Discussion

The output shows pairs of words that are ranked in terms of the sum of standard deviation and frequency. The standard deviation is that of the frequencies of all words used with A, and frequency is that of B used with A. Thus, pairs that are used exclusively with each other but still used frequently will be placed at the top. Observing the output, such requirement is fulfilled. To name a few examples, in the bigram “high school”, “high” is used most frequently with “school” and exclusively that the next word used with “high” is “speed” which is ranked at 194th, though not in the output file. Similarly, “great” is used most frequently with “deal”, ranked 6th, compared to “great number” ranked 549th. One could argue “high” and “great” should not rank high because they can be used to describe many other words. However, according to the Brown corpus, that is not the case. “High” and “great” are used relatively exclusively with “school” and “deal”. Even if the universal use of “high” and “great” is considered, “high school” and “great deal” are exclusive use cases.

I considered frequency in the ranking because word pairs that are always used with each other but appears only once should not rank high. The sample size would be too small. For example, “high speed” and “great number” was ranked low because they were not used frequently enough. In other words, pairs that are used both exclusively and frequently should rank at the top. I have tried different weights for the standard deviation and frequency to check which leads to best results. Since there was no quantitative method to measure the correctness of the ranking, I evaluated it with my own English knowledge. I tried ratios ranging from 1:2 to 5:1 for standard deviation to frequency ratio, but larger the weight of standard deviation, technical terms such as “fiscal planning” and “pulmonary artery” ranked higher. Such results would be the case in which the exclusivity and frequency were not balanced ideally. Through trial-and-error, I concluded that a 1:1 ratio gave the best results.

Potential Improvements

In the previous homework, I could provide “seed” words to find intensifiers and use them as necessary. But because “seed” words are not allowed in this assignment, I took a more implicit approach. I did not explicitly look for words that intensify the word it describes. Instead, I compared the exclusivity and frequency of each word pairs. With adjustment of the weights of the two parameters, intensifiers ended up high in the ranking, namely “only”, “high”, “old”, “great” and “long” etc. Thus, I concluded that, apparently, ranking in terms of exclusivity and frequency is effective in finding exclusively used intensifiers. If there was a more explicit way to determine if an adverb/adjective actually intensifies the word it is modifying, the results may have been more accurate. Also, if plural forms of words that are modified were considered as a single use case, the ranking could have been slightly different. However, I decided that different plural forms of a same word is a different use case and should be counted separately and did not include in the algorithm.