*

* Analysis - perform sentimental analysis on the cleaned data
* Integration - associate the BusinessIds to the BusinessNames
* GUI creation

*Expected challenges ahead:*

Assigning apt sentimental value to the business reviews and classifying them as positive or negative reviews with the respective total count for the businesses.