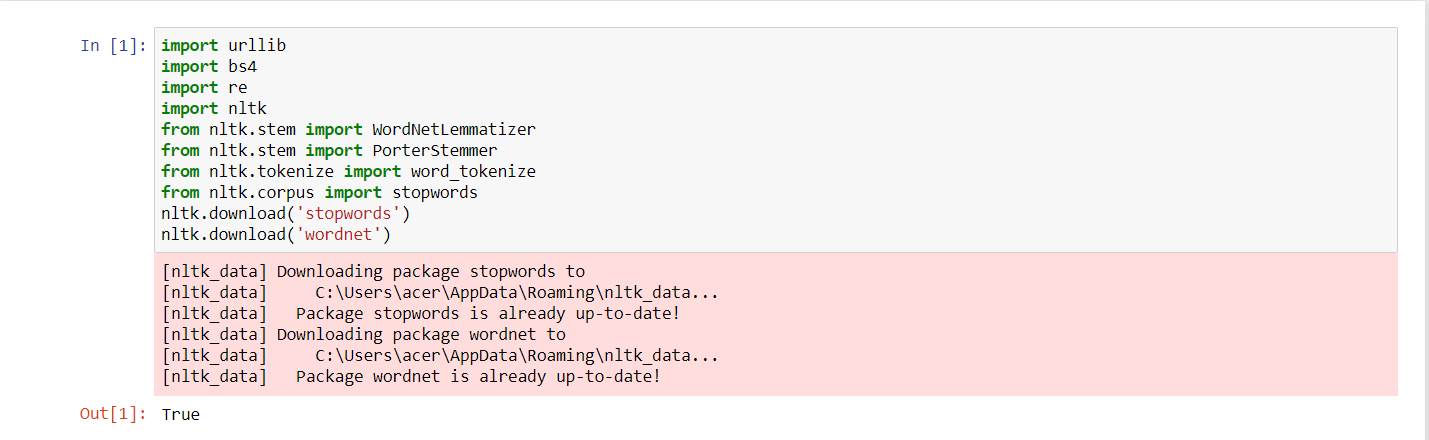
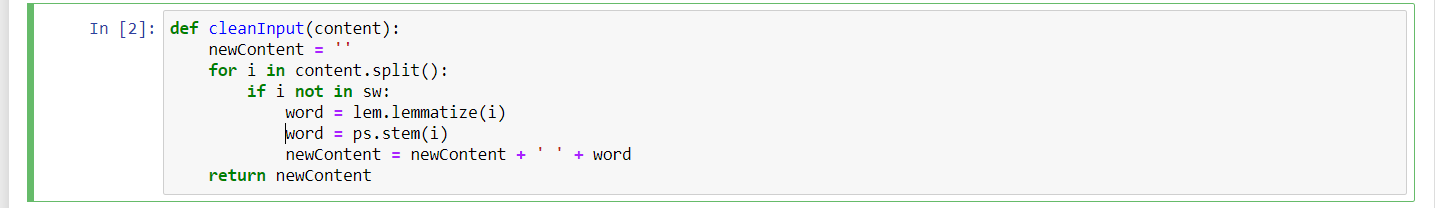
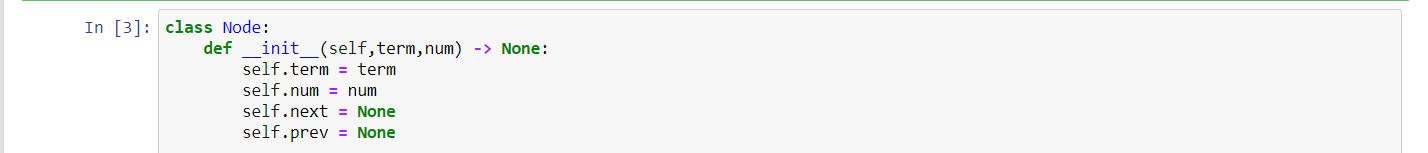
**Web Mining CSE3024**

