Justin Cai

Period 7

7. When I looked at the first 25 uses of “however” in text1 of nltk.book, I noticed that it generally was not used correctly (according to Strunk and White’s *Elements of Style*). Whenever a comma followed the word however if it was the first word of the sentence, it was used improperly. Conversely, it was used in the correct context when it was not followed by a comma.

23. 

I used text1. This confirms Zipf’s Law, as there is clearly an inverse relationship between word frequency and word rank. At the left extreme, the line seems to be approaching roughly 19000, and the right extreme seems to be approaching 0.

27. 

Average polysemy for nouns: 1.2610825311125826

Average polysemy for verbs: 2.1865729898516784

Average polysemy for adverbs: 1.2532916759651864

Average polysemy for adjectives: 1.406536617160948

Essentially what I did was loop through a list of all adjectives and, for each adjective, add the number of synsets for said adjectives to a variable count that I set to 0 before the loop. I then took this count and divided it by the total number of adjectives to get the average polysemy. I repeated this for nouns, verbs, and adverbs.

9. I analyzed the word “though” in text1 and text2. I found this word by first removing all stopwords from both text1 and text2 and, using FreqDist(), looked at the 50 most common words in both texts. Though was one of the first words to appear in both lists of most common words, so I looked at the usage of though in the two texts using the concordance method.

Text2 used though as a word for “even if,” a contrast word. For example:

His manners , though serious , were mild ;

Elinor was obliged , though unwillingly , to believe that

Text1 used though as a continuation word of sorts. I’m not really sure how to describe it, but an example is:

For though I tried to move his arm – unlock

Though true cylinders without -- within ,

Basically, “though” in text2 was usually followed by a comma, indicating contrast, whereas though in text1 was not.

12.

Number of distinct words: 123455

I got this by looping through entries in the CMU Pronouncing dictionary, and add the word to a list I made if it was not already in the list. I then took the length of that list at the end to find the number of distinct words.

7.49% of words have more than one possible pronunciation. I did this by once again looping through the dictionary and adding one to a count I instantiated every time I got a word that I had already seen, and I also accounted for words with over two pronunciations in the process.

15. There are 22339 words in the Brown Corpus that occur 3 or more times. I did this by creating a list of all brown words by using brown.words(). I then created a Frequency Distribution of this list and looped through it, adding 1 to my count every time the frequency of a word was greater than 3.

18. 

I first created a list of all bigrams in text1, and looped through this list. If either words in a bigram was a stopword, I wouldn’t add it to my list. I added the bigram to my list otherwise. I then created a freqdist of this list and printed the 50 most common ones.