**Davis Busteed – LING 360 – HW #4**

Features

I searched the Mini Core corpus to obtain normalized counts (per 1000 words) for the following linguistic features:

* Contractions – words made by shortening and combining two words (ex. don’t, isn’t).
* Possessive Pronouns – pronouns that demonstrate ownership. From what I read online, it appears that possessive pronouns can be split between words like “my” and “mine,” but since most online resources consider both possessive pronouns, I did too.
* Questions marks – the punctuation mark (?) that ends a question.

Hypothesis

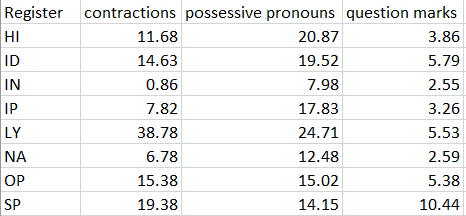
After selecting the features I would be searching for, I made some general predictions about what kind of results I would see.

From a past assignment I learned that contractions are more common in informal text, so I hypothesized that the normalized count for contractions would be highest in song lyrics, interactive discussions, and interview transcripts. I believed these texts are more likely to capture regular, plain speech, so I thought they would have more contractions than other registers. Along those same lines, I guessed that more texts like news reports would have much fewer contractions.

I had a harder time predicting the number of possessive pronouns, since I believed they would be frequently used in all registers. I did think, however, that song lyrics would have a lot of possessive pronouns, since a lot of songs are more personal, often telling stories that include words like “my,” “his,” and “her.” For the other registers, I imagined they will be pretty similar in their normalized counts.

Predicting the number of question marks within a text was a little easier, especially since one of the registers is interview transcripts. I believed this register would have the highest count for obvious reasons. Following that, I hypothesized that interactive discussions would be next highest because questions are an integral part to discussions. I also predicted that the informational register would have the lowest amount of question marks, since these texts generally state facts, rather than asking questions.

Results

After reading though each file in the Mini Core corpus, my Python script calculated normalized counts for each target feature and produced the following summary report.

As I looked back at my original hypotheses, I found that I was correct in most of my predictions, although I may have missed a couple of interesting insights.

I was correct in thinking that song lyrics would have the highest number of contractions, although I didn’t expect it to be almost two times larger than the next highest category. I also correctly predicted that interactive discussions and interviews would have high counts, but I forgot to predict that opinion blogs would also have a high count. I should have realized that blogs are generally written to an informal audience, so that contractions are quite common. I was correct in assuming news reports would have less contractions, but I was really surprised that informational texts had such a low count. I must’ve looked over this register or assumed that because it was Wikipedia there might be an acceptable amount of informal speech.

As with contractions, song lyrics had the highest number of possessive pronouns, something that I had hypothesized. I also predicted that the counts across the other registers would be relatively flat, which wasn’t entirely correct. It seems that both instructional texts and interactive discussions had high counts. Looking back, this makes a lot of sense, because how-to manuals and instructions will often give instructions in relation to the reader, using words like “your.”

I wasn’t surprised that I correctly guessed that the interview register would have the highest normalized count of questions. This value was about double that of interactive discussions, the next highest count. I did correctly predict that discussions would have a high question count, but I was surprised by the high count of song lyrics. Now that I think about it, many song lyrics are actually questions. As for the lowest count, I correctly hypothesized that information readings would have the fewest number of question marks, because their main purpose is to state facts.

Sample Measurements

To assess the general performance of my feature collection methods, I calculated a sample precision and recall score.

|  |  |  |
| --- | --- | --- |
| File: 1+SP+IT+SP-SP-SP-SP+IT-IT-IT-IT+NNNN+3798809.txt | | |
|  | Precision | Recall |
| Contractions | 100% | 92.3% |
| Possessive Pronouns | 100% | 100% |
| Question Marks | 100% | 100% |