**NLP Homework 1: Comparing Corpora with Corpus Statistics**

In this problem set, we are presented the option of picking two texts and comparing them using corpus statistics. For this homework, I decided to use the Caesar and Hamlet from the Gutenberg collection. The reason for my pick is the fact that both texts are written by the same author, that being Shakespeare. By comparing the approaches to both texts, I want to see how much they differ, if at all. If the two texts from the same author do differ substantially, then it will shine a light on how to approach different texts from the same authors. For example, exploring the tweets of the president. Do I need a different approach per month? Per topic? Or do I use the same approach for all of his tweets.

Before finding the frequencies of all the words per text, I went ahead and tokenized both texts using the regular tokenizer shown in class. Unfortunately, I had to change my approach as the basic tokenizer treats a word with an apostrophe as two words. This would cause the bigram analysis to consider words with apostrophe’s and mask the frequency of actual bigrams. This was the case for the Hamlet text. Therefore, I used the TweetTokenizer instead which treats the words with apostrophe as one. The second step I took was converting all the letters to lowercase, and removing the stop-words and punctuation before getting the frequency per word.

|  |  |  |  |
| --- | --- | --- | --- |
| **Caesar** | | **Hamlet** | |
| **word** | **freq** | **word** | **freq** |
| caesar | 0.01719 | ham | 0.021194 |
| brutus | 0.014566 | lord | 0.01327 |
| bru | 0.013842 | haue | 0.011006 |
| haue | 0.0133 | king | 0.010754 |
| shall | 0.011309 | shall | 0.006729 |
| thou | 0.010404 | thou | 0.006603 |
| cassi | 0.009681 | come | 0.00654 |
| cassius | 0.00769 | hamlet | 0.006289 |
| antony | 0.006785 | good | 0.006163 |
| come | 0.006695 | hor | 0.005974 |
| good | 0.006333 | let | 0.005912 |
| know | 0.006152 | thy | 0.00566 |
| enter | 0.00579 | enter | 0.005346 |
| men | 0.00579 | oh | 0.005094 |
| let | 0.00579 | like | 0.004842 |
| vs | 0.005609 | well | 0.004402 |
| thy | 0.005066 | tis | 0.004339 |
| heere | 0.005066 | know | 0.004339 |
| man | 0.004976 | would | 0.004276 |
| thee | 0.004976 | selfe | 0.004214 |
| ant | 0.004343 | may | 0.004088 |
| well | 0.004343 | loue | 0.004088 |
| vpon | 0.004252 | vs | 0.003899 |
| day | 0.003981 | sir | 0.003899 |
| would | 0.003981 | qu | 0.003899 |
| lord | 0.003981 | laer | 0.003773 |
| yet | 0.0038 | giue | 0.00371 |
| go | 0.003709 | thee | 0.003648 |
| selfe | 0.003528 | ile | 0.003648 |
| caes | 0.003528 | must | 0.003648 |
| noble | 0.003528 | hath | 0.003585 |
| like | 0.003528 | ophe | 0.003522 |
| rome | 0.003438 | speake | 0.003459 |
| heare | 0.003438 | make | 0.003396 |
| caesars | 0.003438 | say | 0.003207 |
| cask | 0.003438 | vpon | 0.003144 |
| night | 0.003438 | father | 0.003144 |
| say | 0.003348 | doe | 0.003144 |
| may | 0.003348 | pol | 0.003082 |
| see | 0.003348 | go | 0.003019 |
| brut | 0.003348 | one | 0.002893 |
| tell | 0.003348 | man | 0.002893 |
| must | 0.003257 | see | 0.00283 |
| speake | 0.003167 | heere | 0.00283 |
| stand | 0.003167 | time | 0.002767 |
| giue | 0.003167 | mine | 0.002767 |
| hath | 0.003167 | much | 0.002704 |
| loue | 0.003076 | heauen | 0.002704 |
| one | 0.003076 | tell | 0.002704 |
| vp | 0.002895 | rosin | 0.002704 |

For the Bigram section of the analysis, I was able to follow similar patterns for both texts to get a substantial amount of information back with the most common bigrams. For the raw frequency analysis, I followed a similar pattern where I used the alpha-filter function, removed the stop-words, and got rid of words with length less than 3.

