CSE556 NLP
Assignment 01

Date: 10 Sep, 2022
Deadline: 11:59pm 18 Sep, 2022 **Max Marks: 50**

General Instructions:

1. Allowed programming language: Python.
2. Use classroom discussion for any doubt. No query will be entertained through personal emails.
3. Each group member must do at least one of the following sections. But both should know the working of all the tasks. (Recommended: Divide the sections among yourselves.)
4. The assignment can be submitted in a group of a maximum of two members.
5. For plagiarism, institute policies will be followed.
6. You need to submit a report.pdf and code files (.py or .ipynb) in a single zip on google classroom with the following name: ***A1_Name1_Name2.zip***.
7. Mention methodology, helper functions, preprocessing steps, any assumptions you may have, and the contribution of each member in the report.

Dataset: We have curated a subsample of Twitter Sentiment Analysis dataset and this subsample is attached for your reference. (link)

Dataset description: It contains 3 columns: text, label and datetime. The text is an English tweet, label is a binary set where 0 means negative sentiment for the corresponding text and 1 means positive sentiment. The datetime column is a string of dates in the form "WEEKDAY MONTH DAY HOUR:MINUTE:SECONDS YEAR".

Classes: Positive (1) and Negative (0)

I. REGULAR EXPRESSION [20 Marks]

RegEx or Regular Expression, is a sequence of characters that forms a search pattern. Python has a built-in package called re, that you need to use for this part.

A. Report the following values for each class separately.
   a. average number of sentences and tokens. 2 marks
   b. total number of words starting with consonants and vowels. 2 marks

c. lowercase the text and report the number of unique tokens present before and after lower casing. 2 marks
d. count and list all the usernames. 2 marks
e. count and list all the urls. 2 marks
f. count the number of tweets for each day of the week. Eg Mon: 58, Tues: 20, Wed… 4 marks

B. You will be given a word x and a class label during the demonstration, and your programme must be able to output the following.
   a. total number of occurrences of the given word and sentences containing that word. 2 marks
   b. number of sentences starting with the given word. 2 marks
   c. number of sentences ending with the given word. 2 marks

II. TEXT PREPROCESSING [10 Marks]

Whenever we have textual data, we need to apply several preprocessing steps to transform text into numerical features that work with ML algorithms. The preprocessing steps for a problem depend mainly on the domain and the problem itself.

a. Tokenization 1 mark
b. Spelling correction 1 mark
c. Stemming/Lemmatization 1 mark
d. Punctuations removal 1 mark
e. Using regex remove stopwords 1 mark
f. Using regex remove extra whitespaces 1 mark
g. Using regex remove URL and HTML tag 2 marks

Choose one sentence from each class (positive and negative) and show its output after each preprocessing step (a-g). 2 marks

III. VISUALIZATION [10 Marks]

Data visualization shows how the data looks like and what kind of correlation is held by the attributes of data. A word cloud is a text visualization that displays the most used words in a text from small to large, according to how often each appears.

a. From the clean text obtained post preprocessing above, generate word clouds for both the classes. 6 marks
b. Compare the word clouds and report your observations. 4 marks

IV. RULE-BASED SENTIMENT ANALYSIS [10 Marks]

Sentiment Analysis is the process of 'computationally' determining whether a piece of writing is positive, negative, or neutral. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.

a. Using VADER (in-built package) retrieve a class label for every instance: 6 marks
   i. for preprocessed text (obtained in part II)
   ii. for raw text
b. Write a program for computing 'accuracy'. Report the accuracy for both preprocessed and raw text. (you need to write your own function that computes accuracy. DO NOT use the built-in accuracy function)
   4 marks

$$\text{Accuracy} = \frac{\text{correct classifications}}{\text{all classifications}}$$