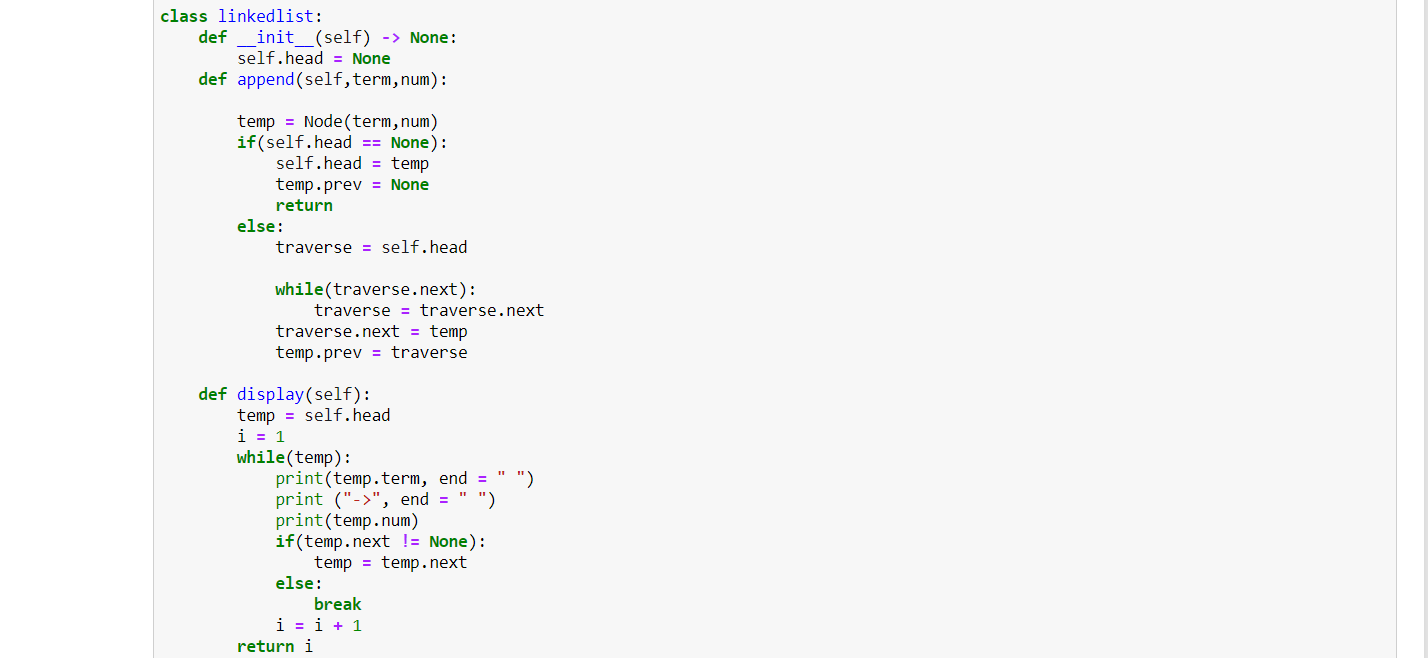
|  |  |
| --- | --- |
| **Student Name:** | **Abhishek Yadav** |
| **Student ID:** | **19BCE1308** |
| **Date:** | **02-08-2000** |

