Library used:

from nltk.tokenize import word\_tokenize

from nltk.corpus import stopwords

from nltk import stem

import string

import json

NLTK

* tokenize a text file in to words
* generate a list of stopwords for removal
* stem the words

STRING

* used to remove punctuations from the tokens

Reference: http://stackoverflow.com/questions/265960/best-way-to-strip-punctuation-from-a-string-in-python

JSON

* write dictionary variable to txt files