Assignment 7 - WRITEUP.pdf

Victor Nguyen

March 13, 2022

1 Introduction:

In this writeup, I will be focusing on the behaviors of changing the amount of words to limit when calculating how closely related the anonymous text to known authors. I was curious to see the behavior of the distances across multiple noise limit specifications and also across the three different distance metrics. I want to first look at the Euclidean distance across noise limits of 100, 1000, 10000. I collected the top 5 authors closely related to the given text. This text that we'll use is from William Shakespeare (which I'll refer to as W), which is located in the resource texts directory. Here's what I've found.

2 Data Tables and Analysis:

2.1 Small.db Euclidean

Noise Limit: 100	Noise Limit: 1000	Noise Limit: 10000
Saxo Grammaticus 0.029267891288144	Thomas Carlyle 0.026505649814581	H. G. Wells 0.037680408705579
Thomas Carlyle 0.029698589149425	Joseph Conrad 0.028163150282689	Saxo Grammaticus 0.040111636287486
Henry van Dyke 0.031161917833862	Saxo Grammaticus 0.028180586640447	Thomas Carlyle 0.040910370525885
Joseph Conrad 0.031478177284140	Henry Fielding 0.028435717552362	Henry Fielding 0.041269828103164
Daniel Defoe 0.032451281509624	Henry van Dyke 0.028533433503289	Henry van Dyke 0.041548902374772

According to the above data, having an absurdly large noise file can help increase the distance between other authors to the W text. This is a good thing since It can minimize false positives. Whats interesting though is that for some reason, when we filtered out about a thousand words, it detected that the authors were more closely related to file W compared to when we filtered out 100 words. The only reason I could think of why this may be the case is that the words contained in the noise.txt file from 101 to 1000 contains words that pertained to just the authors text but not in W. This can reduce the distance between the two texts. We should also check how these statistics may differ with a bigger data base containing more authors and samples.

2.2 Medium.db Euclidean

Noise Limit: 100	Noise Limit: 1000	Noise Limit: 10000
William Shakespeare 0.0	William Shakespeare 0.0	William Shakespeare 0.0
Dante Alighieri 0.026849331782958	Various 0.025674174983406	Charles Dickens 0.036072143201462
Edgar Allan Poe 0.027813156552717	Thomas Carlyle 0.026505649814581	Dante Alighieri 0.037279657630622
Saxo Grammaticus 0.029267891288144	Charles Dickens 0.026692761169669	H. G. Wells 0.037680408705579
Various 0.029394565260006	Dante Alighieri 0.027600452877624	Various 0.039057973424277

With a medium sized data base, the author we used for W is actually contained in here. It makes sense why the top person that's closely related to text W is himself. We can see a similar trend with a medium sized data base with the small data base. The differences between filtering out 100 words v.s. 1000 is marginally smaller though. So far, the bigger the noise limit, the better to reduce false positives. Lets test an even bigger data base.

2.3 Large.db Euclidean

Noise Limit: 100	Noise Limit: 1000	Noise Limit: 10000
William Shakespeare 0.0	William Shakespeare 0.0	William Shakespeare 0.0
William D. McClintock 0.026421981806312	Various 0.025674174983406	Charles Dickens 0.036072143201462
Dante Alighieri 0.026849331782958	Max Beerbohm 0.026101625497083	Dante Alighieri 0.037279657630622
Edgar Allan Poe 0.027813156552717	Tobias Smollett 0.026334790211751	H. G. Wells 0.037680408705579
Johann Wolfgang von Goethe 0.027861620127910	Washington Irving 0.026391571550681	Tobias Smollett 0.038317013348967

The results of this test is actually more similar to the medium data base. I find it most intriguing that some files ended up with the same calculated distance. Other than that, not much more is really interesting. Now, this was only testing 1 type of metric. I think it would be a good idea to test the other metrics to see if the results are any different.

2.4 Small.db Manhattan

Noise Limit: 100	Noise Limit: 1000	Noise Limit: 10000
Thomas Carlyle 1.256027242264269	Thomas Carlyle 1.517450008451044	Thomas Carlyle 1.857994029718847
Saxo Grammaticus 1.261176114273505	Saxo Grammaticus 1.521314432436919	Saxo Grammaticus 1.881410408709309
Unknown 1.287125206767091	Joseph Conrad 1.555874086897223	Joseph Conrad 1.882307334826114
Joseph Conrad 1.293547608536969	Henry Fielding 1.573457424227844	H. G. Wells 1.897684490021930
Henry van Dyke 1.299102579843970	Henry van Dyke 1.578112089239797	Henry Fielding 1.912260562414285

Looking at this data, it appears that there is some sort of positive growth in relation to the noise limit and the calculated Manhattan distance. Compared to the Euclidean metric, it looks like these calculations are more straight forward, whereas the Euclidean had a relative max with a noise limit of 100 and another relative max at the noise limit of 10000. It also seems like this trend continues on even with larger data bases. You can see the results below.

2.5 Large.db Manhattan

Noise Limit: 100	Noise Limit: 1000	Noise Limit: 10000
William Shakespeare 0.0	William Shakespeare 0.0	William Shakespeare 0.0
Christopher Marlowe 1.064019665173879	Christopher Marlowe 1.398049764612895	Christopher Marlowe 1.746960983282467
John Webster 1.122331537957684	Dante Alighieri 1.443163191214434	Dante Alighieri 1.799606953149556
Dante Alighieri 1.174836433744908	Charles Dickens 1.466263260624981	Charles Dickens 1.811452971225383
Alexander Whyte 1.177925622521591	Honore de Balzac 1.466944905519075	Honore de Balzac 1.826573971124117

As expected, we were able to accurately identify text W with its own file. We can see the similar positive growth as with the small data base. Moving onto the metric for Cosine, we can notice something that is very unique.

2.6 Large.db Cosine

Noise Limit: 100	Noise Limit: 1000	Noise Limit: 10000
Elizabeth Barrett Browning 0.998922776695346	William Shakespeare 0.999406634897737	William Shakespeare 0.998805923464217
William Shakespeare 0.998929343147098	John Webster 0.999735030800194	A. A. Milne 0.999792289502788
John Webster 0.999152436570009	John Dryden 0.999766407859951	John Dryden 0.999794540390079
Richard Brinsley Sheridan 0.999210512191703	Christopher Marlowe 0.999814995482465	Christopher Marlowe 0.999887830142763
Christopher Marlowe 0.999217420310796	Richard Brinsley Sheridan 0.999815806827864	John Webster 0.999897354170611

Under the noise limit of 100, the cosine metric actually identified an author by the name of Elizabeth to be more closely related to text W then William Shakespeare himself. That is quite bizarre and left me wondering why. The only thing I could think of is the added bonus of us computing the magnitude of the vector. It goes to show that we should be using multiple metrics to calculate the distance of one author compared to another.

3 Conclusion/What I learned:

1. Use multiple metrics to reduce the chances of false positives while trying to identify plagiarism. It would be a shame to flag someone falsely for plagiarism when in reality they didn't.

2. Like in asgn3, we should be extra careful dealing with decimal numbers since we cannot represent them as accurately as we would like.

3. We need a good balance for our noise words, as this may alter our results. The same attention should be applied to what we identify as a word.

4. We should also be careful with the size of our hash tables. Its important so that we can avoid word collisions. If we didn't make a big enough hash table, we could over write some words in our hash table which can influence our author identification significantly.