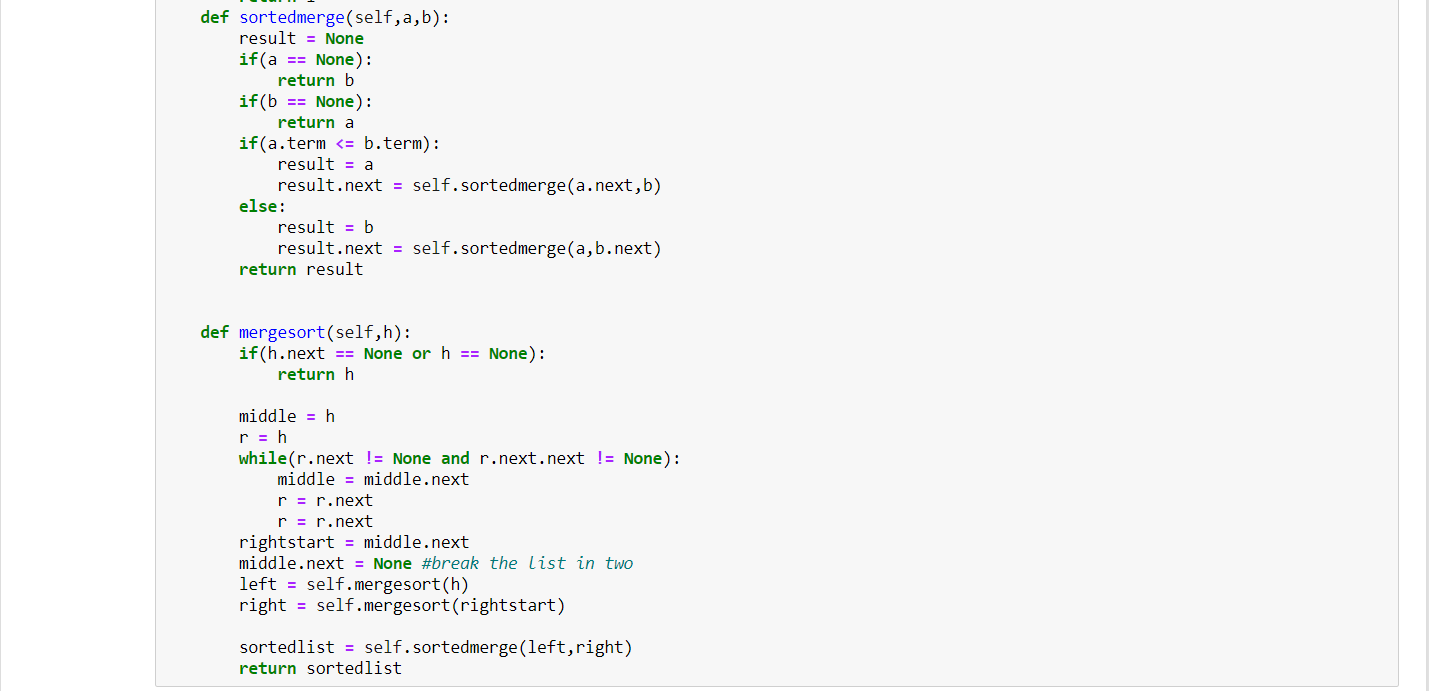
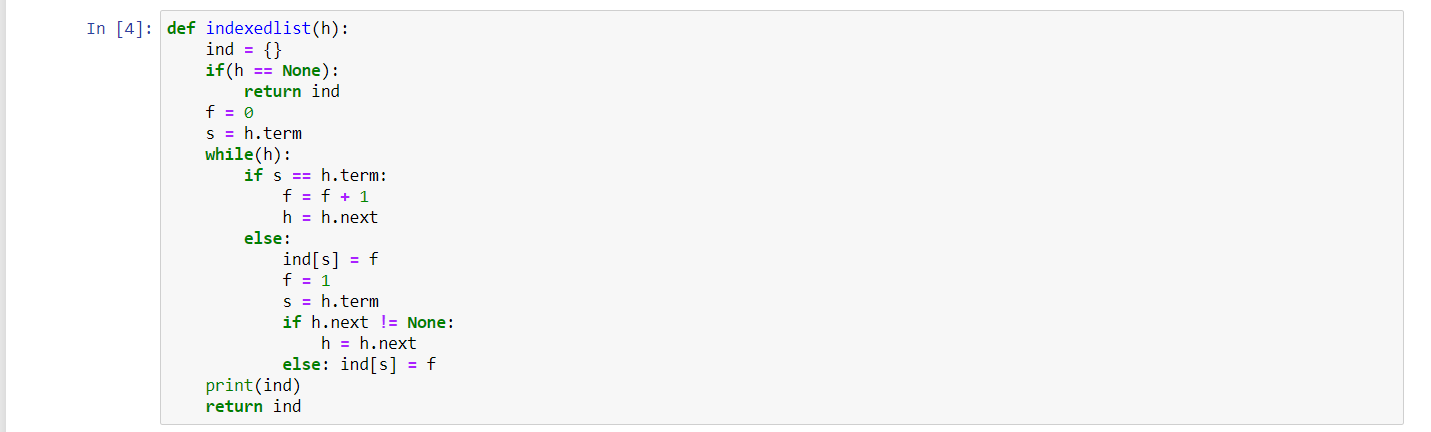
EXP NO. 4  
Pre-processing of Data

