Comparing Vulgarity of Texts

Bryson Gullett

Questions

- How much more vulgar is *Ulysses* by James Joyce (an infamously vulgar book) compared to *Alice's Adventures in Wonderland* by Lewis Carroll?
- How much more vulgar is *Ulysses* by James Joyce (an infamously vulgar book) compared to *Dubliners* by James Joyce?

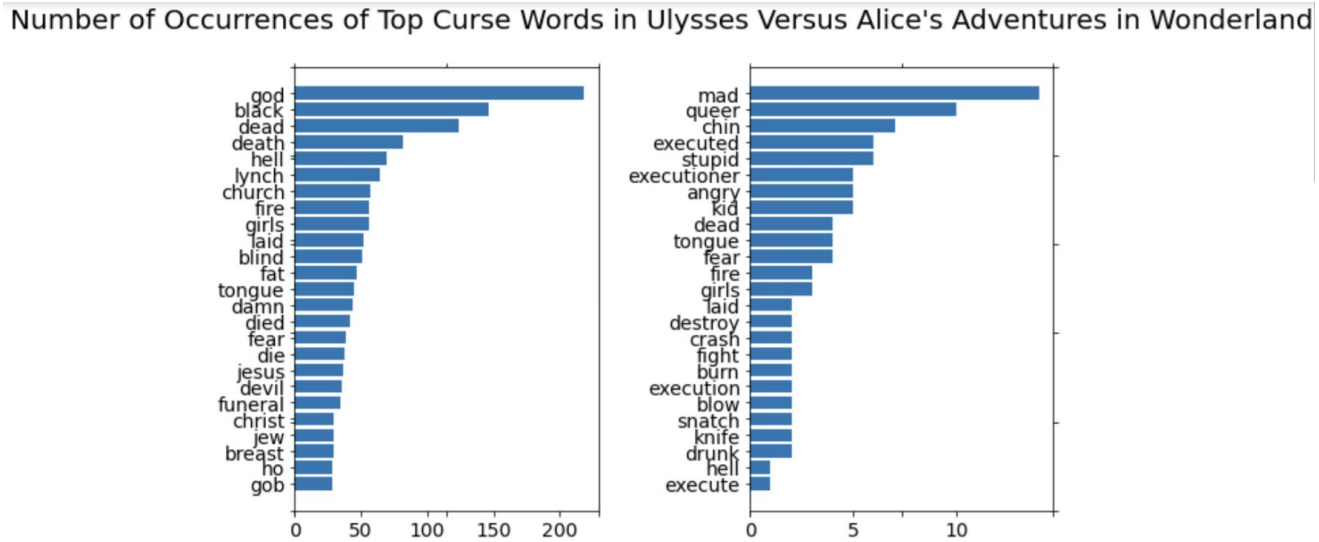
Approach

- Compare vulgarity of texts by comparing the number of occurrences of curse words in each text
- Use dataset of curse words from Kaggle to define and count curse words in each text: https://www.kaggle.com/datasets/nicapotato/bad-bad-words
- Get texts from Project Gutenberg
- Graph top 25 curse words and their number of occurrences for each text in a bar graph
- Graph top 25 curse words and their relative frequency (number of occurrences / total words in text) for each text in a bar graph

Problems Encountered

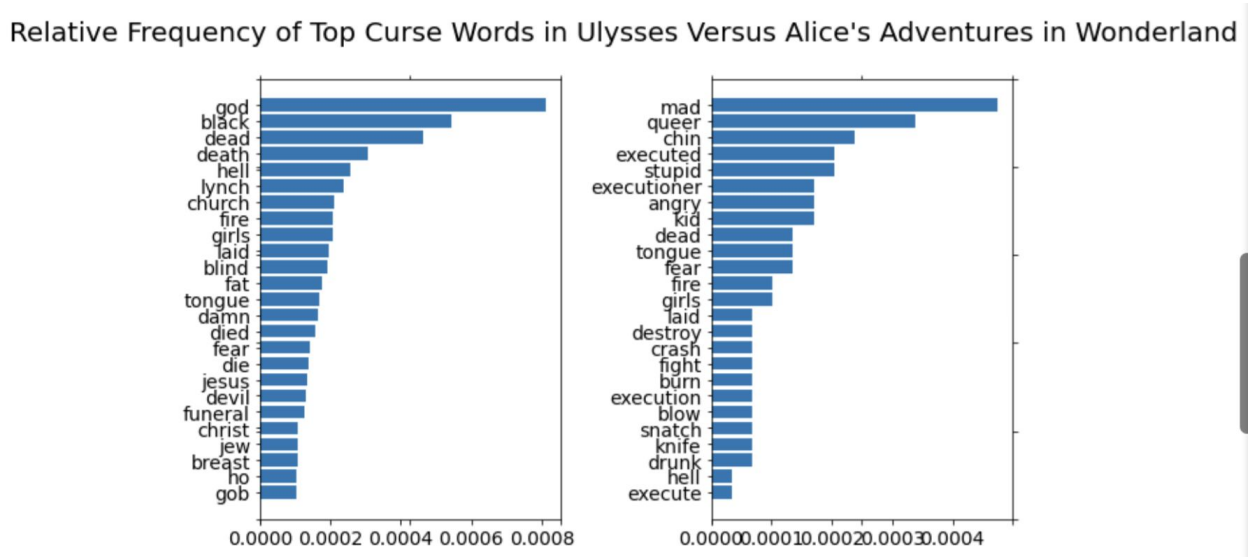
- Kaggle dataset I found considers some rather clean words as curse words (e.g. tongue and angry)
- I weigh every word in the dataset as equally bad
- Formatting Matplotlib plots is not fun

