Introduction

I wanted to complete a text analysis comparing 2 separate texts in Python. This will help us to understand how python can use its algorithms to compare different texts by looking at how many times certain words appear within it. I decided to look at Donte's inferno as well as Shakespeare to see if two classics have a lot of the same words used. I wanted to see if these two classics were similar in many regards.

Methods/R Code

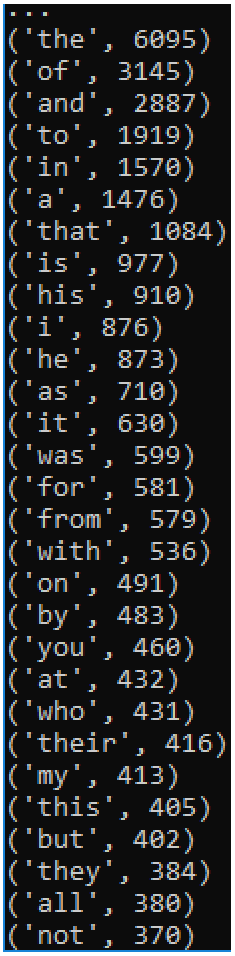
To do this, I had to upload data into Python after downloading it. Once the data was downloaded, we manipulated different operators in order to help let the system know what to do. First, I downloaded the inferno text to analyze and then I downloaded the Shakespeare text to compare. I wanted to make sure that we wrote the correct code in order to compare the word counts as well as the words without spacing.

I decided to do Donte's Inferno and Shakespeare in text format for comparison. I then loaded their txt files. The first time the files were uploaded they had stop-words included and they then had to be removed for a more accurate analysis. Once this was complete both texts could be compared and analyzed.

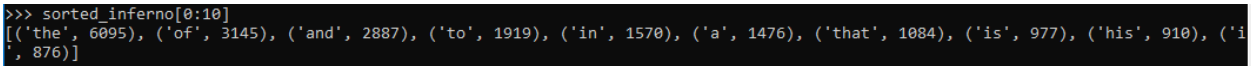
Results/Outputs

Below are the different steps of the process along with the text imported and the final outputs of the two analyses.

Below is the creation of the lists of distinct words that have been sorted of the Donte's Inferno text before it was cleaned with the removal of the stopwords.

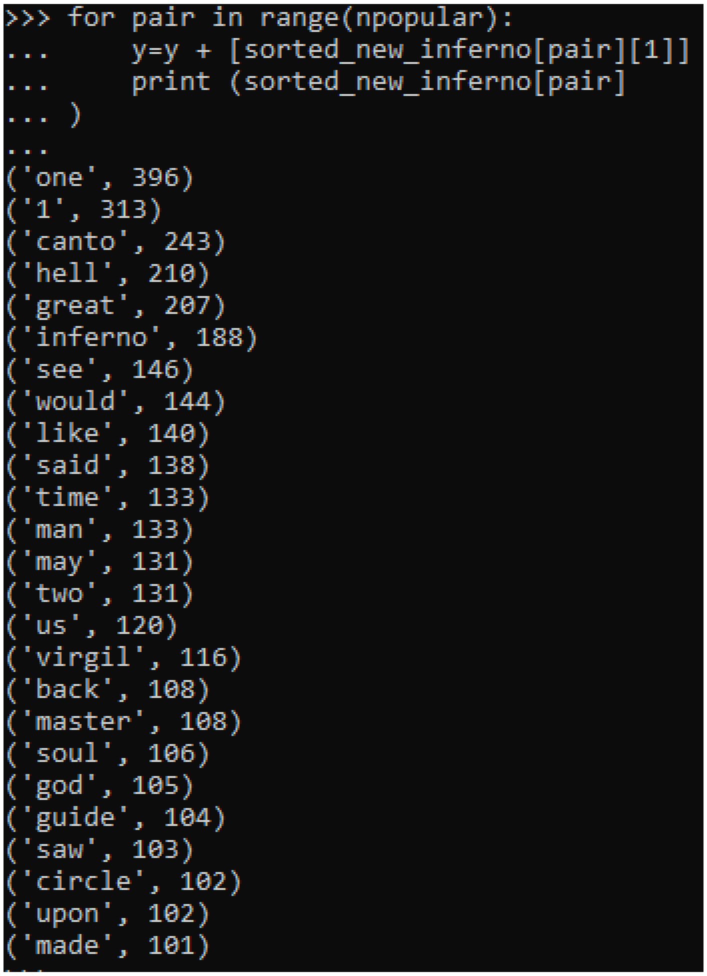


```
...
('the', 6095)
('of', 3145)
('and', 2887)
('to', 1919)
('in', 1570)
('a', 1476)
('that', 1084)
('is', 977)
('his', 910)
('i', 876)
('he', 873)
('as', 710)
('it', 630)
('was', 599)
('for', 581)
('from', 579)
('with', 536)
('on', 491)
('by', 483)
('you', 460)
('at', 432)
('who', 431)
('their', 416)
('my', 413)
('this', 405)
('but', 402)
('they', 384)
('all', 380)
('not', 370)
```