The Functions used and libraries used:



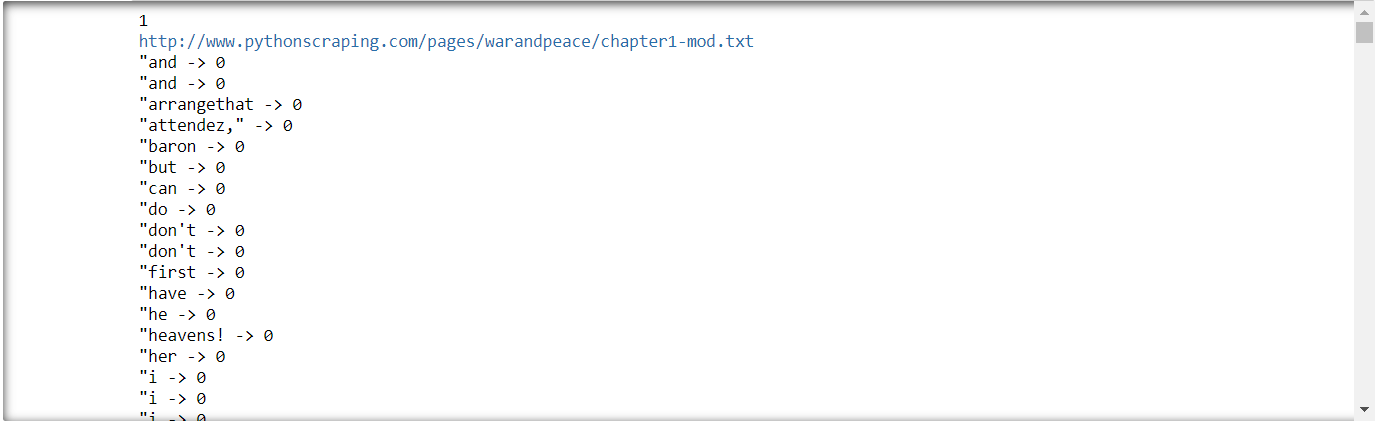


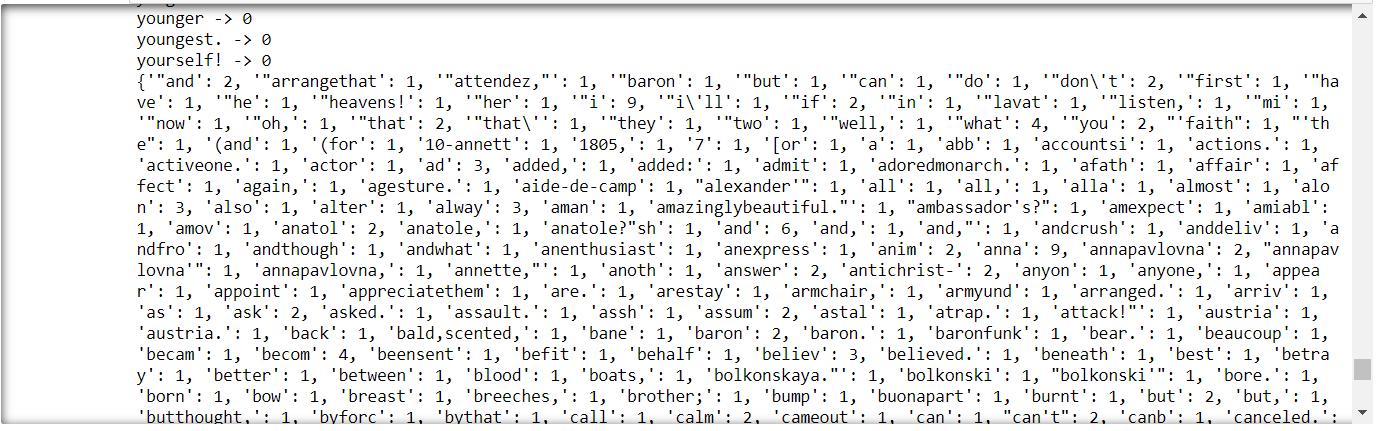
 

Collect any 10 documents (English text documents) from the web and create inverted index by