Results - *Ulysses* vs *Alice's Adventures in Wonderland*



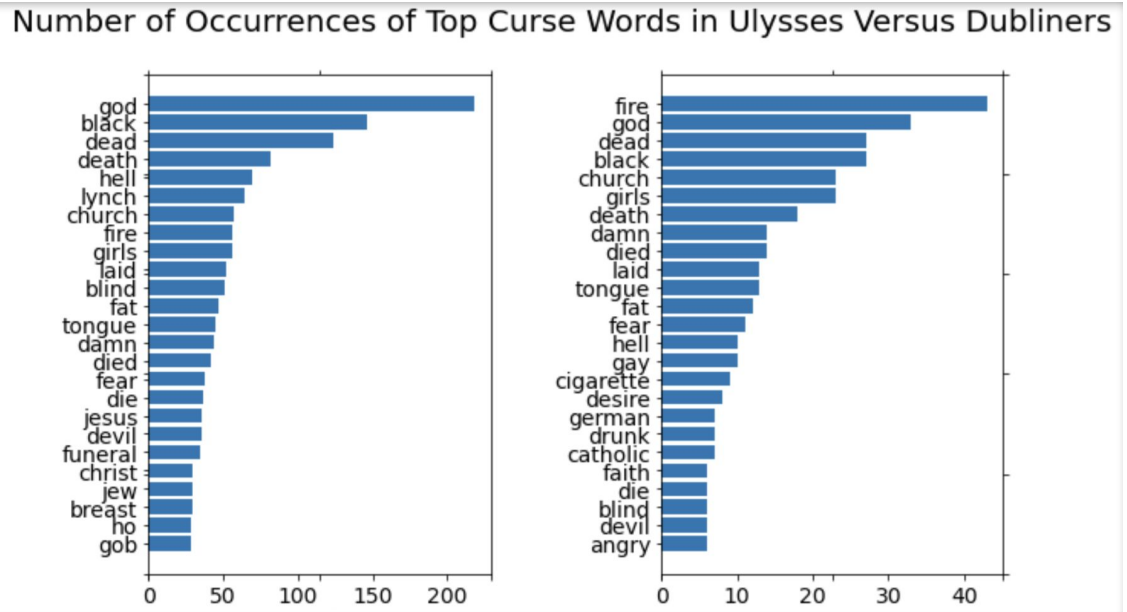
Ulysses seems to have far more occurrences of bad words than *Alice's Adventures in Wonderland*. However, it also has far more total words in the text. Some of the curse words in each text overlap, such as "dead" and "hell." However, they are pretty different lists overall, indicating that the authors' vocabularies are quite different.

Results - *Ulysses* vs *Alice's Adventures in Wonderland* (Cont)



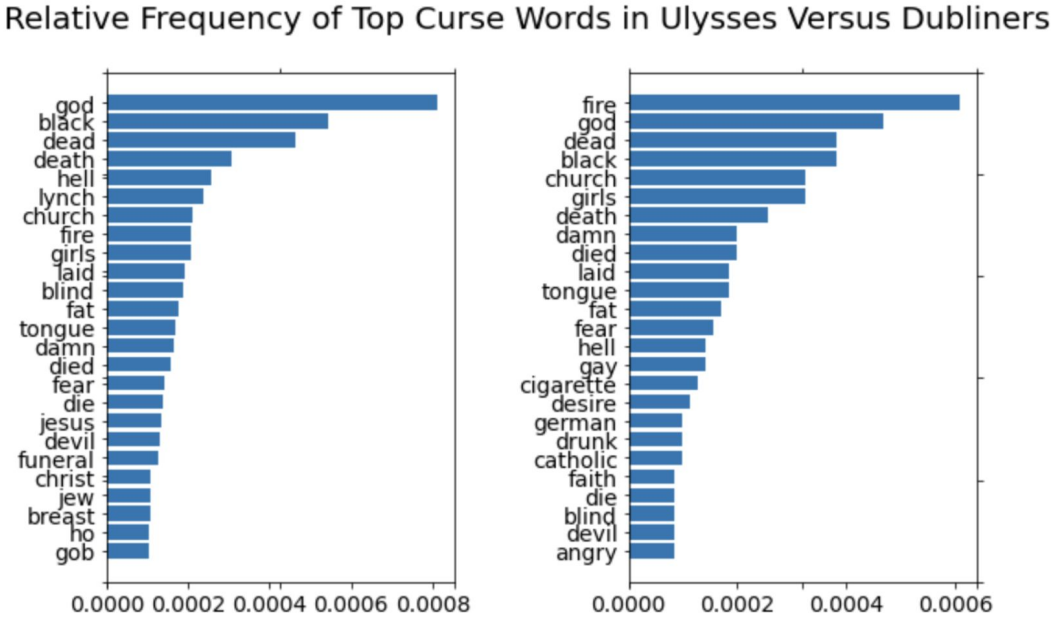
Ulysses seems to have bad words more frequently than *Alice's Adventures in Wonderland*. However, it is much closer than I initially expected because *Alice's Adventures in Wonderland* has curse words occurring more frequently than I thought they would.

Results - *Ulysses* vs *Dubliners*



Ulysses seems to have more occurrences of bad words than *Dubliners*. Note that James Joyce uses a lot of the same words in his texts, such as “god” and “black.”

Results - *Ulysses* vs *Dubliners* (Cont)



Ulysses seems to have its bad words more frequently than *Dubliners* as well. However, the word “damn” is used more frequently in *Dubliners*, and it is one of the “worst” words on this list.

New Ideas

- Could manually assign weights of how bad I believe each word is to calculate a number to compare how vulgar each text is
- Could find texts being protested by parents on K-12 student reading lists and compare them to more socially acceptable texts