**IST 664 – NLP Homework 1**

**5/3/2021**

**David Doman**

**Introduction & Texts**

Poetry has been an important and popular way of writing for many centuries. Styles and topics of these writings have changed over time and have transformed the way people think and feel about the world around them. Two very well-known writings that have shaped and provided perspective to life and world around us are “Paradise Lost” by John Milton and “Leaves of Grass” by Walt Whitman. These writings were written nearly 200 years apart so inevitably one would expect the idea and style between the two would be quite different in many ways. While Milton focuses on the religious and biblical perspective concerning Adam & Eve and the Garden of Eden in “Paradise Lost”, Whitman focuses more on the philosophy of life and humanity, as well as nature and the individual human’s role in it. The styles of these two writings will be explored throughout the rest of this analysis. The following writings are contained within the Gutenberg corpus. The sources of the texts are shown below.

*Paradise Lost:* <https://www.gutenberg.org/files/20/20-0.txt>

*Leaves of Grass:* <https://www.gutenberg.org/files/1322/1322-0.txt>

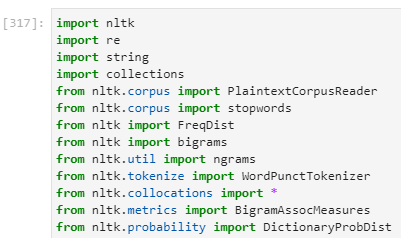
**Questions**

While these two writings are similar in the way that they both deal with life and how we interpret the world around us, each takes a different perspective to explain this. “Paradise Lost” relies on a higher power or a God to explain events in a biblical fashion, while “Leaves of Grass” points to nature and the responsibility the individual human has for why certain events happen. These altering points of view were the basis for the following questions to be answered in the analysis:

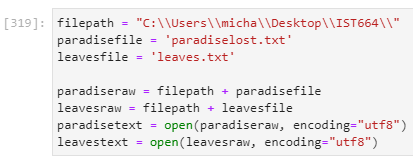
* How do the types of words differ and how are they similar between the two writings?
* How do the styles of writing differ and how are they similar between the two writings?
* Do each of the authors have a specific way of writing that is unique to them?

**Data & Pre-processing**

The very first step before anything else was to load in the necessary packages into Jupyter that will be used for the analysis. Below are the packages that were loaded:

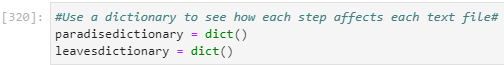


After loading the packages, the two text files were brought into the Jupyter using the following code shown below:

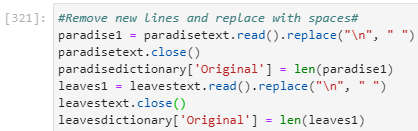


Initially, an error occurred when attempting to load in the files. This was due to the encoding being incorrect, so this was changed to “utf8” after searching for resolutions through google.

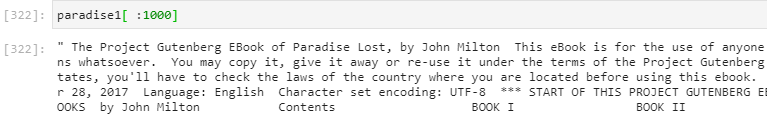
Although not a necessary step, dictionaries for each of the two texts were created to be able to see how each step of the data preprocessing affected the texts.

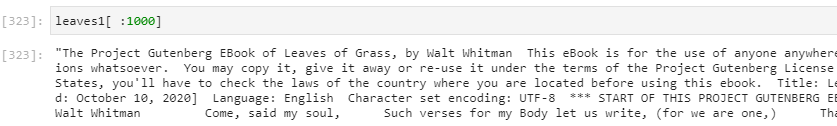


Now that the texts were loaded and the dictionaries were created, the data preprocessing began with removing new lines and replacing them with spaces. Performing this action makes the tokenization process work much better and cleaner. This first action on the texts was stored into the dictionaries. More specifically, the length of the texts was stored.



