# Web Analytics

# BIA-660A

# Topic – Context Extraction

**Name: Bhavik Lodha**

**Project File Names:** GetHotelLinks.py

ExtractReviews.py

FindContext.py

GetPhrases.py

HotelLinks.txt

HotelReviews.txt

contextphrase.txt

nonrelevant.txt

relevant.txt

**Description:**

**GetHotelLinks.py** – This script will extract the links from the yelp website and store in HotelLinks.txt file. It will go on extracting all the links till the mentioned page number in the script

**ExtractReviews.py** – This script will use the extracted links from HotelLinks.txt and open that link. After opening that link it will start extracting each reviews and it will go on till the mentioned page numbers. All reviews will be stored in HotelReviews.txt

**FindContext.py** – After extracting all the reviews, this script will use contextphrase.txt to find the similar phrases from each review. After finding similar phrases, it will give output the very next word of the found phrase. After getting this word, we tag this word using POSTagger and filter all the words with just the NOUN ‘NN’ Tag.

These entire nouns tagged words are passed through the nonrelevant.txt words list to filter out the non-relevant words. The final word is then saved into relevant.txt

**HotelLinks.txt** – This file stores all the Hotel Links from which reviews are to be extracted.

**HotelReviews.txt** – This file stores all the Hotel Reviews

**contextphrase.txt** – This file contains all the phrases which are to be used to find the context words. These were extracted from 30,000 reviews and then tested on another 30,000 reviews

**nonrelevant.txt** – These are the words which are out of context related to this project. All these words should be excluded from our final relevant words list.

**relevant.txt** – This file contains all the relevant words which we get as the final outcome after running the script FindContext.py

**GetPhrases.py** – This script was used to capture the phrases based on the context words and then used to find the new context words from other reviews from these phrases