Report

Strategy:

I have used BeautifulSoup which is a python library to parse the HTML documents. It does its job by pulling the data out of HTML and XML documents. A Hash map like dictionary is used to store the words vs frequency. The logic checks that if the data is encountered for the first time or already visited and then updates the frequency accordingly using a count variable. A file of tokens sorted alphabetically by using the sorted () function which takes keys/words an iterator is returned. A separate file of tokens sorted frequency wise is also maintained by using the items as an iterator and lambda as a function object.

Handling Punctuations:

I made sure that all the non-relevant characters are removed before calculating the frequency. I use the RegexpTokenizer method from the NLTK package to scrape all the words and numbers and then use the isalpha () function to return only words as the tokens as the focus of the assignment.

RegexpTokenizer: (r'\w+')

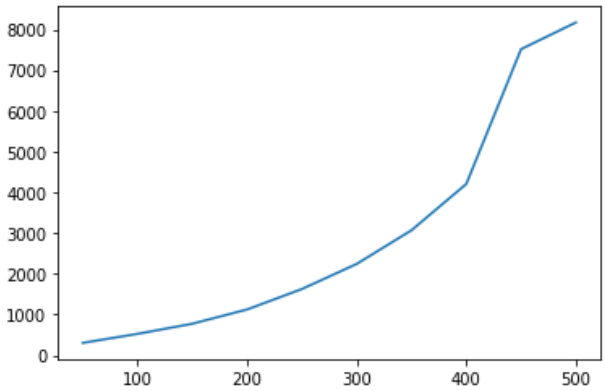
Performance Table:

Number of input documents	Time Taken (ms)
50	307
100	528
150	777
200	1127
250	1629
300	2245
350	3071
400	4206
450	7515
500	8170

```
In [2]: import matplotlib.pyplot as plt

In [3]: plt.plot([50,100,150,200,250,300,350,400,450,500],[307,528,777,1127,1629,2245,3071,4206,7515,8170])

Out[3]: [<matplotlib.lines.Line2D at 0x278e6786548>]
```



Installation Instructions:

I have executed the python script using Anaconda command prompt:

```
(base) C:\Users\Vrindavan\Downloads\IR>python assignment01.py C:\Users\Vrindavan\Downloads\IR\files C:\Users\Vrindavan\Downloads\IR\tokenized
2020-02-18 02:18:43.067715
Input directory is: C:\Users\Vrindavan\Downloads\IR\files
output directory is: C:\Users\Vrindavan\Downloads\IR\tokenized
8170
```

The command line takes two arguments: - <input_dir> and <output_dir>

Comparison:

I compared my program with Madhurya's program. Execution time wise, my program took 8.12 seconds and her program took 23 seconds. Both of used BeautifulSoup python package to extract the data from HTML pages. She used the split () function to tokenize and I used isalpha() .