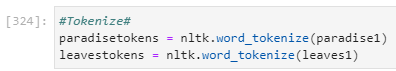
To get an initial view of what the current state of the texts looked after removing new lines with spaces, the following code was run shown below:



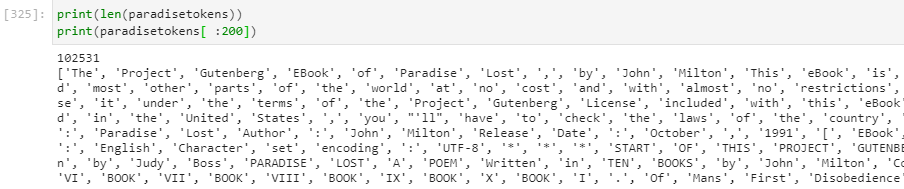


As shown above in both “Paradise Lost” and “Leaves of Grass”, the first 1,000 or so words in the texts are not actually a part of the text but rather disclaimers and verbiage unrelated to the actual writing. This unnecessary and useless text will be handled in future steps.

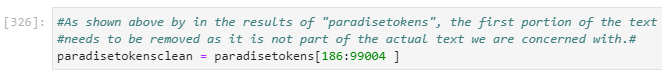
Tokenization was then performed on each of texts using the simple word\_tokenize code shown below:



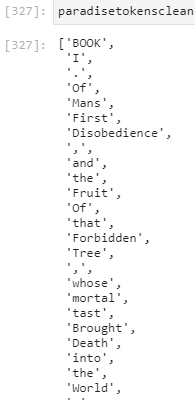
After tokenization, the length of the tokenized text and the first 200 tokens were viewed. The results are shown below:

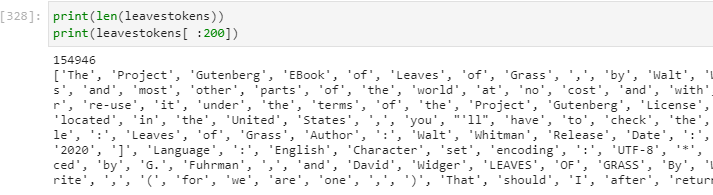


As stated previously, the first portion of the text contained wording that was unrelated to the actual text and needed to be removed to clean up the text. It was also found that there was unnecessary and unrelated wording at the end of the text file, so this was also removed in the same fashion. This process took some trial and error to find exactly where the meaningful text began and when it ended. Below shows the tokens that were kept to be used in the analysis.

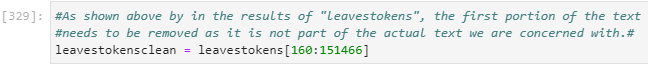


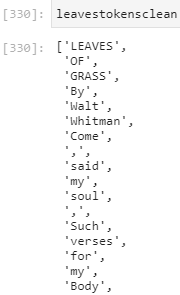
Below is a small example of the beginning of the tokens. One can see that the unnecessary tokens shown previously are now removed and the first token is the start of Book I.



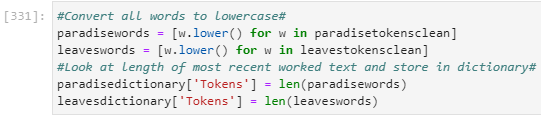


This same process of trimming off the unnecessary wording at the beginning and end of the text file was performed on “Leaves of Grass” as well.