|  |  |  |  |
| --- | --- | --- | --- |
| **Hamlet** | | **Caesar** | |
| **bigram** | **freq** | **bigram** | **freq** |
| (lord, ham) | 0.001223 | (mark, antony) | 0.00052 |
| (enter, king) | 0.000333 | (marke, antony) | 0.00048 |
| (enter, hamlet) | 0.000278 | (thou, art) | 0.00044 |
| (haue, seene) | 0.000278 | (enter, brutus) | 0.00036 |
| (lord, hamlet) | 0.000278 | (noble, brutus) | 0.00036 |
| (good, lord) | 0.00025 | (thou, hast) | 0.00036 |
| (thou, hast) | 0.00025 | (caesar, caes) | 0.00032 |
| (enter, polonius) | 0.000222 | (good, morrow) | 0.00032 |
| (fathers, death) | 0.000222 | (good, night) | 0.00032 |
| (lord, polon) | 0.000222 | (haue, done) | 0.00032 |
| (dost, thou) | 0.000195 | (lord, bru) | 0.00032 |
| (good, friends) | 0.000195 | (antony, ant) | 0.00028 |
| (let, vs) | 0.000195 | (enter, lucius) | 0.00028 |
| (haue, heard) | 0.000167 | (come, downe) | 0.00024 |
| (ile, haue) | 0.000167 | (euery, man) | 0.00024 |
| (mine, owne) | 0.000167 | (haue, seene) | 0.00024 |
| (thou, art) | 0.000167 | (would, haue) | 0.00024 |
| (enter, horatio) | 0.000139 | (caesar, shall) | 0.0002 |
| (hamlet, ham) | 0.000139 | (caius, cassius) | 0.0002 |
| (set, downe) | 0.000139 | (caius, ligarius) | 0.0002 |
| (tell, vs) | 0.000139 | (decius, brutus) | 0.0002 |
| (thy, selfe) | 0.000139 | (did'st, thou) | 0.0002 |
| (wilt, thou) | 0.000139 | (enter, antony) | 0.0002 |
| (would, haue) | 0.000139 | (fell, downe) | 0.0002 |
| (come, hither) | 0.000111 | (great, caesar) | 0.0002 |
| (dead, body) | 0.000111 | (haue, beene) | 0.0002 |
| (enter, ghost) | 0.000111 | (haue, heard) | 0.0002 |
| (father, lost) | 0.000111 | (honourable, men) | 0.0002 |
| (get, thee) | 0.000111 | (metellus, cymber) | 0.0002 |
| (giue, vs) | 0.000111 | (mine, owne) | 0.0002 |
| (good, horatio) | 0.000111 | (shall, finde) | 0.0002 |
| (good, night) | 0.000111 | (shall, haue) | 0.0002 |
| (hath, made) | 0.000111 | (caesar, doth) | 0.00016 |
| (let's, follow) | 0.000111 | (caesar, hath) | 0.00016 |
| (shall, heare) | 0.000111 | (euery, one) | 0.00016 |
| (shall, liue) | 0.000111 | (heere, comes) | 0.00016 |
| (sir, ham) | 0.000111 | (honourable, man) | 0.00016 |
| (take, thy) | 0.000111 | (messala, messa) | 0.00016 |
| (thou, shalt) | 0.000111 | (mine, eyes) | 0.00016 |
| (thy, soule) | 0.000111 | (noble, antony) | 0.00016 |
| (tis, true) | 0.000111 | (noble, caesar) | 0.00016 |
| (would'st, thou) | 0.000111 | (tell, thee) | 0.00016 |
| (art, thou) | 0.000083 | (thou, shalt) | 0.00016 |
| (christian, buriall) | 0.000083 | (wee'l, heare) | 0.00016 |
| (come, againe) | 0.000083 | (wilt, thou) | 0.00016 |
| (come, away) | 0.000083 | (yong, octauius) | 0.00016 |
| (enter, ophelia) | 0.000083 | (brother, cassius) | 0.00012 |
| (enter, queene) | 0.000083 | (brutus, sayes) | 0.00012 |
| (follow, thee) | 0.000083 | (caesar, lou'd) | 0.00012 |
| (good, faith) | 0.000083 | (come, hither) | 0.00012 |
|  |  |  |  |

Now, for the bigram analysis using the PMI metric, I did have to have a different approach per text, but the only difference is the fact that I had to get rid of the word ‘thou’ from the analysis in the Caesar text because it showed up too frequently.

|  |  |  |  |
| --- | --- | --- | --- |
| **Hamlet** | | **Caesar** | |
| **bigram** | **freq** | **bigram** | **freq** |
| (fathers, death) | 8.683937 | (caius, ligarius) | 10.06086 |
| (enter, polonius) | 7.403829 | (metellus, cymber) | 10.06086 |
| (mine, owne) | 7.306483 | (mine, owne) | 9.079473 |
| (good, friends) | 7.079866 | (fell, downe) | 8.822698 |
| (haue, seene) | 6.683937 | (mark, antony) | 8.380475 |
| (haue, heard) | 6.461545 | (marke, antony) | 7.64351 |
| (tell, vs) | 6.076615 | (good, morrow) | 7.573121 |
| (enter, horatio) | 5.725757 | (honourable, men) | 7.123867 |
| (enter, hamlet) | 5.403829 | (haue, seene) | 6.994584 |
| (lord, polon) | 5.166121 | (haue, beene) | 6.924195 |
| (enter, king) | 4.892867 | (caius, cassius) | 6.821392 |
| (lord, ham) | 4.476876 | (enter, lucius) | 6.716209 |
| (lord, hamlet) | 4.092121 | (euery, man) | 6.605542 |
| (good, lord) | 3.969264 | (haue, heard) | 6.561625 |
| (would, haue) | 3.918402 | (come, downe) | 6.461241 |
| (hamlet, ham) | 2.416615 | (good, night) | 6.232084 |
|  |  | (haue, done) | 5.88606 |
|  |  | (shall, finde) | 5.877975 |
|  |  | (antony, ant) | 5.602868 |
|  |  | (decius, brutus) | 5.512843 |
|  |  | (noble, brutus) | 5.1629 |
|  |  | (lord, bru) | 4.892475 |
|  |  | (caesar, caes) | 4.754036 |
|  |  | (enter, antony) | 4.702404 |
|  |  | (great, caesar) | 4.660927 |
|  |  | (would, haue) | 4.535153 |
|  |  | (enter, brutus) | 4.448302 |
|  |  | (shall, haue) | 2.765766 |
|  |  | (caesar, shall) | 2.395582 |

The only problem with the PMI metric, is that only a few bigrams met the full criteria.

Finally, the when we look at the results of both texts, we find that the texts are similarly analyzed which can account for an author signature or type of writing. In this case the main difference was the results from the analysis itself. Both tragedies revolve around death, but with different characters. We also see how the same characters are address differently (probably by different people) with nicknames like ‘Ham’ instead of ‘Hamlet’. Conversely, the Caesar text shows more respectfulness in its results. Mainly from the repeated use of words like noble and good. We can actually get some context on the type of characters we have from the bigrams selected.

All in all, the two texts written by the same author show similar signatures that can be interpreted using the same style of algorithms, but with a different resulting context. Not because of the different characters, but because of how the characters are addressed by the author.