doing necessary preprocessing steps using python.



Output:





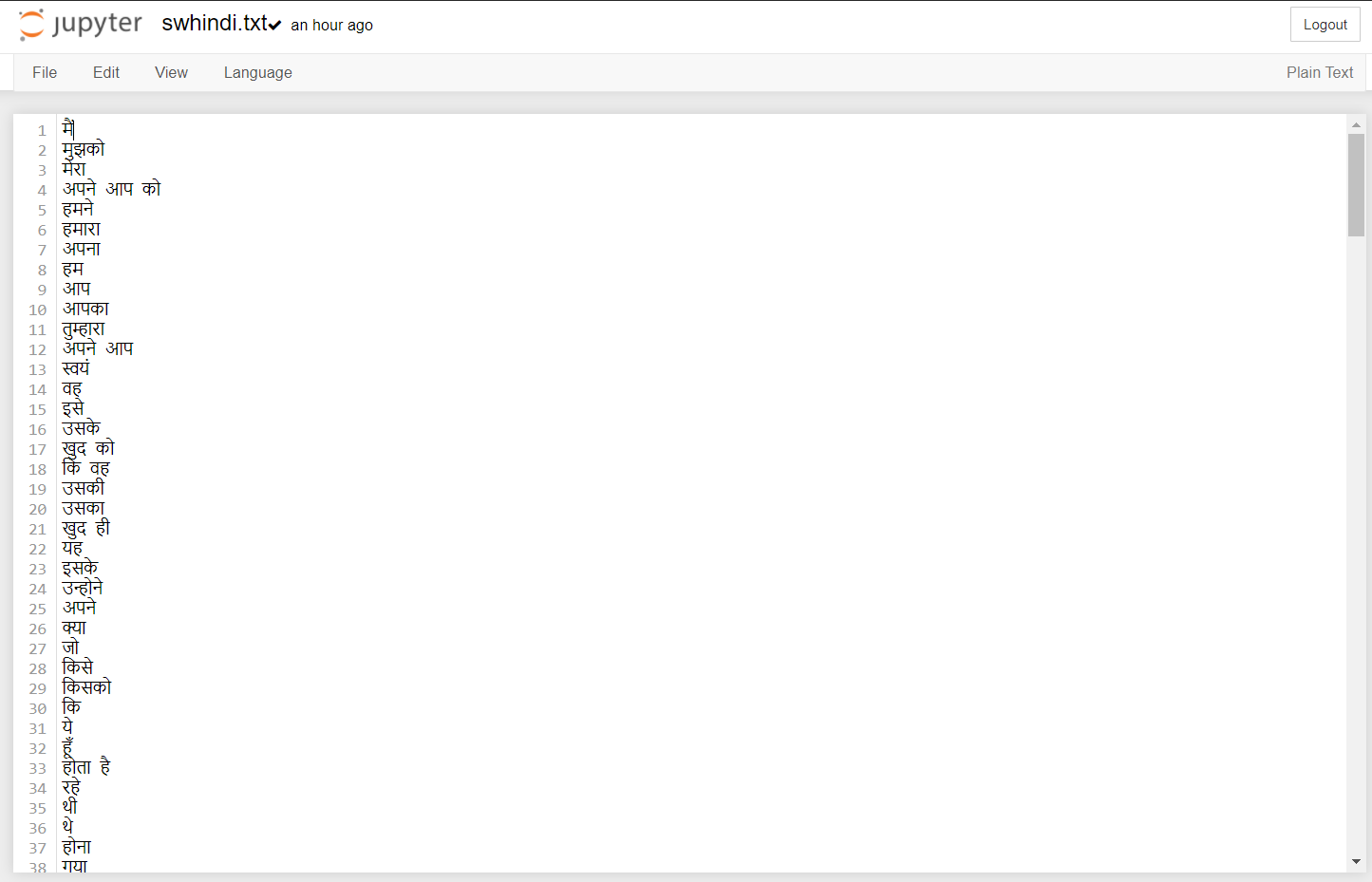
(i sampled this outputs only on one txt file I found on: <http://www.pythonscraping.com/pages/warandpeace/> .)

