Lauren Bassett
Jme9rt
DS 5001 Final Project

Corpse Corpus

A Text Analysis on Classic Tales of Horror

The Horror genre has scared and delighted the public for decades. Classic horror novels set the groundwork for the bone-chilling thriller movies we see today. In this assignment, I will be analyzing three classic scary tales; *The Phantom of the Opera,* by Gaston Leroux; *The Picture of Dorian Gray,* by Oscar Wilde; and *Dracula,* by Bram Stoker. In the preface for Dorian Gray, Oscar Wilde writes, "*Those who find beautiful meanings in beautiful things are the cultivated."* By analyzing the text of these novels, I aim to do just that.

1. **CONVERT**

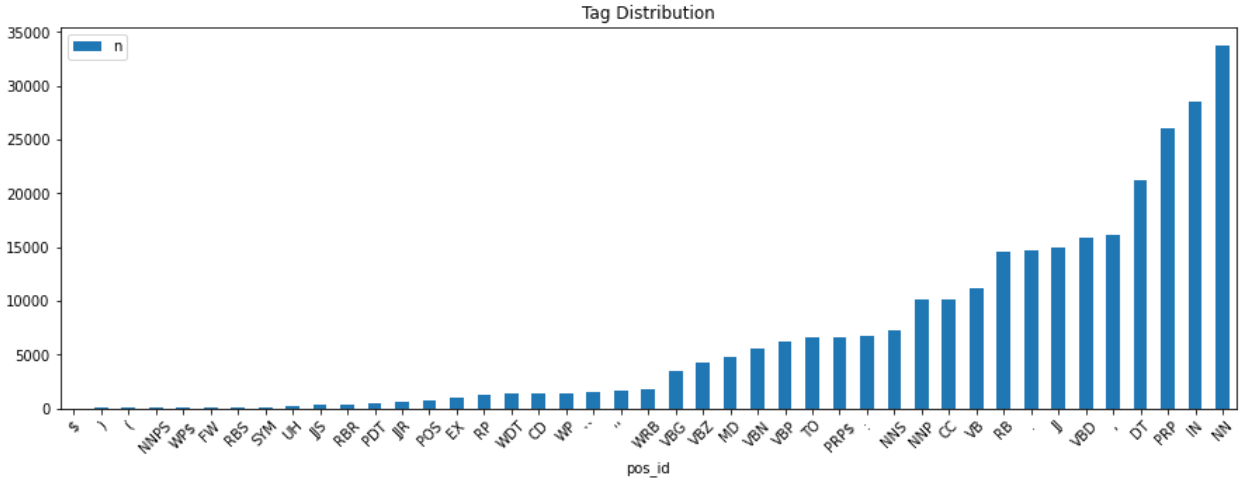
The convert phase was simple enough. To populate the DOC table, I used regex expressions to identify each chapter, and iterated line by line to group each book into the Standard Text Analytic Data Model (F2) and to the Machine Learning Corpus Format (F1). To generate the LIBRARY Table, I extracted the author, book title, and file directory. To create the basic TOKEN table, I used a function to turn each paragraph into a sentence, and then a sentence to a token. I also used NLTK to pick the tokenizer for each word, which gave the TOKEN table the appropriate linguistic features. I decided to use the whitespace Tokenizer to ensure that all contractions were properly tokenized. I ensured that the index was set to OHCO. Once I ensured that the tables were properly loaded, I moved onto the next stage.

2. **ANNOTATE**

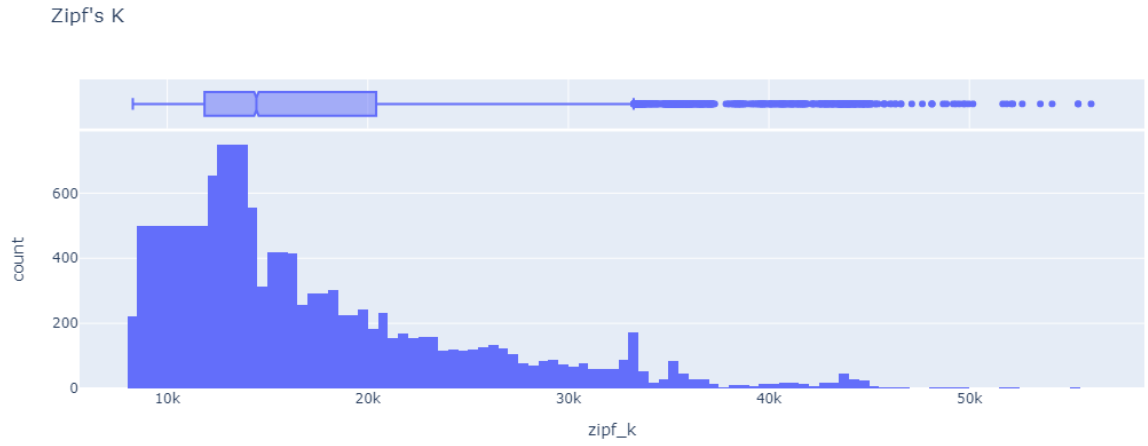
The annotate phase pulled in the TOKEN table created in the convert stage, and added the 'term string' row. To create the VOCAB table, I first used a count of each of the words in the TOKEN table, Then, I created a boolean value to identify the stopwords in the corpus. Using NLTK, I calculated the Stem Porter, Snowball Stemmer, and Lancaster Stemmer, and got the POS Max for each row.

The 10 most frequent non-stopwords are: one, said, would, could, know, us, must , like, time, and see.

The distribution of POS counts in the corpus is shown below. The top 3 parts-of-speech in the corpus are nouns, prepositions, and determiners.

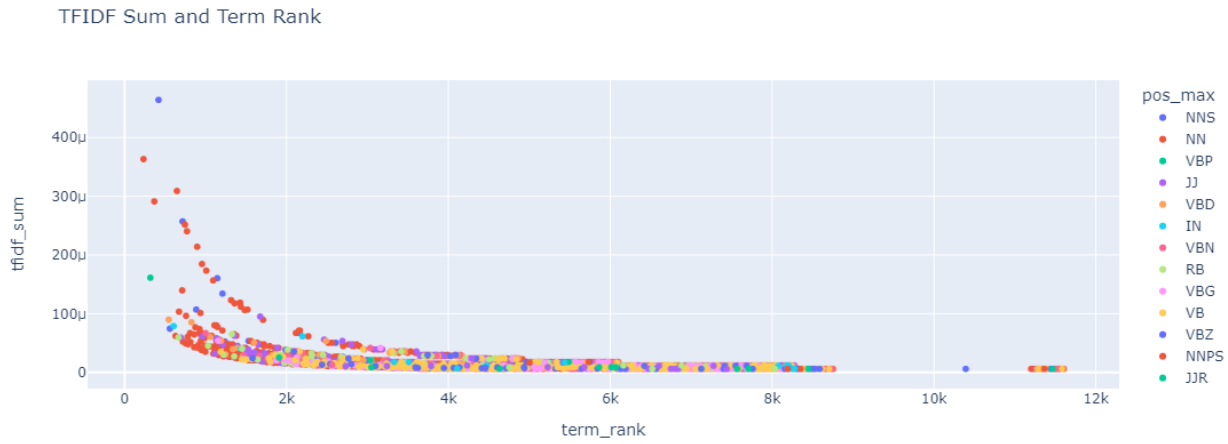


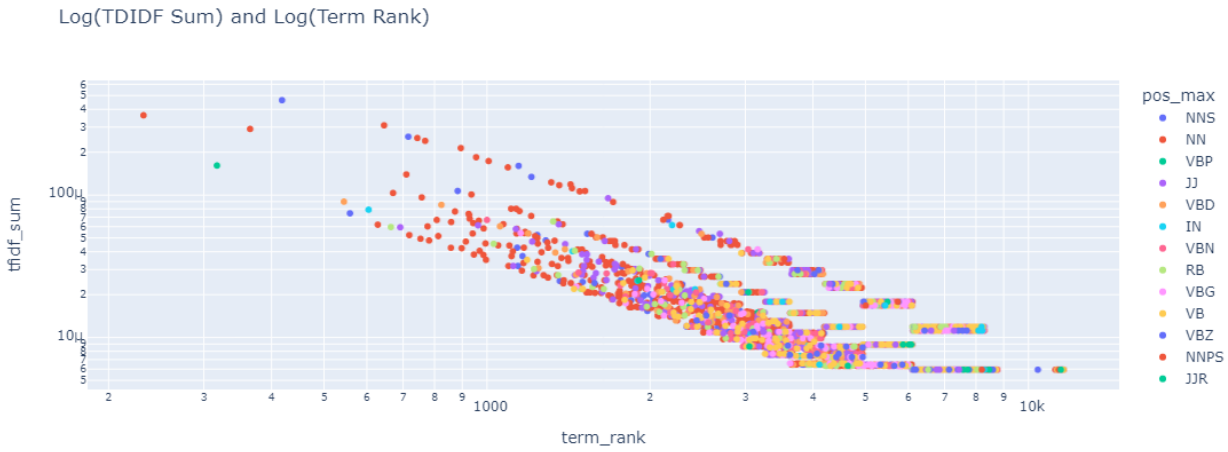
I also calculated Zipfs' Law for Zipf K, K2, and K3, and added that to the VOCAB table. Zipf's Law tells us that the meaning of a word is inversely proportional to it's ranking, and the output of our analysis shows that our corpus is following this law. The common, frequent words are associated with low meaning, while the infrequent ones are associated with high meaning.



3. Create

The next step is to create the TFIDF Vector and add the corresponding values to the TOKEN and VOCAB tables. To do this, I re-used the function we created earlier in the semester. I found that the TFIDF Sums were negatively exponentially distributed inversely with the term rank.





4. Extend

The most interesting part of this project comes in the analysis of the text. I used Principal Component Analysis, Topic Modeling, Word Embeddings, and Sentiment Analysis to explore the corpus.

Principal Components

I created 2 PCA models, using two different methods. The first I calculated manually, and came up with the following loadings for PC0, PC1, and PC2. I found it most interesting that “portrait” was sorted both positive and negative, and that ghost was positive while wolf and vampire were “negative”.

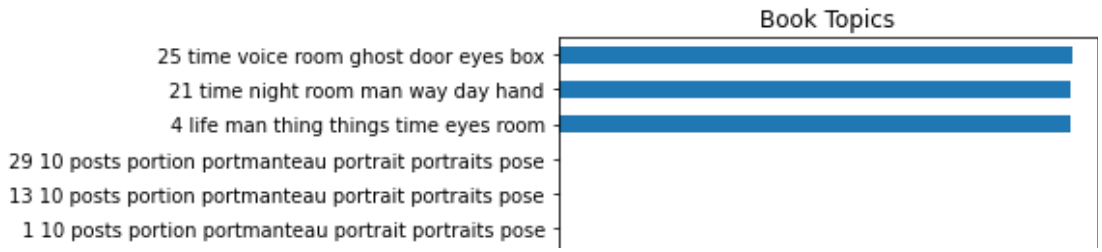
```
PC0+ managers ghost francs daroga viscount ballet commissary foyer cellars theater
PC0- portrait whilst painter sins cannot club castle studio shallow duchess
PC1+ portrait painter sins club studio shallow duchess mode boyhood coloured
PC1- whilst castle wolves comfort sunrise harbour wolf vampire mate seemingly
PC2+ consternation wishing honest duly messages choke brute cook gs condescend
PC2- specks dwelling jewels astonish scratchin steeply bewilder organized
anonymous architect
```

I also used Scikit-learn's PCA Tool. I was interested that the results were very simil;ar, especially for PC0. PC0 completely identical. PC1 had the positive and negative switched (I had to double check to make sure my code wasn't switched!), but PC2 was entirely different.

```
PC0+ managers ghost francs daroga viscount ballet commissary foyer cellars theater
PC0- portrait whilst painter sins cannot club castle studio shallow duchess
PC1+ whilst castle wolves comfort sunrise harbour wolf vampire seemingly mate
PC1- portrait painter sins club studio shallow duchess mode boyhood coloured
PC2+ managers francs daroga viscount ballet commissary foyer cellars theater
mademoiselle
PC2 - ghost whilst portrait painter cannot sins castle club wolves studio
```

I saved the first LOADINGS File as "LOADINGS.csv"

Topic Modeling
The Book Topic Modeling weights' were very limited. However, I find it interesting that I can't predict which of the three topics belong to which book. Each one discusses 'time' 'room' and a body part; either 'hands' or 'eyes'

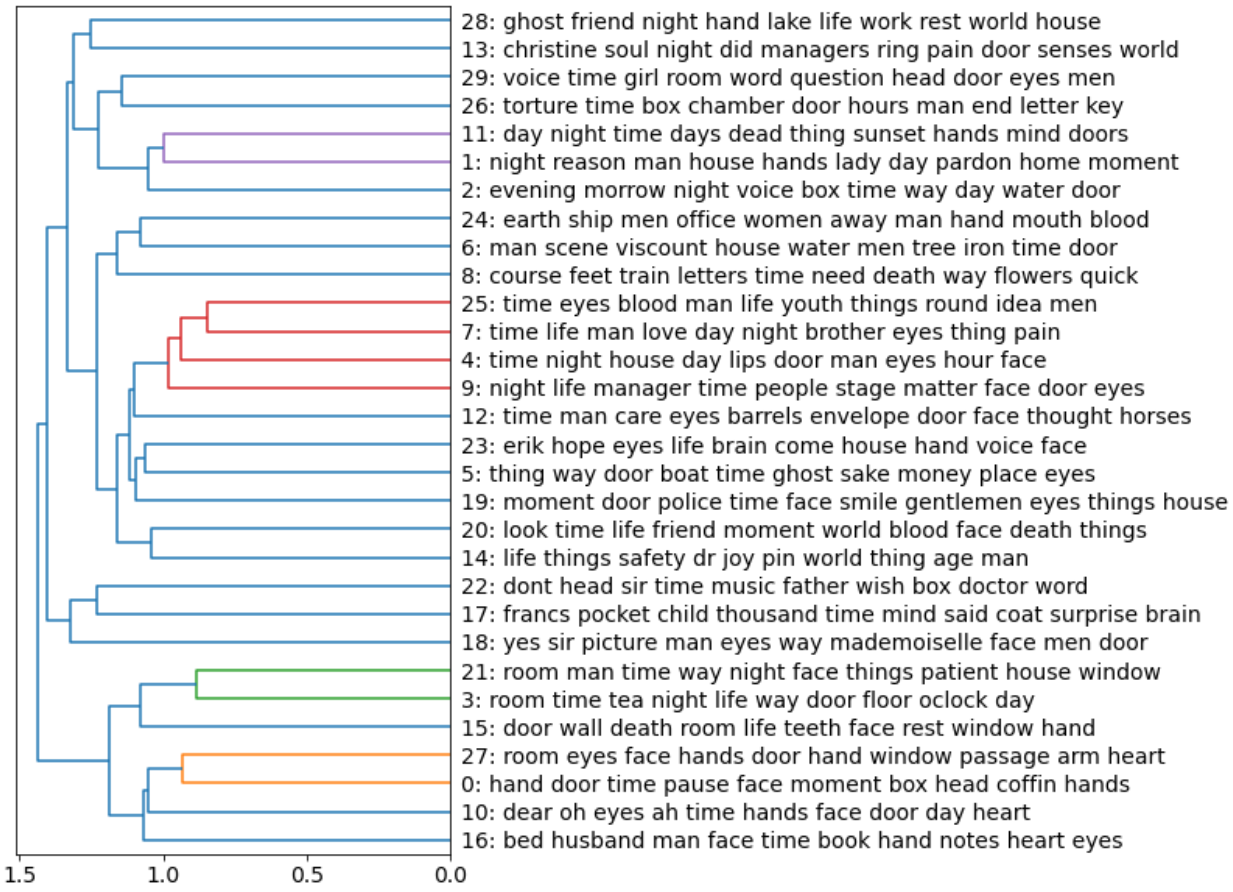


The paragraph topics are a bit more varied. The plot tree shows how the paragraph topics are connected together.

I want to point out a few of the topics:
15. door wall death room life teeth face rest window hand - These words draw out feelings of the 'monster in the room'.

25, 7, 4, 9 (Red in the Tree) - These topics are about time, humans, and parts of a building.

At an author/book level, these terms have similar representations. These books, while different, still have some of the same broad topics.



Word Embeddings
Dorian Grey

Dorian Grey has three main sections in the word embedding. The top topics are about feelings and personas; hideous, passion, death, beautiful. The left wing is about time and perception; face, passed, through, dorian.

The right wing is all negative terms; don't, never, quit, not. I find it interesting that the algorithm missed 'Dorian' as a proper noun, and placed it in the same group as the words about transformation, considering the one he goes through over the course of the novel.

The other comparison I find interesting are the terms about the painting grouped with 'years', and 'death', as the painting documents that for Dorian.

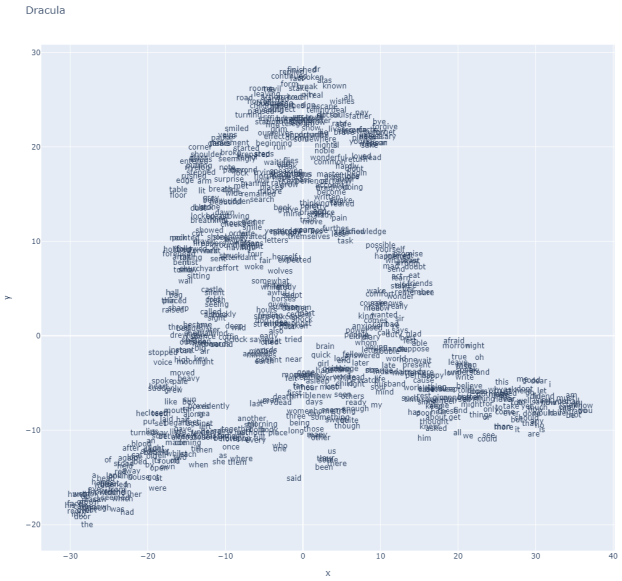
Phantom of the Opera

The Phantom of the opera has two distinct directions, one that goes up and down, and one that crosses from left to right. The vertical line that does not cross the horizontal line contains words like ‘he’, ‘you’, ‘from’ , ‘to’, and other singular experiences. The right side of the horizontal line has words like ‘staircase’, ‘building’, and ‘theatre’.
The really interesting group sits where the lines intersect. These words are ‘give’, touch’, ‘return’, around’.

Essentially, the vertical line refers to people, the horizontal line refers to places, things, and ideas,. The lines cross where people, ideas, and places intersect.

. **Dracula**

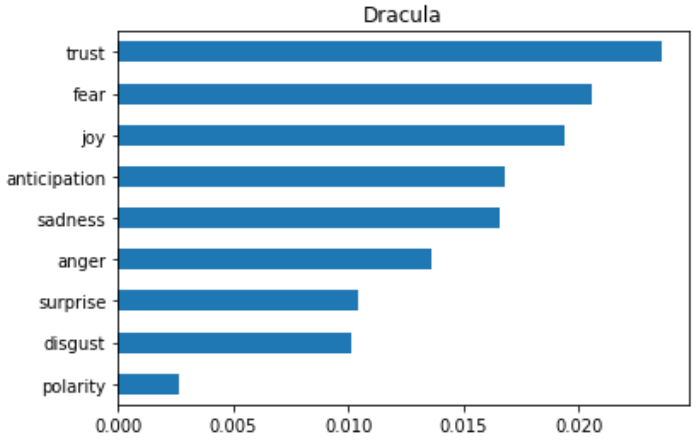
Dracula is almost the inverse of Dorian Gray, a Triangle with almost nothing in the center.
The top of the triangle contains relating to speech and communication: finished, replied, speaking, telegram, question, laugh. The bottom-left point has lots of words about bodies and the actions they take: hand, took, eyes, look, into. The bottom right point has words about emotions and thoughts; believe, better, dear, know.
Traversing the triangle from point to point transitions between the main ideas. The tiny cluster in the center has words like strength, danger, shock, cried, and answer, which I think can be connected to any of the main 3 areas.

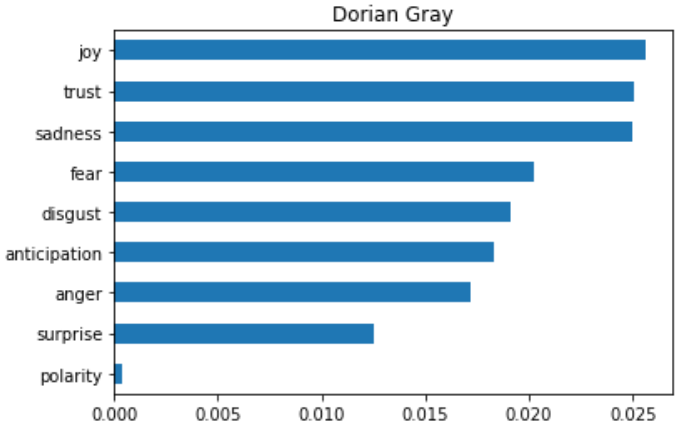


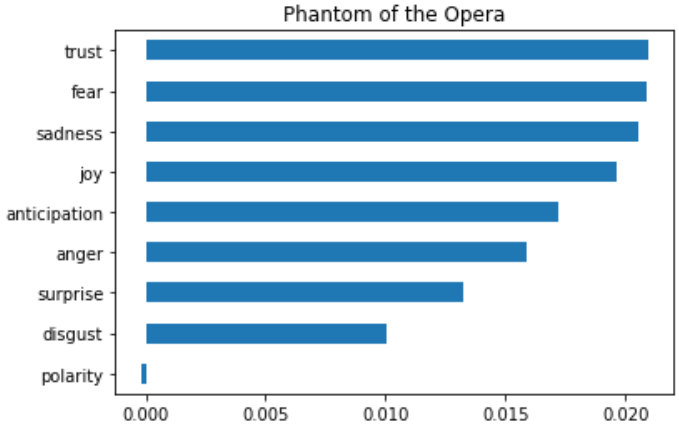
Sentiment Analysis

The results of the sentiment analysis really shocked me. My assumption would be that for all three books, the overwhelming feeling would be ‘fear’. Instead, a different theme seemed toi connect the novels. That theme was ‘Trust’. Ultimately, I think this does make sense, despite the corpus being centered around scary stories and thrillers. I think that trust, or lack of trust, can be the foundation of fear. In Dracula, the team of Light must trust one another to defeat evil. There are many quotes in Dracula that reference trust directly. “If you trust me not, then I must tell what I think; and that is not perhaps well.” (Chapter 23, Dracula).

Phantom of the Opera and Dorian Gray have these themes as well. The Phantom wants Christine to fall in love with him, which requires trust. In Dorian Gray, Lord Henry immediately trusts Dorian, because “...something in his face that made one trust him at once.” (Chapter 2, Dorian Gray). The other results of the sentiment analysis were expected and followed the general plot of each novel.







Conclusion:

Overall, I think this corpus shows a common theme across Gothic novels. The classic works by Oscar Wilde, Bram Stoker, and Gaston Leroux have terrified and delighted audiences for decades. All three of these works were adapted to live theatre plays and muscials, and were all turned into movies early in the 20th century. The 1931 Movie adaptation of Dracula has been credited with creating the cinematic horror genre (Oliver, 2016).. This movie may have been inspired by the 1925 silent film Version of the Phantom of the Opera, which was released 8 years after the silent film adaptation of Dorian Gray. The Picture of Dorian Gray (1945) earned an academy award for best black-and-white cinematography (Lansbury, n.d.).. When analyzing their texts, their themes of life, death, love, and change come across clearly, and help provide context to how these works helped to build the foundation of horror as we know it today.

References

Oliver, J. (2016, October 20). *Why I love... Bela Lugosi's Dracula*. BFI. Retrieved December 9, 2022, from https://www2.bfi.org.uk/news-opinion/news-bfi/features/why-i-love-bela-lugosis-dracula