**REPORT FOR ASSIGNMENT 4**

1. **Your report should list the number of sentences, number of words, number of non-stopword unique words, and the relative frequency distribution of the top 20 non-stopword words for each file.**

The number of sentences and words in file 1865-Lincoln.txt are **26** and **779** respectively.

The number of sentences and words in file 1945-Roosevelt.txt are **26** and **633** respectively.

The number of non-stopwords (including punctuations) in 1865-Lincoln.txt are **447** .

The number of non-stopwords (with no punctuations) in 1865-Lincoln.txt are **369.**

The number of unique words (with no punctuations) in 1865-Lincoln.txt are **280**.

The number of non-stopwords (including punctuations) in 1945-Roosevelt.txt file are **358** .

The number of non-stopwords (with no punctuations) in 1945-Roosevelt.txt file are **284.**

The number of unique words (with no punctuations) in 1945-Roosevelt.txt file are **216**.

The relative frequency distribution of the top 20 non-stopword words is

1865-Lincoln.txt 1945-Roosevelt.txt

|  |  |
| --- | --- |
| war | 12 |
| God | 6 |
| Shall | 5 |
| years | 4 |
| Union | 4 |
| Let | 4 |
| The | 3 |
| would | 3 |
| interest | 3 |
| right | 3 |
| His | 3 |
| may | 3 |
| us | 3 |
| offenses | 3 |
| must | 3 |
| less | 2 |
| occasion | 2 |
| address | 2 |
| four | 2 |
| public | 2 |

|  |  |
| --- | --- |
| We | 11 |
| shall | 7 |
| peace | 6 |
| learned | 5 |
| I | 4 |
| today | 4 |
| men | 4 |
| test | 3 |
| The | 3 |
| way | 3 |
| simple | 2 |
| courage | 2 |
| democracy | 2 |
| presence | 2 |
| fellow | 2 |
| God | 2 |
| days | 2 |
| work | 2 |
| war | 2 |
| achieve | 2 |

1. **You should manually examine the results of sentence segmentation and stopword removal to comment on the respective performance of NLTK functions. For example, the percentage of sentences that are correctly segmented for each file.**

The sentences in both the files are 100% sentence segmented by using the sent\_tokenize and the stopwords are removed completely using the **NLTK** function such stopword and tokenize. The number of words are determined through the word\_tokenize function in both the files and calculated, but it may not be accurate because when I executed the program it contains 779 words in the 1865-Lincoln.txt file but when I looked up in the word file it shows that it contain 699 words. Similarly when I executed the program it contains 633 words in the 1945-Roosevelt.txt file words but when I looked up in the word file it shows that it contain 571 words. Hence the word\_tokenize in the NLTK function is not accurate.

1. **Based on the word-level analysis you conducted, can you answer the following question: how are the two files similar to each other and how are they different? Present your answer in the report.**

The two files are similar in terms of number of sentences, both files have 26 sentences. But the numbers of words differ after removing the stop words from both the files. Also both of the files contain different number of unique words like the number of unique words (with no punctuations) in 1865-Lincoln.txt are 280 and thenumber of unique words (with no punctuations) in 1945-Roosevelt.txt file are 216.These two file are related to the category of "news". Here when we compared the frequency distribution of the top 20 words in both the files, there are some common words like "God", "war", "shall" and "the" with different frequency distribution, but the word "The" contain same frequency Distribution number.

|  |  |
| --- | --- |
| **1865-Lincoln.txt file** | **1945-Roosevelt.txt file** |
| The - 3 | The-3 |
| God - 6 | God-2 |
| shall - 5 | shall-7 |
| war-16 | war-2 |

**You should report any problems and issues you have for the assignment, and your solutions to them.**

When trying to import nltk package in idle of python, it gave an error like package cannot be found. Before that I have came to know that we should also install pip in order to install nltk. For that I had to set the path in the environment variable of my system properties so that it can be recognized at the command prompt of my system. Then I have installed the nltk package using the commands import nltk and nltk.download() in idle. The python version is 3.4 and the package is downloaded in python 3.5 32-bit version. So I installed the nltk package in the location of python 3.5 lib/site-packages folder.