Natural Language Processing
Assignment 01

Deadline: 07/09/2021 11:59:59 PM **Maximum Marks: 50**

Instructions:

- The assignment is to be attempted individually.
- Language allowed: Python
- You are allowed to use libraries such as NLTK for data preprocessing.
- For Plagiarism, institute policy will be followed. Refer: Academic Dishonesty Policy
- You need to submit README.pdf, Code files (it should include both .py files and .ipynb files), and Output.pdf.
- Mention methodology, preprocessing steps, and assumptions you may have in README.pdf.
- Mention your sample outputs in the output.pdf.
- You are advised to prepare a well-documented code file.
- Submit code, readme, and output files in ZIP format with the following name: A1_<roll_no>.zip
- Use classroom discussion for any doubt.

Task:

Write a python program that accepts as input text messages from the spam message classification dataset and performs the following tasks on the message field in the dataset.

1. Print the number of words starting with consonants and the number of words starting with vowels in all messages.
2. Compute the percentage of capitalized words in spam and ham messages.
3. List all email ids and phone/mobile numbers and their respective counts in the dataset. Compute percentage for spam and ham messages for both email and mobile numbers.
4. List all monetary quantities, e.g., £1000. Compute the percentage of spam and ham messages with monetary quantities.
5. Count and print all emoticons (use NLTK only) in the dataset.
6. Print a list of words with clitics (a subword unit that cannot stand on it own, e.g., ‘ve, n’t, ‘d)
7. Print the messages and number of messages starting with a given word as a custom input.
8. Print the sentences and number of sentences ending with a given word as a custom input.
9. Spam/Ham classification
   – Utilizing the features in 2, 3, and 4, identify the category of the message.
     Note: For a given message, identify whether it contains features as per points 2, 3, and 4. Propose heuristics ([only 2], [only 3], [only 4], [2,3], [2,4], [3,4], or [2,3,4]) to obtain the maximum accuracy for the spam/ham classification. You should report accuracy for all cases.

Task [1-8] will be evaluated on random input (not necessarily on the messages from the dataset.) Given a word (string) and a .txt file as input, print the count of that word and sentences containing that word in the input file.