**Davis Busteed – LING 360 – HW #7**

Results

Although the Python script I wrote only focused on identifying “noun pairs,” I was still able to pick up on the general topics and feelings of the six registers. Unlike previous assignments, I have no idea what these six different registers are, as well as what their abbreviations stand for (I couldn’t find any information about this on the course site or online). This actually made the analysis more interesting, since I was making unbiased inferences.

Just because the frequency lists are sorted from largest to smallest didn’t necessarily mean I could simply skim the top 20 noun pairs for each register to hypothesize its topic. There are many noun pairs that only show up once in the register, so at a certain point, all of the noun pairs will share the same normalized frequency value. I focused my attention on the “top” noun pairs that clearly appeared in the registers more than once.

With this in mind, I started to make some hypotheses about the registers. For example, I found that three registers that end with “HI” are quite similar. It seems that the “HI” registers share similar topic of civics, society, and politics. Although there might be some slight differences between the three in their primary focus, they share a couple of frequent noun pairs, making me think that they are pretty similar in content. For example, JA\_HI appears to mention more economic terms than PS\_HI, but they both share phrases relating to the government / society as a whole.

Similar to this, the three registers with the “BI” suffix seemed to share a common topic, but in this case, they were rooted in the sciences. For example, JA\_BI’s top noun pairs were “mass loss,” “motor neuron,” and “heart disease,” indicating that many of the articles might be scientific journals or lab studies. PS\_BI on the other hand, seems to be more focused on ecology and environmental studies. The noun pair “cell wall” reminds me of biology class, so it seems to me that TB\_BI’s main focus is biology.

Sample Measurements

To assess the performance of my script, I read through one of the files (TB\_BI\_02.txt) and manually counted noun pairs. I modified my script to give me a result for only this file so that I could calculate a precision and recall score. The calculations for these measurements are shown below.

I also calculated the normalized count for the noun pairs in my sample. The normalized count per 1000 words for a noun pair that occurred only once in the sample was about 1.92, but when the script calculates this normalized count in relation to all the words in the register, it is no surprise that this normalized count drops down to 0.08.