Pre-processing:

- Convert the string to lower case.
- Remove the digits. For this used the module “re”
- Remove the punctuations. For this imported the “string” module. Used translator to remove the punctuations.
- Remove the stop words.
- Apply stemming. Used the module “nltk.stem.snowball” , “EnglishStemmer”.
- Merge the stemmed list.

Stop Words:

Removed the top 30 (30 t0 50) terms from index list. The size of positional index has reduced. As most frequent terms have been removed.

Modules:

os: used to get list of files in dictionary (os.listdir()). Used in “task 1” and “task 2”.

sys: used to read the command line arguments. Used in “task 1”, “task 2” and “task 3”.

re: It is regex module. Used in preprocessing of text. Used in “task 1”, “task 2” and “task 3”.

string: used to remove the punctuations during pre-processing. Used in “task 1” and “task 2”.

nltk.stem.snowball.EnglishStemmer: Used for stemming. Used in “task 1”, “task 2” and “task 3”.

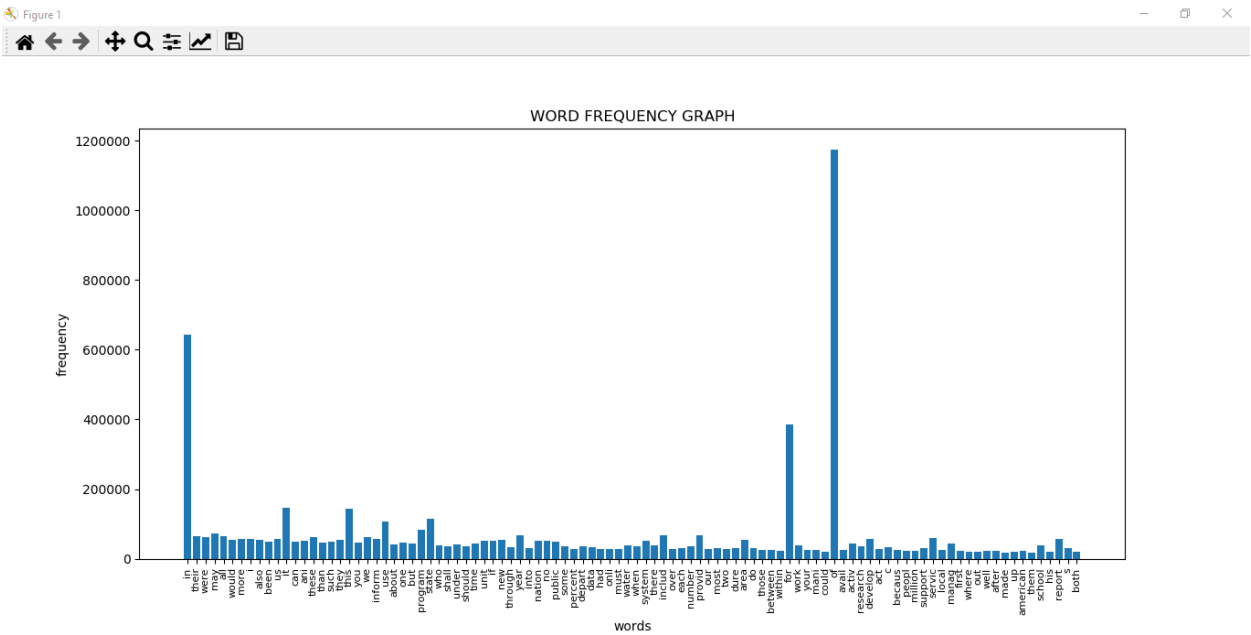
Linecache: Used to read specific line number in text file. Used in “task 3”.

matplotlib.pyplot: Used to show graph. Used in “task 4”.

math: Used to take log of frequency. Used in “task 4”.

Word Frequency Graph:

Only first 100 words printed.



Word Log Frequency Graph: