Anthony Johnson

IST664  
NLP Homework 1

1. I chose Moby Dick and Paradise Lost from the NLTK Gutenberg collection.
2. For finding the top 50 frequencies, I filtered out NLTK stop words and added the stop words from week 2 lab. I also filtered out non-alphabetical characters.

Top 50 words and bigrams in Moby Dick:

('whale', 1086) ('one', 912)

('like', 580)

A screen shot of a computer code

Description automatically generated ('upon', 565)

('ahab', 508)

('man', 490)

('ship', 463)

('old', 443)

('ye', 438)

('sea', 384)

('though', 383)

('yet', 344)

('time', 326)

('captain', 324)

('long', 318)

('still', 312)

('said', 304)

('great', 303)

('two', 288)

('boat', 287)

('seemed', 283)

('head', 277)

('last', 275)

('see', 269)

('thou', 268)

('whales', 267)

('way', 264)

('stubb', 254)

("n't", 252)

('queequeg', 252)

('little', 249)

('white', 248)

('round', 245)

('say', 241)

('sperm', 240)

('three', 239)

('men', 237)

('may', 237)

('first', 235)

('every', 232)

('us', 228)

('much', 223)

('well', 220)

('never', 206)

('hand', 203)

('good', 197)

('almost', 195)

('starbuck', 195)

('ever', 193)

('go', 192)

​

Top 50 words and bigrams in Paradise Lost:

A computer code with many black and white text

Description automatically generated with medium confidenceA close-up of a computer code

Description automatically generated

Top 50 bigrams with Mutal Information scores

A white background with black text

Description automatically generatedMoby Dick Paradise

A screen shot of a computer code

Description automatically generated

2a. I decided that extra words like ‘the’ or ‘in’ would be extraneous, so I filtered those, along with contractions. I figured these filters would allow the more meaningful words to be analyzed.

2b. I think I may have forgot to apply the filter to my bigrams for Moby Dick, so I could fix that to get a better result. I could have also found the most frequently used words that might be considered extraneous and added those to the filter.

2c. I’ve found that the two different bigram types seem to be pretty similar in that they’re all logically related in some way. It makes sense for ‘ten’ and ‘thousand’ to be found together frequently.

1. I wanted to compare the language of two texts that were separated by the eras they were written in, and I found that Moby Dick and Paradise Lost were 200 years apart. I wanted to see how the language in each text differed.

Based on the results of the frequency distributions of words and bigrams, I can see that the language between the two are very different. I expected a difference, but nothing as major as this. For example, Paradise Lost uses the words ‘thou,’ ‘shalt,’ and a few other archaic words that have long since been replaced with more modern iterations. Moby Dick seems a little more modern in its language, though it’s harder to parse through with its nautical vocabulary. It also seems as though Moby Dick is much more descriptive and ‘flowery’ with its words than Paradise Lost, which seems like it’s mostly addressing someone (several iterations of bigrams beginning with ‘thou’).

It seems like the vocabulary of Moby Dick is much more expanded than that of Paradise Lost, which has many repeating words in its bigrams.