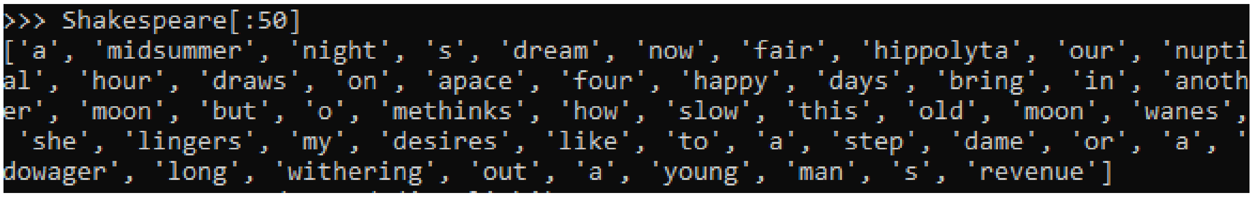
```
>>> sorted_inferno[0:10]
[('the', 6095), ('of', 3145), ('and', 2887), ('to', 1919), ('in', 1570), ('a', 1476), ('that', 1084), ('is', 977), ('his', 910), ('i
', 876)]
```

Below is the list for comparison after the stop-words have been taken out of the Donte's Inferno text for comparison.



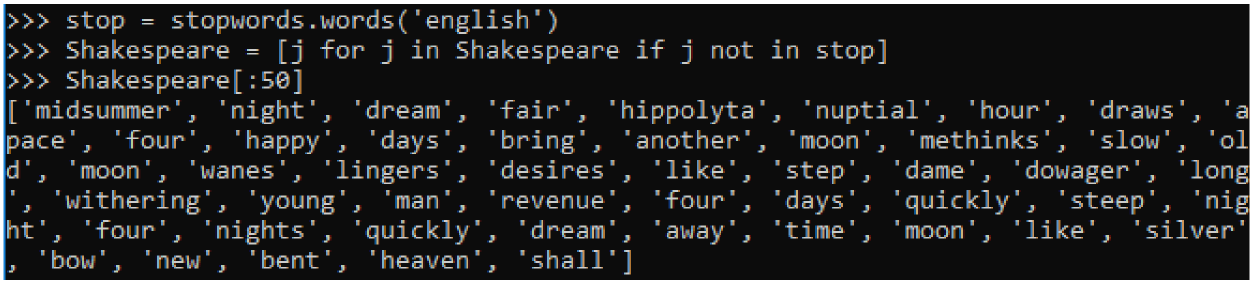
```
>>> for pair in range(npopular):
...     y=y + [sorted_new_inferno[pair][1]]
...     print (sorted_new_inferno[pair]
... )
...
('one', 396)
('1', 313)
('canto', 243)
('hell', 210)
('great', 207)
('inferno', 188)
('see', 146)
('would', 144)
('like', 140)
('said', 138)
('time', 133)
('man', 133)
('may', 131)
('two', 131)
('us', 120)
('virgil', 116)
('back', 108)
('master', 108)
('soul', 106)
('god', 105)
('guide', 104)
('saw', 103)
('circle', 102)
('upon', 102)
('made', 101)
```

Below are the listed words for comparison before the stop-words were removed in the clean up process.



