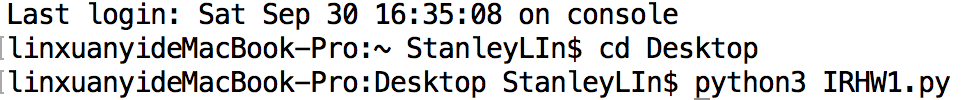
Information Retrieval and Text Mining HW1 B03705002 林軒逸

1. **Working Environment**

The program is ran under **Mac OS Sierra, version 10.12.6.** The method to run the program is open the terminal, change the directory to the file where the program is, and execute it by entering **“python3 IRHW1.py”.**

(or **“python3 IRHW1\_nltk.py”** for another program)

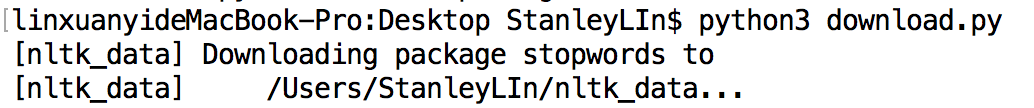
(PS: all the “3”s in the paper are for specifying python 3 version in mac, depends on personal computer and python version, 3 can be omitted.)

****For instance, if the program is saved at the Desktop, the command will be:

1. **Programming Language and Tools Used**

For this programming assignment, I use **Python 3.6.0** to finish all the programs. Also, I used a special package - “NLTK (Natural Language Toolkit)” to support me with stemming by Porter’s Algorithm.

To install the packages and data, enter **“pip3 install -U nltk”**

 To install the nltk stop words, please run ”download.py” program attached in the file. It will download the stopword data of nltk for further use. To run the download program, enter **“python3 download.py”** under the correct directory.

1. **File Description**
2. Source Code (.py):
3. IRHW1.py

The program that outputs the tokens using the stop words provided by professor. Each token is divided by “\n”. The output file is “**Output.txt**”.

1. IRHW1\_nltk.py

The program that outputs the tokens using the stop words provided by nltk. Each token is divided by “\n”. The output file is “**Output\_nltk.txt**”.

1. download.py

The program is used for downloading nltk stop words.

1. Text files:
2. paragraph.txt

The original paragraph provided by the professor.

1. stoplist.txt

The stop words provided by the professor.

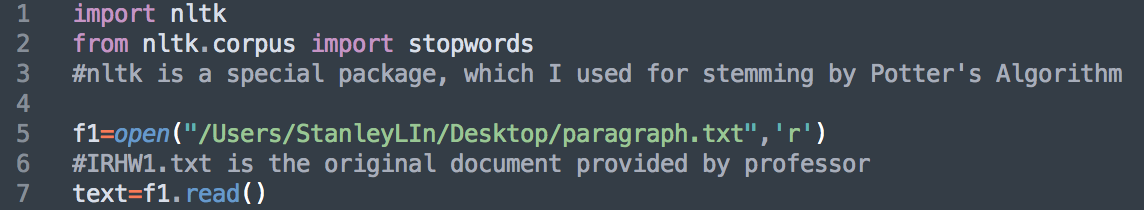
1. Output.txt

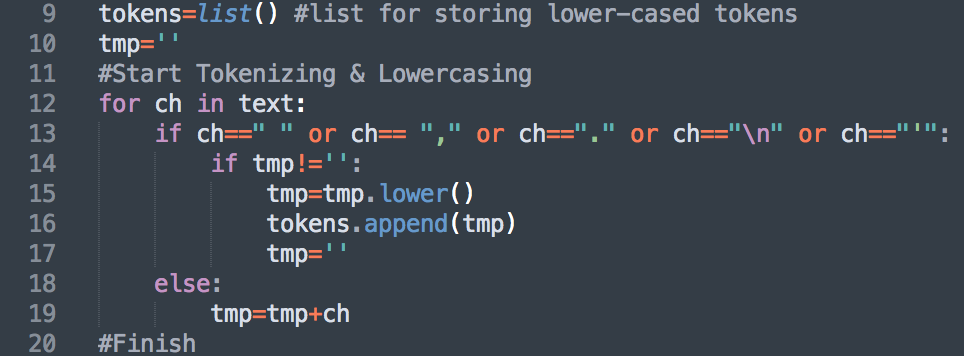
The output of “**IRHW1.py**”, each token is divided by “\n”.

