NLP Assignment - 1 Monsoon 2020
ReadMe File

This file contains all the assumptions which are made for completing the assignment.

1) For counting the number of words and number of sentences, I am using the libraries “word_tokenizer” and “sent_tokenizer”.

2) For counting the vowels and consonants, I am considering two regex for both vowels and consonants for dealing with lower and upper case alphabets. After tokenizing the words, trying to match the words with the regex defined and corresponding counters have been incremented.

3) For getting email ids considered, I have written two regular expressions and corresponding outputs are also posted in the output file. These regex are matched with the list of words obtained using split() method.

4) After taking a required word with which the sentence has to start with, I am iterating on the list of sentences obtained using sent_tokenizer and matching with the first word.

5) Similar to the above task.

6) For counting the words, I am also considering some other options of occurrences of the same word. For example, consider the word ‘natural’, I am also taking the occurrences like ‘natural.’ , ’natural? ’, ’natural!’ , ’<natural>’ , ’natural,’ , ‘natural..’ . But for counting the occurrences in sentences, if ‘natural’ is present twice in a sentence, I am considering this as only one increment of sentence as both the occurrences are in the same sentence.

7) For getting questions, I am taking the sentences obtained by sent_tokenizer and checking whether the last symbol is a Question mark or not. If the last symbol is a question mark, then I am considering it as a question.

8) After getting the words matching with the regex, trimming down the second and third item of time as they are the minutes and seconds.

9) I am using the spacy library and AbbreviationDetector to do this task. The objects are trained using a pre-trained model and the results are obtained based on the library.

References:-

- ➔ https://www.nltk.org/api/nltk.tokenize.html
- ➔ https://www3.ntu.edu.sg/home/ehchua/programming/howto/Regexe.html
- ➔ https://pypi.org/project/scispacy/
- ➔ https://stackoverflow.com/questions/201323/how-to-validate-an-email-address-using-a-regular-expression

****