NLP Assignment - 1 Monsoon 2020

This file contains the screenshots of the output for the questions in the assignment.

Task 1:- Count of words and sentences

```
[9]  if(len(filedata)>0):
       words_list = word_tokenize(filedata)
       #print(words_list)
     print("The number of words in the file \'{}\' is:: {}".format(file1,len(words_list)))
     sent_list = sent_tokenize(filedata)
     #print(sent_list)
     print("The number of sentences in the file \'{}\' is:: {}".format(file1,len(sent_list)))
```

```
The number of words in the file '101725' is:: 373
The number of sentences in the file '101725' is:: 16
```

Task 2:- Count of words starting with consonants and vowels

```
[10] vowel_word_count = 0
     consonant_word_count = 0
     vowel_pattern_upper = '^[AEIOU]'
     vowel_pattern_lower = '^[aeiou]'
     consonant_pattern_upper = '^[BCDFGHJKLMNPQRSTVWXYZ]'
     consonant_pattern_lower = '^[bcdfghjklmnpqrstvwxyz]'
     for i in words_list: #iterating over every word and checking for the pattern
         if(re.match(vowel_pattern_upper,i) or re.match(vowel_pattern_lower,i)):
             vowel_word_count = vowel_word_count+1
         if(re.match(consonant_pattern_upper,i) or re.match(consonant_pattern_lower,i)):
             consonant_word_count = consonant_word_count+1
     print("\n The number of words in the file  \'{}\' starts with vowels :: {}".format(file1,vowel_word_count)
     print(" The number of words in the file  \'{}\' starts with consonants :: {}".format(file1,consonant_word_
```

```
The number of words in the file  '101725' starts with vowels :: 63
The number of words in the file  '101725' starts with consonants :: 197
```

Task 3:- Email id's present in file

```
[11] email_pattern = '[a-zA-Z0-9-_.<\W]+@[a-zA-Z0-9-_.>\W]+'
     #email_pattern = '^[a-zA-Z0-9_.+-]+@[a-zA-Z0-9-]+\.[a-zA-Z0-9-.]+$'
     words = filedata.split()
     print("THE LIST OF E-MAILS PRESENT IN THE FILE::")
     for i in words:
         if(re.match(email_pattern,i)):
             print(i.strip('<>'))
```

```
THE LIST OF E-MAILS PRESENT IN THE FILE::
1993Mar22.215141.28352@mri.com
jeff@mri.com
jeff@mir.com
C41soE.M62@ns1.nodak.edu
C41soE.M62@ns1.nodak.edu
wilken@plains.NoDak.edu
wilken@plains.nodak.edu
WILKEN@PLAINS
jeff@mri.com
```

```
[12] #email_pattern = '[a-zA-Z0-9-_.<\W]+@[a-zA-Z0-9-_.>\W]+'
     email_pattern = '^[a-zA-Z0-9_.+-]+@[a-zA-Z0-9-]+\.[a-zA-Z0-9-.]+$'
     words = filedata.split()
     print("THE LIST OF E-MAILS PRESENT IN THE FILE::")
     for i in words:
         if(re.match(email_pattern,i)):
             print(i.strip('<>'))
```

```
THE LIST OF E-MAILS PRESENT IN THE FILE::
jeff@mri.com
jeff@mir.com
wilken@plains.NoDak.edu
wilken@plains.nodak.edu
jeff@mri.com
```

Task 4:- Sentences starting with a given word

```
[13] sentence_start=0
     start_word = input("Enter the required word to be matched with begining of sentence:: ")
     for i in range(len(sent_list)):
         individual_words_list=[]
         individual_words_list = word_tokenize(sent_list[i])
         if(individual_words_list[0]==start_word):
             sentence_start=sentence_start+1
     print("THE NUMBER OF TIMES THE WORD \'{}\' OCCURED AT THE BEGINNING OF THE SENTENCE IN THE FILE {} IS::
```

```
Enter the required word to be matched with begining of sentence:: in
THE NUMBER OF TIMES THE WORD 'in' OCCURED AT THE BEGINNING OF THE SENTENCE IN THE FILE 101725 IS:: 0
```

Task 5:- Sentences ends with a given word

```
[14] sentence_end=0
     end_word = input("Enter the required word to be matched with ending of sentence:: ")
     for i in range(len(sent_list)):
         individual_words_list=[]
         individual_words_list = word_tokenize(sent_list[i])
         if(individual_words_list[-1]==end_word):
             sentence_end=sentence_end+1
     print("THE NUMBER OF TIMES THE WORD \'{}\' OCCURED AT THE BEGINNING OF THE SENTENCE IN THE FILE {} IS
```

```
Enter the required word to be matched with ending of sentence:: of
THE NUMBER OF TIMES THE WORD 'of' OCCURED AT THE BEGINNING OF THE SENTENCE IN THE FILE 101725 IS:: 0
```

Task 6:- Count of words and sentences containg the word

```
[15] word_count_t6 = 0
     sen_count_t6 = 0
     req_word = str(input('Enter the required word: '))
     for i in words_list:
       if( i==req_word or i==req_word+"." or i==req_word+"?" or i==req_word+"!" or i=="<"+req
             word_count_t6 = word_count_t6+1

     for i in sent_list:
       if((" "+req_word+" " in i) or (" "+req_word+"." in i) or (" "+req_word+"?" in i) or ("
         sen_count_t6 = sen_count_t6+1
         #print(i)

     print("The count of the word \'{}\' in the file {} IS: {}".format(req_word,file1,word_co
     print("The count of the sentences with the word \'{}\' in the file {} is: {}".format(req
```

```
Enter the required word: in
The count of the word 'in' in the file 101725 IS: 3
The count of the sentences with the word 'in' in the file 101725 is: 3
```

Task 7:- Questions present in the file

```
[16] print("The questions present in the given file are:- ")
     count_t7 =0
     for i in sent_list:
       if(i[-1]=='?'):
         print(i)
         count_t7=count_t7+1
     print()
     print("The number of questions present in the file is: {}".format(count_t7))
```

```
The questions present in the given file are:-
Path: cantaloupe.srv.cs.cmu.edu!das-news.harvard.edu!husc-news.harvard.edu!kuhub.cc.ukans.
Newsgroups: rec.motorcycles
Subject: Re: Lexan Polish?
>
>Can anyone recommend a polish to use on it that is safe for lexan?

The number of questions present in the file is: 2
```

Task 8:- Minutes and Seconds

```
[17] regex_t8 = '[0-9][0-9]:[0-9][0-9]:[0-9][0-9]'
     print("The minutes and seconds in the file \'{}\' are:- ".format(file1))
     for word in words_list:
       if re.match(regex_t8,word):
         time = word.split(':')
         print("{} min, {} sec for the time {}".format(time[1],time[2],word))
```

```
The minutes and seconds in the file '101725' are:-
51 min, 41 sec for the time 21:51:41
00 min, 00 sec for the time 07:00:00
```

Task 9:- Abbrevations present in the file

Reference:- https://pypi.org/project/scispacy/

```
[20] #pip install scispacy
```

```
[32] import spacy
     from scispacy.abbreviation import AbbreviationDetector
     print("The abbreviations in the file \'{}\' are:".format(file1))
     #loading the pre-trained model for abbrevations
     abbrevation_t9 = spacy.load("en_core_web_sm")

     abbrev_pipe = AbbreviationDetector(abbrevation_t9)
     abbrevation_t9.add_pipe(abbrev_pipe)

     doc = abbrevation_t9(filedata)
     print(doc._.abbreviations)
```

```
The abbreviations in the file '103209' are:
[]
```

****