Maria Frey
LING 492B
18 February 2021

Lab 1 Questions

Part 2 Questions:

Given these words, it's pretty hard to guess what the topics could be because the words at the very top of the list don't necessarily speak for the whole list. I would guess corpus 1 is about food/eating, corpus 2 is about language, corpus 3 is maybe about animals/pop culture, corpus 4 is about politics, corpus 5 is about sports/football/super bowl, and the mystery corpus is also about food/eating. Corpus 3 is definitely the hardest to guess because there doesn't seem to be an obvious pattern.

A couple of things that I found really interesting and funny about the data were that in both corpus 1 and the mystery corpus, lunch was much less frequently mentioned than breakfast and dinner (do people not like lunch as much?) and I also found it funny that the name of the kpop idol 'Jungkook' appears at varying frequencies in all of the corpora except corpus 4. So I guess it's good to know that kpop has infiltrated all aspects of life except politics…

Part 3 Questions:

Based on my scatterplot, this data does not follow Zipf's law, as my scatter plot does not show a linear relation between the log frequency and log rank. I'm not sure if something went wrong here and where it went wrong, because looking at the relative word frequencies, it does seem as if the most frequent word occurs twice as often as the second most frequent word.

Part 4 Questions:

a) The value of N that worked well for me was 200.
b) The topics of some of the corpora is clearer, or at least easier and more efficient to see. I still stand by my guesses for all of the corpora, except for corpus 3, which I still don't really know what the topic could be. In order to answer the Part 2 question, I was sorting out the stop words myself, so I kind of did the same thing as the program in terms of filtering out stop words. Interesting side note here, it seems that because of the high frequency of "Trump" cross the corpora, it was considered a stop word.
c) We didn't necessarily need to use more corpora, but it works better with more information to pull from. If we had just removed the top N words from each corpus, then it wouldn't have gotten rid of enough stop words, or it would have gotten rid of more words than we want it to. For example, if I had filtered out the first 200 words of just corpus 1, then it would have filtered out useful content words as well as the stop words.
d) If each topic had a list of key words, the program could compare those words with the high frequency content words of the mystery file. The topic with the most matches to the mystery file would be the topic where the mystery file belongs.