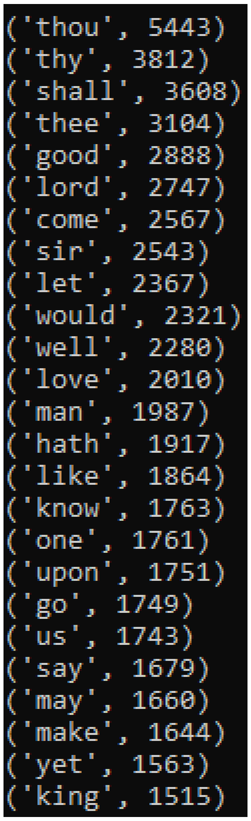
```
>>> Shakespeare[:50]
['a', 'midsummer', 'night', 's', 'dream', 'now', 'fair', 'hippolyta', 'our', 'nupti
al', 'hour', 'draws', 'on', 'apace', 'four', 'happy', 'days', 'bring', 'in', 'anoth
er', 'moon', 'but', 'o', 'methinks', 'how', 'slow', 'this', 'old', 'moon', 'wanes',
 'she', 'lingers', 'my', 'desires', 'like', 'to', 'a', 'step', 'dame', 'or', 'a', '
dowager', 'long', 'withering', 'out', 'a', 'young', 'man', 's', 'revenue']
```

The below is when the stop words were taken out of the list for clean up and comparison.



```
>>> stop = stopwords.words('english')
>>> Shakespeare = [j for j in Shakespeare if j not in stop]
>>> Shakespeare[:50]
['midsummer', 'night', 'dream', 'fair', 'hippolyta', 'nuptial', 'hour', 'draws', 'a
pace', 'four', 'happy', 'days', 'bring', 'another', 'moon', 'methinks', 'slow', 'ol
d', 'moon', 'wanes', 'lingers', 'desires', 'like', 'step', 'dame', 'dowager', 'long
', 'withering', 'young', 'man', 'revenue', 'four', 'days', 'quickly', 'steep', 'nig
ht', 'four', 'nights', 'quickly', 'dream', 'away', 'time', 'moon', 'like', 'silver'
, 'bow', 'new', 'bent', 'heaven', 'shall']
```

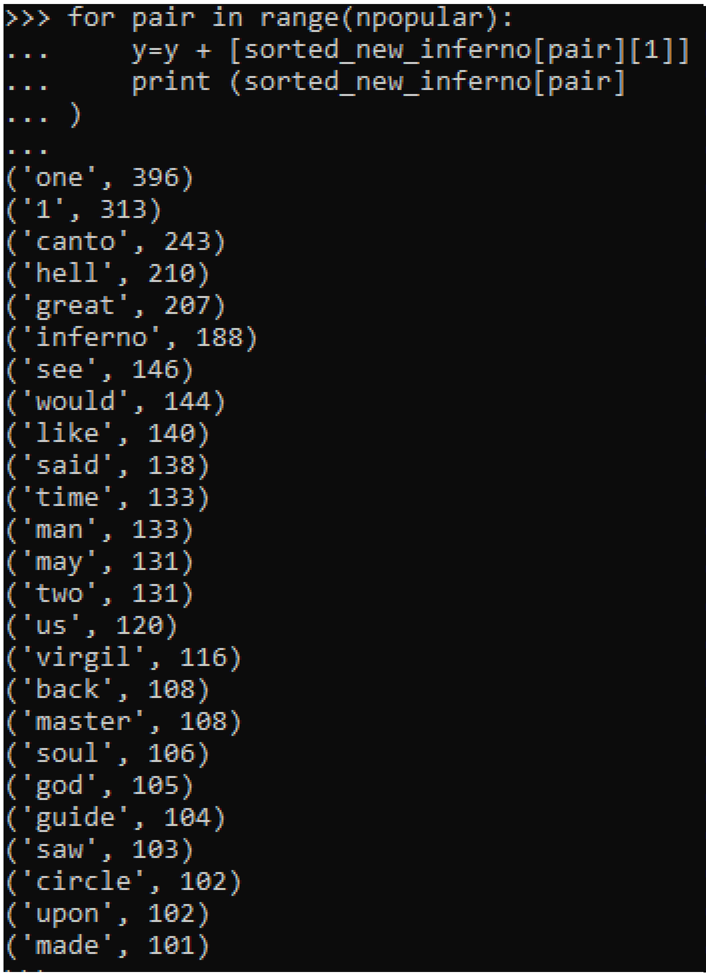
Below is the final Shakespeare text for comparison after the stop words have been taken out.



