**Scope**

This is a project for the Natural Language Processing Course taught in the graduate programm of the Computer Science department of the Univerisity of Texas at Daallas during the fall semester 2016.

**Goal**

The objective of this project is to apply various sentiment analysis techniques(NLP) on the restaurant reviews and assess if they can correctly identify the tone of the reviews as positive/negative/neutral.

**Data**

Yelp has publicly released a sample of their data (including over 2.7 million reviews) as part of their [Dataset Challenge](http://www.yelp.com/dataset_challenge/). This data can be used for the project as it is easy/quick to acquire. This solution uses the data sets provided by Yelp in the Yelp Dataset Challenge. It is available online at Yelp dataset, located at [1]

It includes the data

* 2.7M reviews and 649K tips by 687K users for 86K businesses
* 566K business attributes, e.g., hours, parking availability, ambience.
* Social network of 687K users for a total of 4.2M social edges.
* Aggregated check-ins over time for each of the 86K businesses
* 200,000 pictures from the included businesses

The data for the project was taken from following two files

* yelp\_academic\_dataset\_review.json
* yelp\_academic\_dataset\_business.json

**Tools and Technologies used**

Python, NLTK

**Proposed Solution**

**Defining Sentiment**

For the purpose of project, we define sentiment to be "a personal positive or negative feeling." Here are some examples:

|  |  |
| --- | --- |
| **Sentiment** | **Review** |
| Positive | The food here is very good. |
| Neutral | The ambience is okay and the food was usual. |
| Negative | I am never coming to this restaurant again. The food was tasteless. |

**High Level Steps**

The high-level sequence involved in processing is as follows:

1) Raw data collection from Yelp Dataset

2) Sentiment labeling

3) Transform into train/test sets for classifier

4) Bag of Words

5) Transform train/test sets for final classification by classifier

6) Adjust classifier and repeat until best model

**Dependencies**

Make sure you have the following libraries installed before running the code.

* [NLTK](http://www.nltk.org/)
* [Numpy](http://www.numpy.org/)

Also you must have installed the stopword corpora of NLTK. Run the following in a python console for NTLK downloader.

import nltk

nltk.download()

**Extracting reviews**

**This step must be done before running any of the classifiers below.**

You need to run the DataCreator file to extract the reviews for businesses of category restaurant and generate samples for each review class (pos/neg/neutral). The script creates three json files one for each class and a file which contains all the restaurant id and names.

python data\_Creator.py

(Make sure to have input data files in folder yelp\_dataset\_challenge\_academic\_dataset . For created sample files check yelp\_dataset\_challenge\_academic\_dataset folder)

Input Data files –

[yelp\_academic\_dataset\_business.json](https://github.com/snehabangar/Sentiment-Analysis-NLP/blob/master/ReviewAnalyzer/src/yelp/review/yelp_dataset_challenge_academic_dataset/._yelp_academic_dataset_business.json)

[yelp\_academic\_dataset\_review.json](https://github.com/snehabangar/Sentiment-Analysis-NLP/blob/master/ReviewAnalyzer/src/yelp/review/yelp_dataset_challenge_academic_dataset/._yelp_academic_dataset_review.json)

[ngram\_words.txt](https://github.com/snehabangar/Sentiment-Analysis-NLP/blob/master/ReviewAnalyzer/src/yelp/review/yelp_dataset_challenge_academic_dataset/ngram_words.txt)

[YELP\_Restaurant\_Categories.txt](https://github.com/snehabangar/Sentiment-Analysis-NLP/blob/master/ReviewAnalyzer/src/yelp/review/yelp_dataset_challenge_academic_dataset/._YELP_Restaurant_Categories.txt)

**Naive Bayes**

It trains one classifier for feature extraction filter (single words, stopwords removal, stemming, n-gram) and prints the predicted and actual rating for each restaurant along with the overall accuracy.

python main.py

**Maximum Entropy**

It trains maximum entropy classifier for feature extraction filter (single words, stopwords removal, stemming, n-gram) and prints the predicted and actual rating for each restaurant along with the overall accuracy.

python max\_entropy.py

**Results**

For detailed analysis and result, please check Project\_report.docx