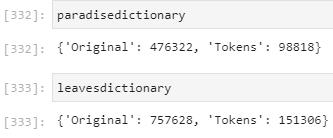




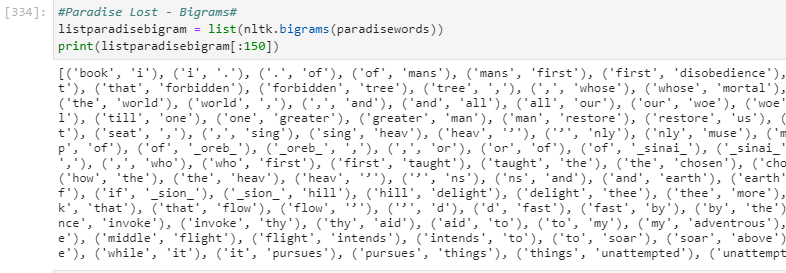
After trimming the texts, the next step was to convert all cases within both texts to lowercase to provide uniformity throughout each of the texts. These most up to date trimmed and converted to lowercase texts were then stored into the dictionaries.



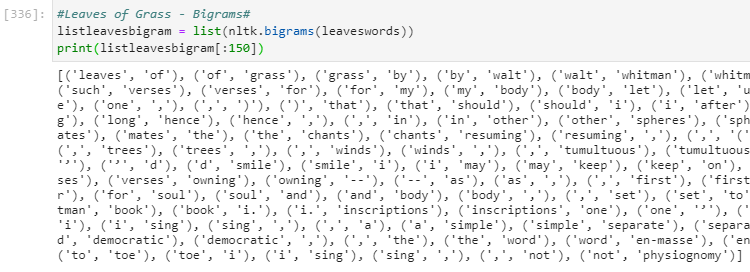
The dictionaries shown below contain the original document’s length and the number of tokens after the trimming was performed.

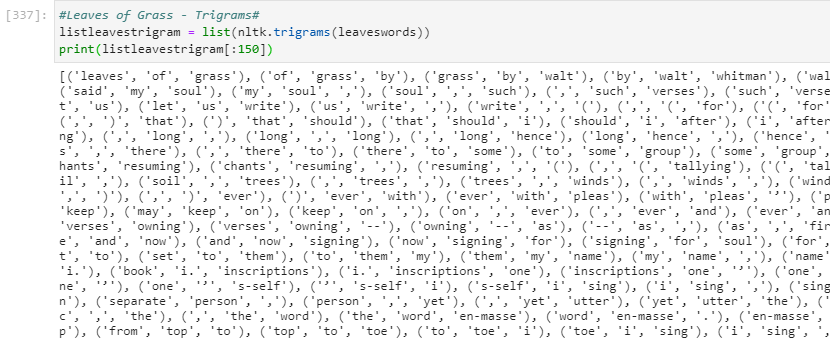


An initial look at the bigrams and trigrams were then viewed to see if further data preprocessing needed to be done.

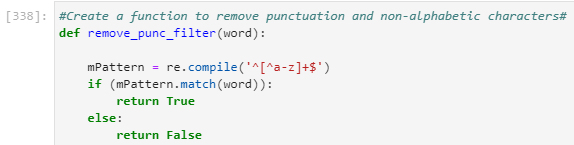


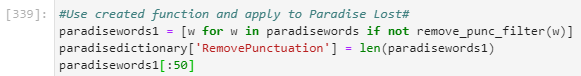


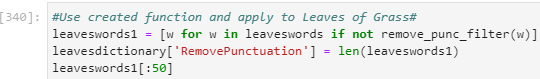




As shown in the bigrams and trigrams above, there are still commas, periods, etc. still included in the texts. This is not beneficial to the analysis, so punctuation and non-alphabetic characters need to be removed. A function was created to do this and the code for this function is shown below:



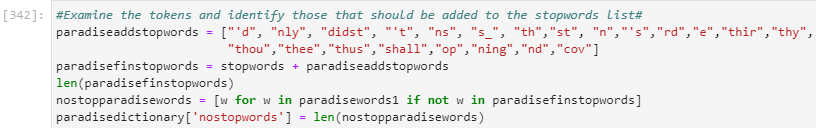


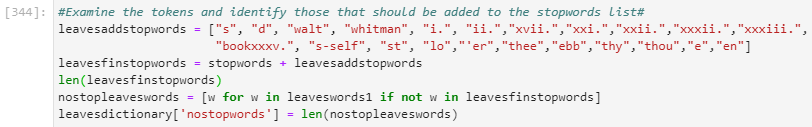


Now that punctuation and non-alphabetic characters were removed, many words that provide no insight into the writings, also known as stop words, were removed next. The nltk corpus stopwords were first called upon as this list can easily be obtained with the code shown below:



