**MACHINE LEARNING**

**Ques :2**

|  |  |
| --- | --- |
| **2.1) Find the number of characters, words and sentences for the mentioned documents. (Hint: use .words(), .raw(), .sent() for extracting counts)** | **3** |
| **2.2) Remove all the stopwords from the three speeches. Show the word count before and after the removal of stopwords. Show a sample sentence after the removal of stopwords.** | **3** |
| **2.3) Which word occurs the most number of times in his inaugural address for each president? Mention the top three words. (after removing the stopwords)** | **3** |
| **2.4) Plot the word cloud of each of the three speeches. (after removing the stopwords)** | **3** |

**2.1) Find the number of characters, words and sentences for the mentioned documents. (Hint: use .words(), .raw(), .sent() for extracting counts)**

**Ans:-** We have loaded the all the required packages. In this case we have worked on the inaugural corpora from the nltk in Python. We will be looking at the following speeches of the Presidents of the United States of America:

President Franklin D. Roosevelt in 1941

President John F. Kennedy in 1961

President Richard Nixon in 1973.

We have downloaded inaugural corpus. There are in total 59 text files in this document.

'1789-Washington.txt',

'1793-Washington.txt',

'1797-Adams.txt',

'1801-Jefferson.txt',

'1805-Jefferson.txt',

'1809-Madison.txt',

'1813-Madison.txt',

'1817-Monroe.txt',

'1821-Monroe.txt',

'1825-Adams.txt',

'1829-Jackson.txt',

'1833-Jackson.txt',

'1837-VanBuren.txt',

'1841-Harrison.txt',

'1845-Polk.txt',

'1849-Taylor.txt',

'1853-Pierce.txt',

'1857-Buchanan.txt',

'1861-Lincoln.txt',

'1865-Lincoln.txt',

'1869-Grant.txt',

'1873-Grant.txt',

'1877-Hayes.txt',

'1881-Garfield.txt',

'1885-Cleveland.txt',

'1889-Harrison.txt',

'1893-Cleveland.txt',

'1897-McKinley.txt',

'1901-McKinley.txt',

'1905-Roosevelt.txt',

'1909-Taft.txt',

'1913-Wilson.txt',

'1917-Wilson.txt',

'1921-Harding.txt',

'1925-Coolidge.txt',

'1929-Hoover.txt',

'1933-Roosevelt.txt',

'1937-Roosevelt.txt',

'1941-Roosevelt.txt',

'1945-Roosevelt.txt',

'1949-Truman.txt',

'1953-Eisenhower.txt',

'1957-Eisenhower.txt',

'1961-Kennedy.txt',

'1965-Johnson.txt',

'1969-Nixon.txt',

'1973-Nixon.txt',

'1977-Carter.txt',

'1981-Reagan.txt',

'1985-Reagan.txt',

'1989-Bush.txt',

'1993-Clinton.txt',

'1997-Clinton.txt',

'2001-Bush.txt',

'2005-Bush.txt',

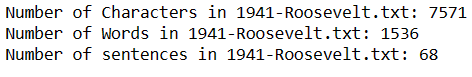
'2009-Obama.txt',

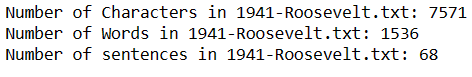
'2013-Obama.txt',

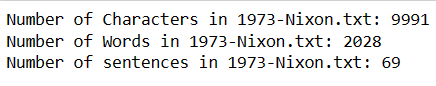
'2017-Trump.txt',

'2021-Biden.txt'

Here we are dealing with 1941-Roosevelt.txt, 1961-Kennedy.txt and 1973-Nixon.txt







**2.2) Remove all the stopwords from the three speeches. Show the word count before and after the removal of stopwords. Show a sample sentence after the removal of stopwords.**

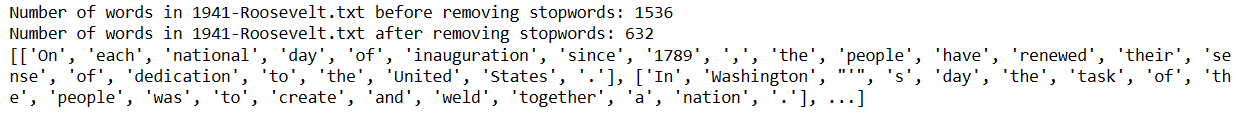
**Ans:-**  Defining a variable 'stopwords' which contains the list of punctuations from the string library and the english stopwords from nltk.

Below is the list of stopwords and punctuations.

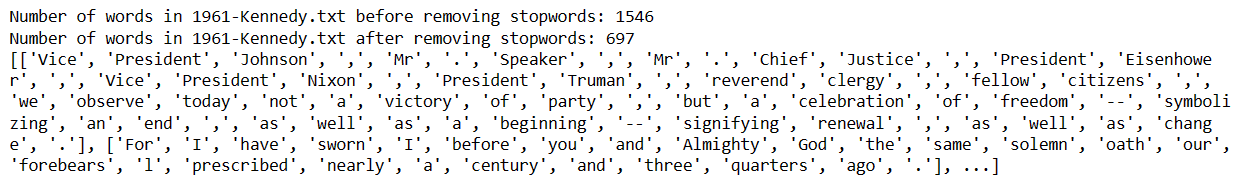
Before removing stopwords from each text file we have first converted the file into lowercase.

Below is the result for each file along with the sample sentence after removal of the stopwords.

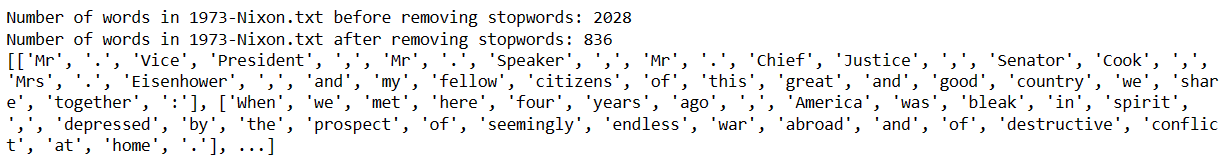
1. We can see that there are 1536 words before removing stopwords and 632 words after removing stopwords from 1941-Roosevelt.txt.



1. We can see that there are 1546 words before removing stopwords and 697 words after removing stopwords from 1961-Kennedy.txt.



1. We can see that there are 2028 words before removing stopwords and 836 words after removing stopwords from 1973-Nixon.txt.



**2.3) Which word occurs the most number of times in his inaugural address for each president? Mention the top three words. (after removing the stopwords)**

**Ans:-**

1. Below is the list of word frequency in 1941-Roosevelt.txt.



1. Top three words in 1941-Roosevelt.txt are.



1. Below is the list of word frequency in 1961-Kennedy.txt.



1. Top three words in 1961-Kennedy.txt are



1. Below is the list of word frequency in 1973-Nixon.txt.

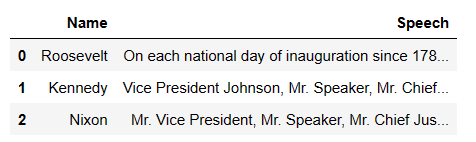


1. Top three words in 1973-Nixon.txt are

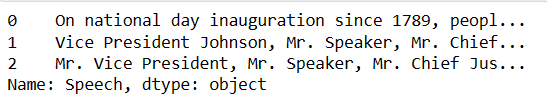


**2.4) Plot the word cloud of each of the three speeches. (after removing the stopwords)**

**Ans:-** We have loaded Project\_Speech file.

****

After removing stopwords from Speech column,the file looks like





In above Word cloud, there are all the Clean words from all three speeches after removing stopwords from them.