1. Output\_nltk.txt

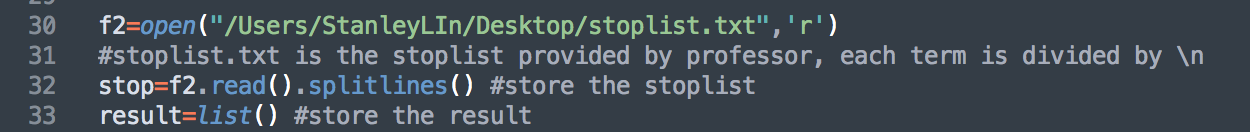
The output of “**IRHW1\_nltk.py**”, each token is divided by “\n”.

1. **Processing Methods and Code Description**

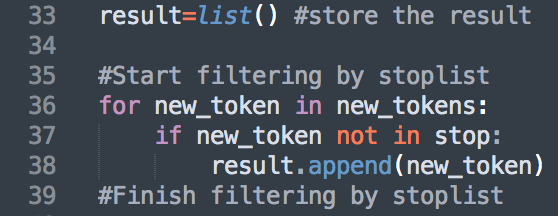
First, I store the document professor provided as “**paragraph.txt**”, and read the file in the program. Also, importing nltk for further use.

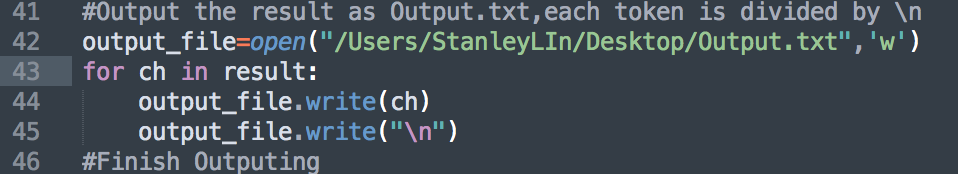
Second, I tokenize the document by symbols and blanks in the document. Also, lowercasing each token simultaneously, and store them in **“tokens”.**

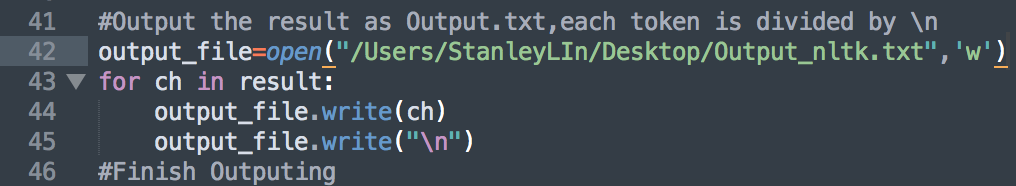
 As for the stemming part, I use nltk’s package to do it. **“porter”** stores the stemmer of nltk, and in the for loop, we use **porter.stem** to filter each token, deciding whether each of them can remain, and store new tokens in **“new\_tokens”.**

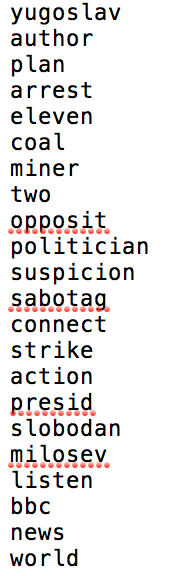
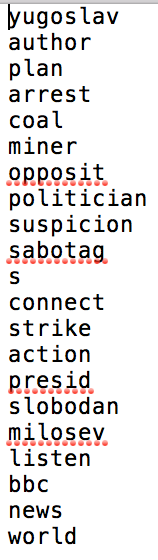
 Next, in **“IRHW1.py”,** we read the stop words provided by the professor from **“stoplist.txt”**, and store it in **“stop”** as a list object.

As for **“IRHW1\_nltk,py”,** we use the stop words provided by nltk.corpus, and store it in **“stop” i**n a set.

 Afterwards, we filter the stemmed tokens **“new\_tokens”** by the stop words **“stop”** in the for loop, checking whether each token is in the stoplist, thereby filtering them and store in the list **“result”**.

 Last, we output the result as **“Output.txt”** in **“IRHW1.py”.**

 Besides, we output **“Output\_nltk.txt”** in **“IRHW1\_nltk.py”**.

Each token is divided by “\n” (line break) in the output file. The left is the output using professor’s stop words, and the right is the output using nltk’s stop words.

1. **Notes**

In the program, I read files from my local directory. For instance, **“/Users/StanleyLIn/Desktop/”.** Thus, it is necessary to change the directory in the program to local directory in your personal computer.

There are **three** places that you need to change the directory:

1. f1, the path of “paragraph.txt”.
2. f2 in “IRHW1.py”, the path of “stoplist.txt”.
3. output\_file, the path of where you want the txt file be at.

Sorry for the inconvenience, you can assume all the **“/Users/StanleyLIn/Desktop/”** can be replace to a common file path in your personal computer. Thanks again for your patience.