Although this original list of stopwords contains many of the frequent and common words typically removed from various texts, there were many more that were found by reviewing the bigram and trigram results, so these were added to the original list of stopwords to create a full list.

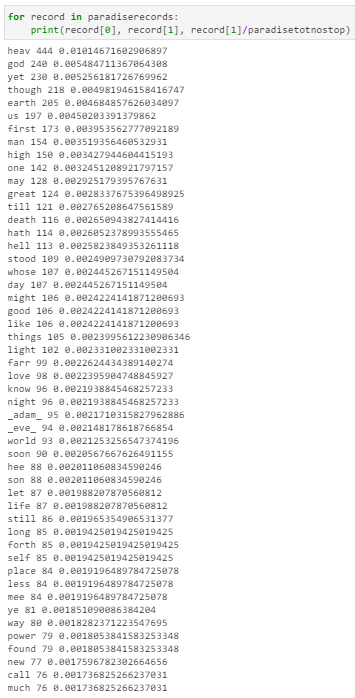
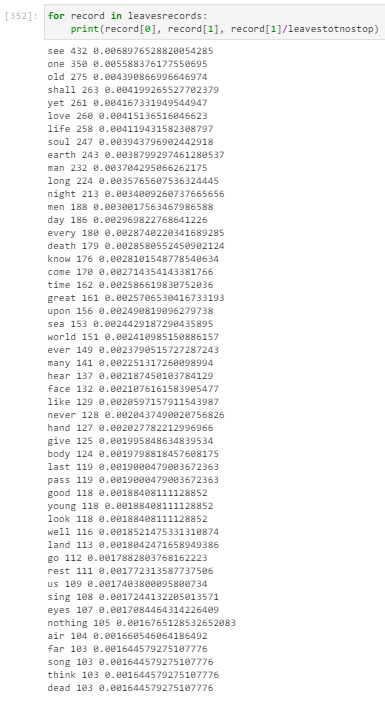




As shown above, many single letter tokens, such as “s”, “d”, “ ‘t” and many other tokens that make no sense were appended to the original list. “Leaves of Grass” also contained chapter names at the beginning of each chapter, so these were included in the stopword list. This was the last step in the data preprocessing, so the analysis of the writings was ready to be performed.

**Analysis**

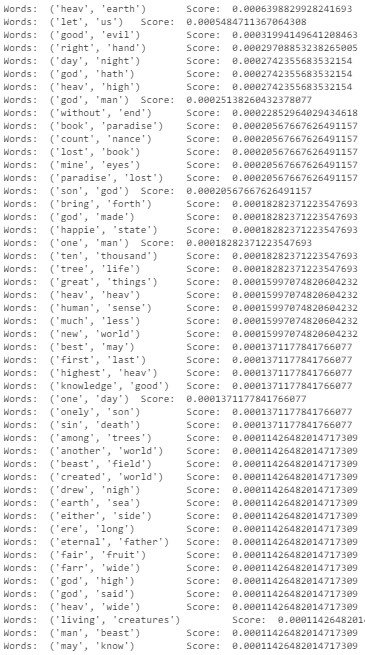
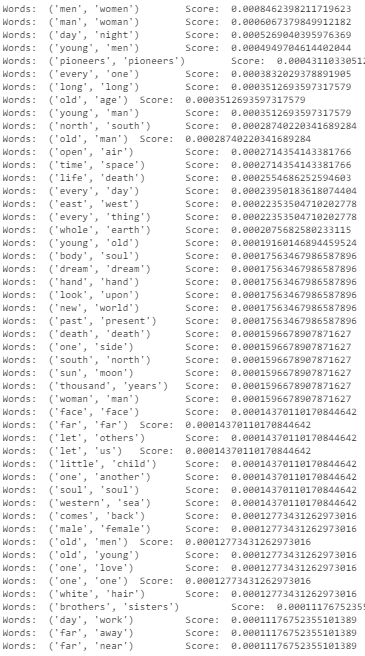
Now that the writings for “Paradise Lost” and “Leaves of Grass” have been cleaned up, tokenized, and the stopwords have been removed, the analysis of the texts can begin. The first part of the analysis involves looking at the frequency of words and to find the 50 most frequent words used in each of the writings. The FreqDist and most\_common() commands were used to perform this.

