# **Design Document for Creating a profession specific information retrieval system**

The application created by the group is a profession specific search engine created based on the:

**The program/application can be broken down into the various subparts (actual file names also added) :**

1. **code.py:**

* Stores the tokenized words of each document as lists and then the corresponding list is stored in a json file.
* Stores all the unique words present in the corpus
* creates a dictionary which contains the words in the vocabulary as the key and the value as another dictionary.

### **Order of executing the files.**

$ sudo python3 code.py

## **Installation:**

Run the following in terminal.

$ sudo pip install -r requirements.txt

$ jupyter-notebook

If you face any problem, install nltk separately

**Installing nltk**

$ pip3 install nltk

$ python3

>>> import nltk

>>> nltk.download()

Packages: all

## **DATA STRUCTURES USED:**

### **Document\_tokens\_list (named as *files*)**

Contains lists enclosed within a list It will contain the stemmed tokens from each file in the corpus as individual lists. All are appended to make a list. Example:

[[‘i’,’am’,’good’],[‘thank’,’you’],[‘india’,’is’,’best’]]

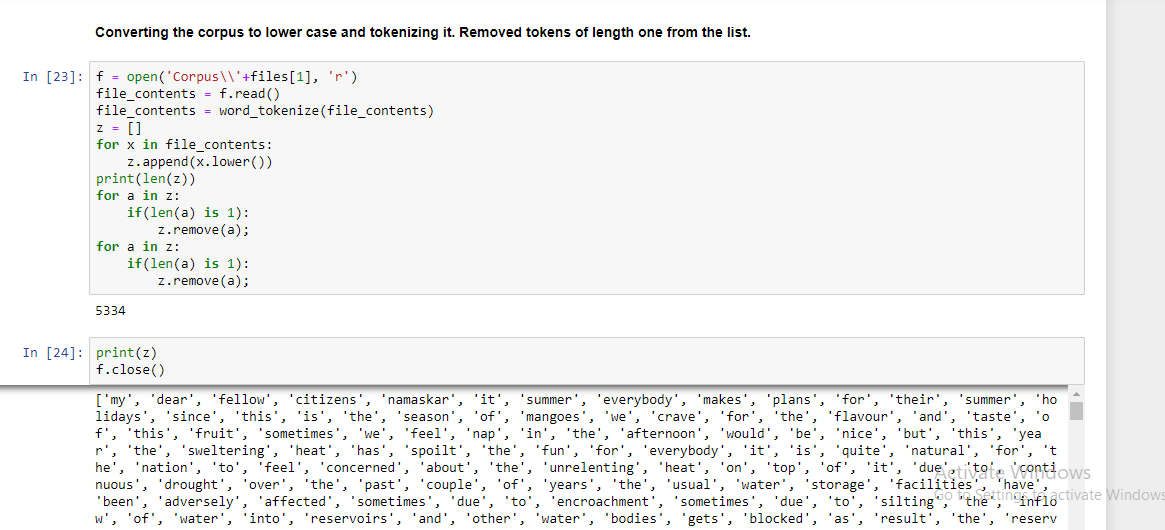
### **Vocabulary**

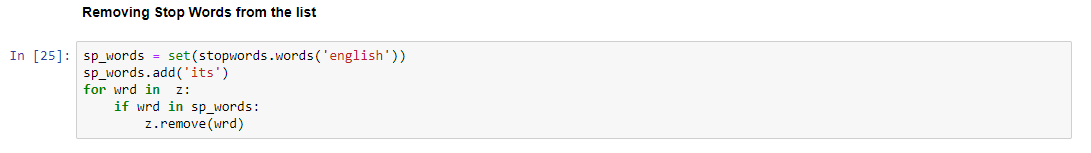
Will contain a dictionary of all the unique words in the corpus. Example:

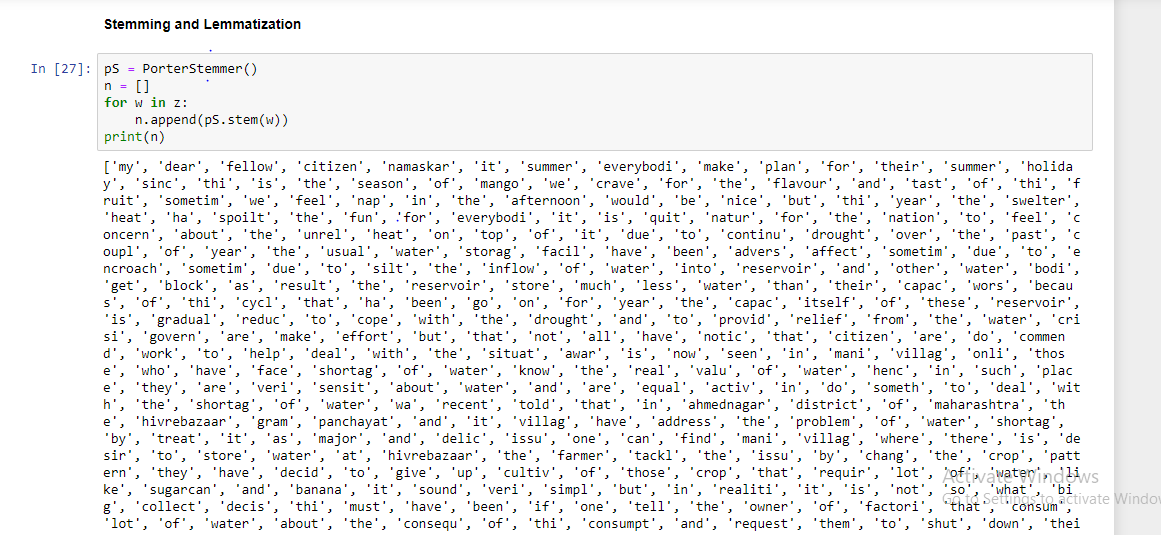
{‘i’: 1, ‘am’:2, ‘good’:3, ‘thank’:4, ‘you’ :5, ‘india’:6 , ‘is’ :7, ‘best’:8]

### **Steps:**

1. Tokenization



1. Removing Stopwords
2. Stemming and Lemmatization

****