**Assignment 1 Report**

**Regno: 1706038**

This assignment has been entirely done by me, to understand the calculations I took the help from the notes available at Nlp Notes : <https://gist.github.com/ttezel/4138642>

**Ex1:**

In the Ex1, I have used urllib and beautiful soup to get the data from the provided guardian link, once the data is received then all the text with in p tags is extracted using the regex function  
pos\_tagging is done using nltk post tagging function.

# Identification of Errors:

# 1. Pos taging is not identify punctuation marks

# 2. Some words are not identified correctly

Ex2:

A regex is written to extract the phone numbers from a user provided url, Ex2 also uses beautiful soup and url to fetch the data from the website.

Part b is not attempted.

Ex3:

I have a written added separate function to perform a different calculation, in the main method all the formatting has been done, to calculate the unigram I have used the notes above to the unigram. Calculations has been in the functions accordingly.

These are the following formulas I have used my code.

Calculating unigram probabilities:

P( wi ) = count ( wi ) ) / count ( total number of words )

Calculating bigram probabilities:

P( wi | wi-1 ) = count ( wi-1, wi ) / count ( wi-1 )

Laplace Smoothing

P ( wi | cj ) = [ count( wi, cj ) + 1 ] / [ Σw∈V( count ( w, cj ) ) + |V| ]