Collect any 10 documents (Indian Language text Documents in Unicode) from the web and create inverted index by doing necessary preprocessing steps using python.

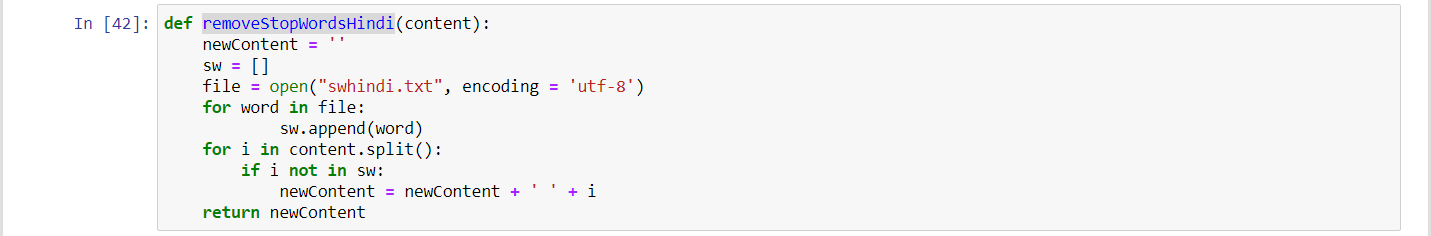


Since this is not English, lemmatization and stemming won’t work and mess with the real words we don’t use those functions for this one: -

For stop words in hindi, I made a file with some said stop words : -



Hence the new clean input function for hindi will replaced with: -



The new main function would look like: -



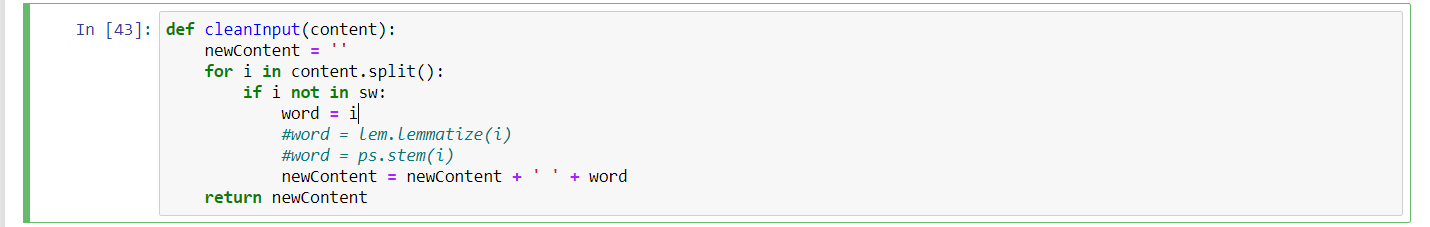
Output:

**I wasn’t able to find any sample hindi text file on the internet and wasn’t able to try this code on doc files but I tried this on vs code by copying some random hindi text and saving them in a txt file and it worked for that.**

Collect any 10 documents (Documents in different formats such as PDF, DOC, ODF) from the web and create inverted index by doing necessary preprocessing steps using python.