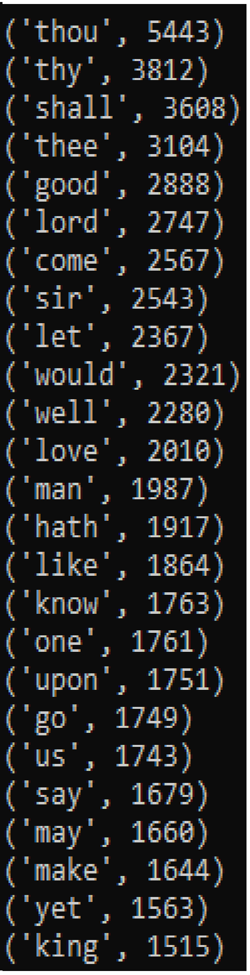
```
('thou', 5443)
('thy', 3812)
('shall', 3608)
('thee', 3104)
('good', 2888)
('lord', 2747)
('come', 2567)
('sir', 2543)
('let', 2367)
('would', 2321)
('well', 2280)
('love', 2010)
('man', 1987)
('hath', 1917)
('like', 1864)
('know', 1763)
('one', 1761)
('upon', 1751)
('go', 1749)
('us', 1743)
('say', 1679)
('may', 1660)
('make', 1644)
('yet', 1563)
('king', 1515)
```

Analysis of Results

For the comparison of the two classics:



```
>>> for pair in range(npopular):
...     y=y + [sorted_new_inferno[pair][1]]
...     print (sorted_new_inferno[pair]
... )
...
('one', 396)
('1', 313)
('canto', 243)
('hell', 210)
('great', 207)
('inferno', 188)
('see', 146)
('would', 144)
('like', 140)
('said', 138)
('time', 133)
('man', 133)
('may', 131)
('two', 131)
('us', 120)
('virgil', 116)
('back', 108)
('master', 108)
('soul', 106)
('god', 105)
('guide', 104)
('saw', 103)
('circle', 102)
('upon', 102)
('made', 101)
```



```
('thou', 5443)
('thy', 3812)
('shall', 3608)
('thee', 3104)
('good', 2888)
('lord', 2747)
('come', 2567)
('sir', 2543)
('let', 2367)
('would', 2321)
('well', 2280)
('love', 2010)
('man', 1987)
('hath', 1917)
('like', 1864)
('know', 1763)
('one', 1761)
('upon', 1751)
('go', 1749)
('us', 1743)
('say', 1679)
('may', 1660)
('make', 1644)
('yet', 1563)
('king', 1515)
```

We can see that on the left there is the Donte's Inferno text compared with the Shakespeare text on the right. We can see that some of the differences are in the language used. The Shakespeare text uses a lot older English as well as speaks more about love. Shakespeare also has God in the text a lot more than in the Inferno text. The Inferno text seems to have more of a modern English language to it, but it is also more about man and hell. There were also less words in the Inferno text that are in the list with each being in the hundreds compared to the thousands in the Shakespeare text. We can see over all that in the Shakespeare classic there is a different time period being identified through the language being used even though both books are classics. That may have something to do with the regions that they were written in as well with Shakespeare being written in the old England, and Donte's Inferno being written in Italy.

Conclusion

Here I just wanted to show the process of Python and how it can be very beneficial in the mining of large text documents in order to separate out certain words, phrases, and comparing texts that are output. I think that this will come in handy for future use, because There will be times in which large data will need to be compared and maybe even checked. This can be done across platform with many different texts and there are many possibilities. This algorithm could allow us to rapidly find comparisons and differences within large seas of information. It was interesting to see the algorithm at work on such an easy to operate interface that involved more precise code writing.

Resources:

1. Stack Overflow for queries. https://stackoverflow.com/questions/13000455/error-in-python-ioerror-errno-2-no-such-file-or-directory-data-csv
2. Python Text Analysis. https://worldclass.regis.edu/content/enforced/233321-CG_MSDS650-XIN_X71_19S8W2/Course%20Resources/W6-Assignment-PythonTextAnalysis.pdf?_&d2lSessionVal=jFfvQgpTzkQqiuAAS4zZ61SF8&ou=233321