The normalized 50 most frequent words used in each of the writings is shown above. As expected, in “Paradise Lost”, which is biblically based, the two most frequent words used are heav (heaven) and god. Also included in the results for Paradise are words like hell, adam, eve, light and power which are words that would be expected in biblical writings. As for “Leaves of Grass”, some of the most frequent words used include old, love, life, soul, earth, man, death, world, and body. These types of words also make sense in that this type of writing is focused more on the human aspect and nature. There were some similarities between the two writings. Words like earth, night, world, life, and one were words included both the top 50 most frequently words for each writing.

Next the Bigrams and Trigrams were viewed to determine if anymore meaning could be pulled from the writings. Below are the results from the Bigram Analysis. A frequency filter of 4 was used for “Paradise Lost” and a frequency filter of 3 was used for “Leaves of Grass” as these were the filters that resulted in at least 50 results for each of the writings. Trial and error was a necessary part in this portion of the analysis to find the optimal filter.

*Paradise Lost Leaves of Grass*

The Bigram results showed typical biblical sayings in “Paradise Lost”. For example, (heav, earth), (god, made), (tree, life), (new, world), (created, world), (eternal, father), (sin, death), (right, hand). These are all typical word combinations one would find in biblical writings. As for “Leaves of Grass”, we see (young, man), (day, work), (white, hair), (western, sea), (sun, moon), (south, north), (life, death), (east, west). These word combinations relate more to nature and are more closely related to the types of words that are more similar to current day writings. Also included in Leaves like “pioneers” and “dreams”, which also relate more to current day words. A similarity between the two writings included (day, night) which makes sense since these are terms that could typically be used in any type of writing from any time era.

After viewing the results from the Bigrams, the Trigrams were viewed to see if any addional insights could be determined. These are shown below:

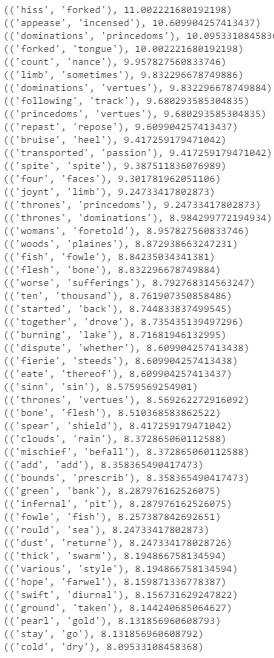
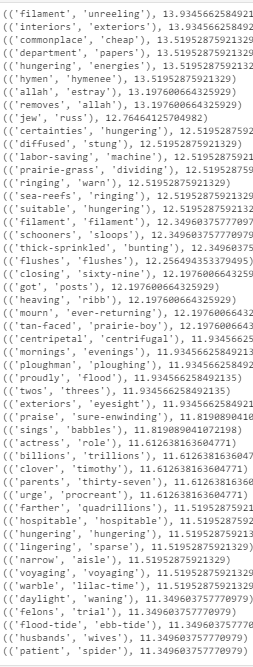
*Paradise Lost Leaves of Grass*

The Trigrams resulted in more precise and more descriptive words. In “Paradise Lost”, trigrams like (dominations, princedoms, vertues), (beast, bird, insect), (bruise, serpents, head), (confess, humbly, faults), (god, creator, wise), (herbs, fruits, flours) appeared and provided words not found in the bigrams. In “Leaves of Grass”, (beat, drums, blow), (fly, like, bird), (lands, western, shore), (splendid, silent, sun), (aboard, ship, helm). Again, as was the case with Paradise, deeper and more descriptive words appeard in the trigrams as opposed to the bigrams.