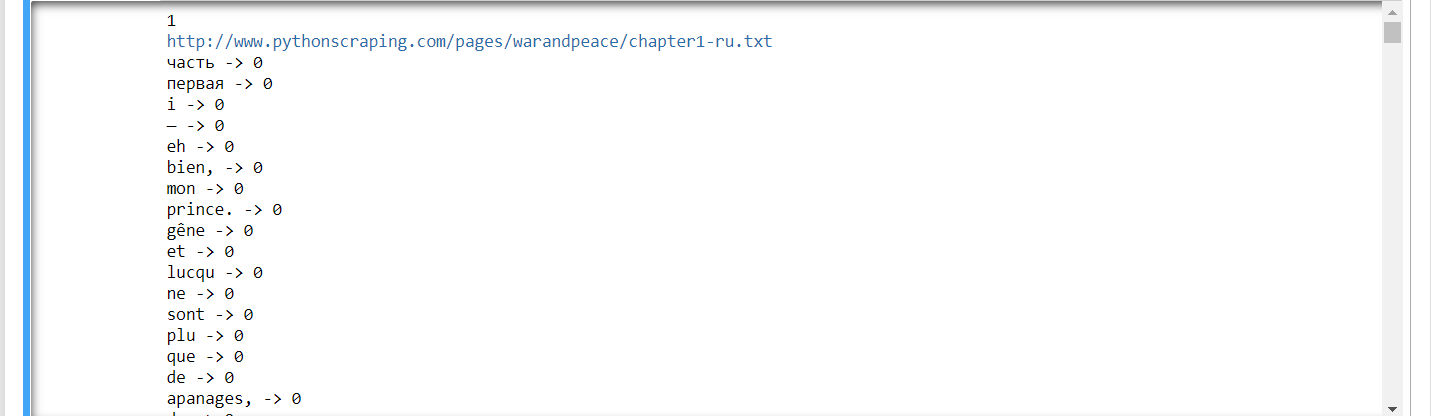
Since this is not English, lemmatization and stemming won’t work and mess with the real words we don’t use those functions for this one: -

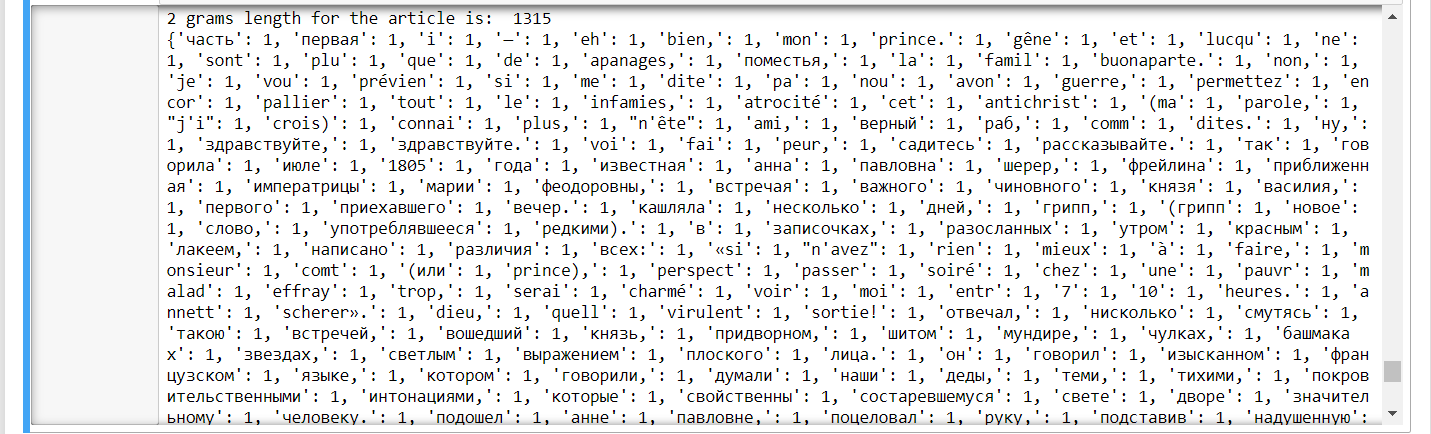


Although for Russian, nltk does support stop words, we can update the main function like: -



Output: -





(i sampled this outputs only on one txt file I found on: <http://www.pythonscraping.com/pages/warandpeace/> )

**thank you**