I. 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:

```
linxuanyideMacBook-Pro:~ StanleyLIn$ cd Desktop
linxuanyideMacBook-Pro:Desktop StanleyLIn$ python3 IRHW1.py
```

II. 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"**

```
linxuanyideMacBook-Pro:Desktop StanleyLIn$ 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.

```
linxuanyideMacBook-Pro:Desktop StanleyLIn$ python3 download.py
[nltk_data] Downloading package stopwords to
[nltk_data]     /Users/StanleyLIn/nltk_data...
```

III. File Description

A. Source Code (.py):

1. 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**".

2. 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**".

3. download.py

   The program is used for downloading nltk stop words.

B. Text files:

1. paragraph.txt

   The original paragraph provided by the professor.

2. stoplist.txt

   The stop words provided by the professor.

3. Output.txt

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

4. Output_nltk.txt

   The output of "**IRHW1_nltk.py**", each token is divided by "\n".

IV. 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.

```
1  import nltk
2  from nltk.corpus import stopwords
3  #nltk is a special package, which I used for stemming by Potter's Algorithm
4
5  f1=open("/Users/StanleyLIn/Desktop/paragraph.txt",'r')
6  #IRHW1.txt is the original document provided by professor
7  text=f1.read()
```

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

```
9   tokens=list() #list for storing lower-cased tokens
10  tmp=''
11  #Start Tokenizing & Lowercasing
12  for ch in text:
13      if ch==" " or ch== "," or ch=="." or ch=="\n" or ch=="'":
14          if tmp!='':
15              tmp=tmp.lower()
16              tokens.append(tmp)
17              tmp=''
18      else:
19          tmp=tmp+ch
20  #Finish
```

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".**

```
22  new_tokens=list() #list for tokens after stemming by Porter's Algorithm
23  #Start Stemming
24  porter=nltk.PorterStemmer() #Store the stemmer
25  for token in tokens:
26      token=porter.stem(token)
27      new_tokens.append(token)
28  #Finsh Stemming
```

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.

```
30  f2=open("/Users/StanleyLIn/Desktop/stoplist.txt",'r')
31  #stoplist.txt is the stoplist provided by professor, each term is divided by \n
32  stop=f2.read().splitlines() #store the stoplist
33  result=list() #store the result
```

As for **“IRHW1_nltk,py”,** we use the stop words provided by nltk.corpus, and store it in **“stop” i**n a set.

```
31  stop = set(stopwords.words('english')) #the stoplist of nltk.corpus
```

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”**.

```
33  result=list() #store the result
34
35  #Start filtering by stoplist
36  for new_token in new_tokens:
37      if new_token not in stop:
38          result.append(new_token)
39  #Finish filtering by stoplist
```

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

```
41  #Output the result as Output.txt,each token is divided by \n
42  output_file=open("/Users/StanleyLIn/Desktop/Output.txt",'w')
43  for ch in result:
44      output_file.write(ch)
45      output_file.write("\n")
46  #Finish Outputing
```

Besides, we output **“Output_nltk.txt”** in **“IRHW1_nltk.py”**.

```
41  #Output the result as Output.txt,each token is divided by \n
42  output_file=open("/Users/StanleyLIn/Desktop/Output_nltk.txt",'w')
43  for ch in result:
44      output_file.write(ch)
45      output_file.write("\n")
46  #Finish Outputing
```

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.

```
yugoslav
author
plan
arrest
coal
miner
opposit
politician
suspicion
sabotag
s
connect
strike
action
presid
slobodan
milosev
listen
bbc
news
world
```

```
yugoslav
author
plan
arrest
eleven
coal
miner
two
opposit
politician
suspicion
sabotag
connect
strike
action
presid
slobodan
milosev
listen
bbc
news
world
```

V. 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.