The PMI’s were then performed on both writings to see how the results compared to those of the Bigrams and Trigrams previously run. A frequency filter of 3 was applied to all PMI Bigrams and Trigrams with a window size of 5.

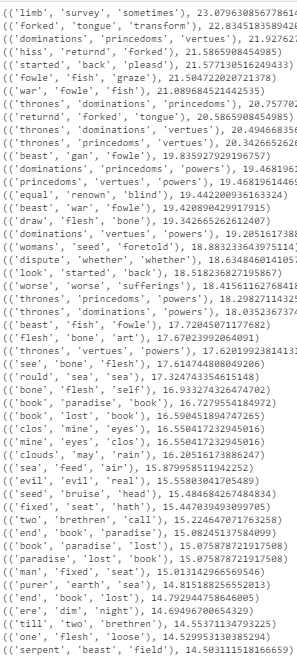
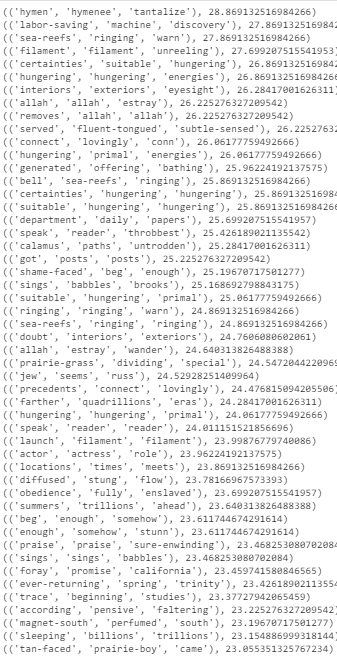
*Paradise Lost Leaves of Grass*

What seems apparent in the PMI results for both of the writings shown above is that the more descriptive words that we only saw in the trigrams in the previous analysis are showing up in the Bigram using PMI. Again, we do see a lot of those typical biblical terms for Paradise but it does not include words like god, heaven, good, evil, etc. Compared to the Bigram used in the previous section, the PMI Bigram results are definitely harder to interpret Paradise as a biblical writing. One could identify biblical terms in the PMI results but it seems to be not as clear as it was for the initial Bigram analysis. The same was the case for “Leaves of Grass”. The results for the Leaves PMI Bigram do provide more actual locations, such as india, ontario, asia, europe, africa. There are also additional terms about nature that would tell us this is a writing about nature.

The PMI Trigrams were also produced, and the results are shown below:

*Paradise Lost Leaves of Grass*

The PMI Trigrams seem to be more similar to the original trigrams in that they provide much more desciptive words. There are some similarities between the PMI and original trigrams but it seems the PMI results take a step further.

**Results & Conclusion**

In conclusion, the top 50 word frequencies produced results that I was not too surprised about. In “Paradise Lost”, biblical words like god, heaven, earth, hell, adam and eve were present so this made sense. Similarly, In “Leaves of Grass”, words like life, soul, man, death and sea were present in the top 50, which makes a lot of sense being a writing about humans and nature. The original Bigrams and Trigrams produced results that I believe were not as descriptive as the PMI Bigrams Trigrams. Meaning of the writings was able to be interpreted from the original Bigrams and Trigrams easier than it was from the PMI’s but the PMI Bigrams and Trigrams were more precise in their results and provided a step further in terms of descriptive words. I would use a combination of both methods if I were to examine these writings further.