Assignment 7: Author Identification

Aniket Pratap

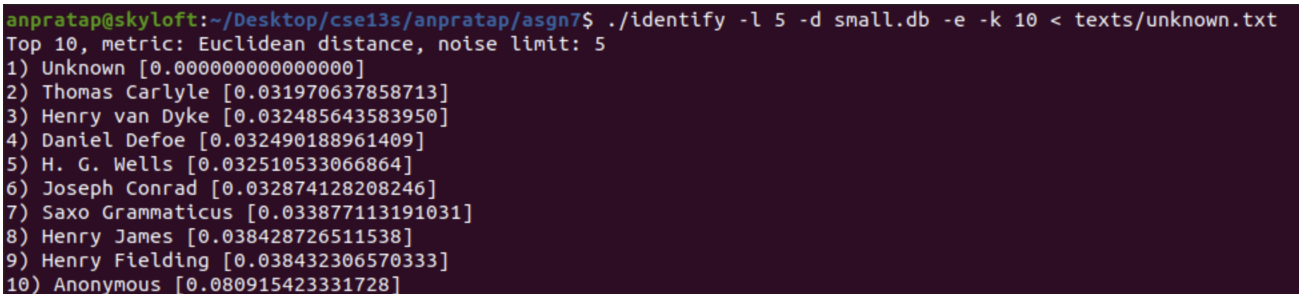
March 2022

1 Introduction

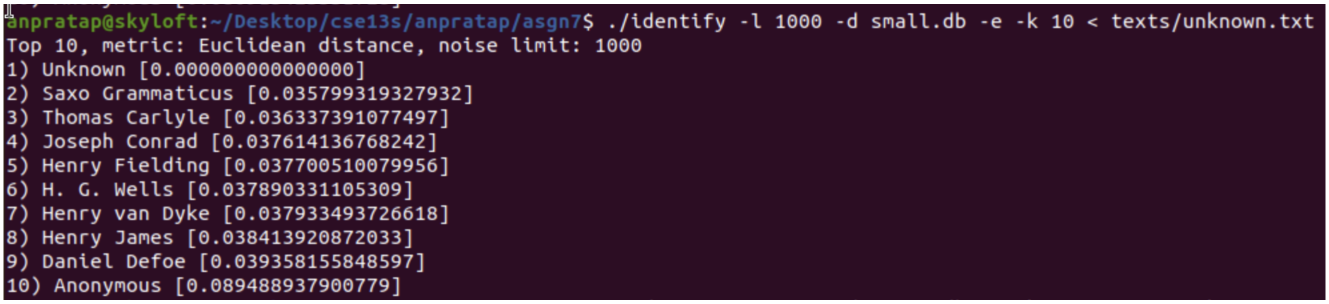
The propose of this write up is to find how accurate this program is based on different variables. These variables will consist of how many noise words used throughout the program and the size of the text inputted. This write up will also check how the different metrics (Euclidean, Manhattan, and Cosine) compare with each other.

2 Tuning Noise

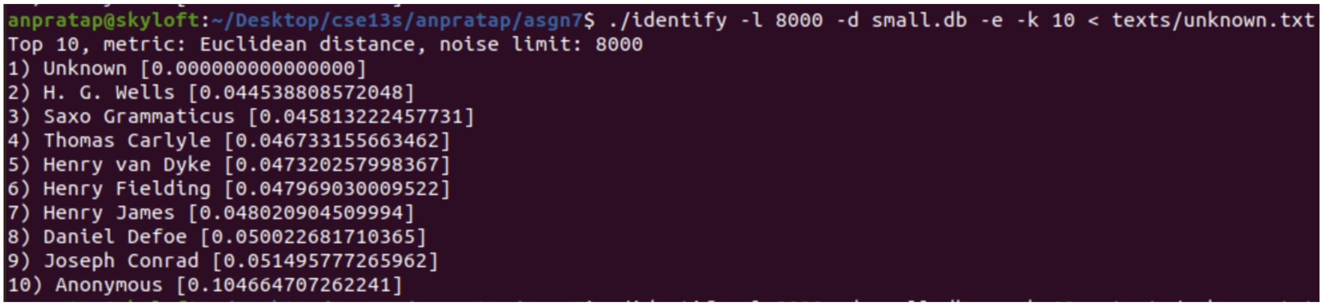
A noise text file determines which words in a given text are counted, and which are not. The noise words, in the case of this assignment, are the 10,000 most common used words. These words make sense as the noise, because common words such as "the" and "and" don’t really help identify the author——since all authors use these words. In order to combat this, a noise text is used to filter out common words. When using texts/unknown.txt as the anonymous text, it could be seen that the program gets it right every single time. The distance is always 0, and it always appears number 1:



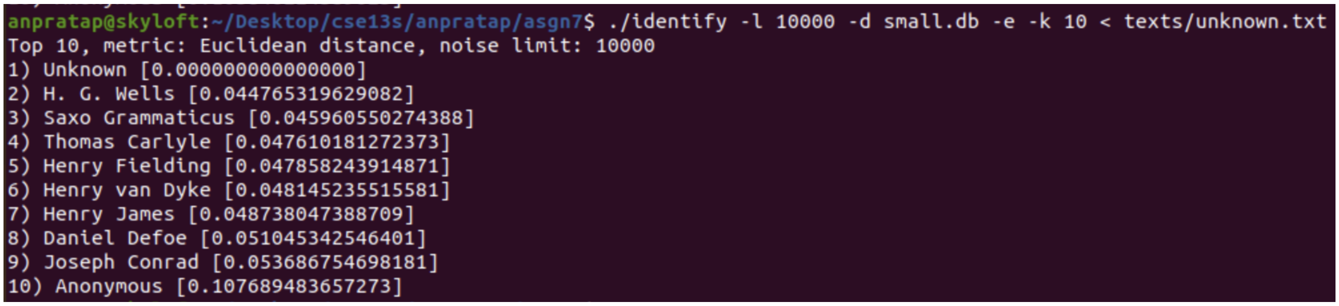
```
anpratap@skyloft:~/Desktop/cse13s/anpratap/asgn7$ ./identify -l 5 -d small.db -e -k 10 < texts/unknown.txt
Top 10, metric: Euclidean distance, noise limit: 5
1) Unknown [0.000000000000000]
2) Thomas Carlyle [0.031970637858713]
3) Henry van Dyke [0.032485643583950]
4) Daniel Defoe [0.032490188961409]
5) H. G. Wells [0.032510533066864]
6) Joseph Conrad [0.032874128208246]
7) Saxo Grammaticus [0.033877113191031]
8) Henry James [0.038428726511538]
9) Henry Fielding [0.038432306570333]
10) Anonymous [0.080915423331728]
```



```
anpratap@skyloft:~/Desktop/cse13s/anpratap/asgn7$ ./identify -l 1000 -d small.db -e -k 10 < texts/unknown.txt
Top 10, metric: Euclidean distance, noise limit: 1000
1) Unknown [0.000000000000000]
2) Saxo Grammaticus [0.035799319327932]
3) Thomas Carlyle [0.036337391077497]
4) Joseph Conrad [0.037614136768242]
5) Henry Fielding [0.037700510079956]
6) H. G. Wells [0.037890331105309]
7) Henry van Dyke [0.037933493726618]
8) Henry James [0.038413920872033]
9) Daniel Defoe [0.039358155848597]
10) Anonymous [0.089488937900779]
```

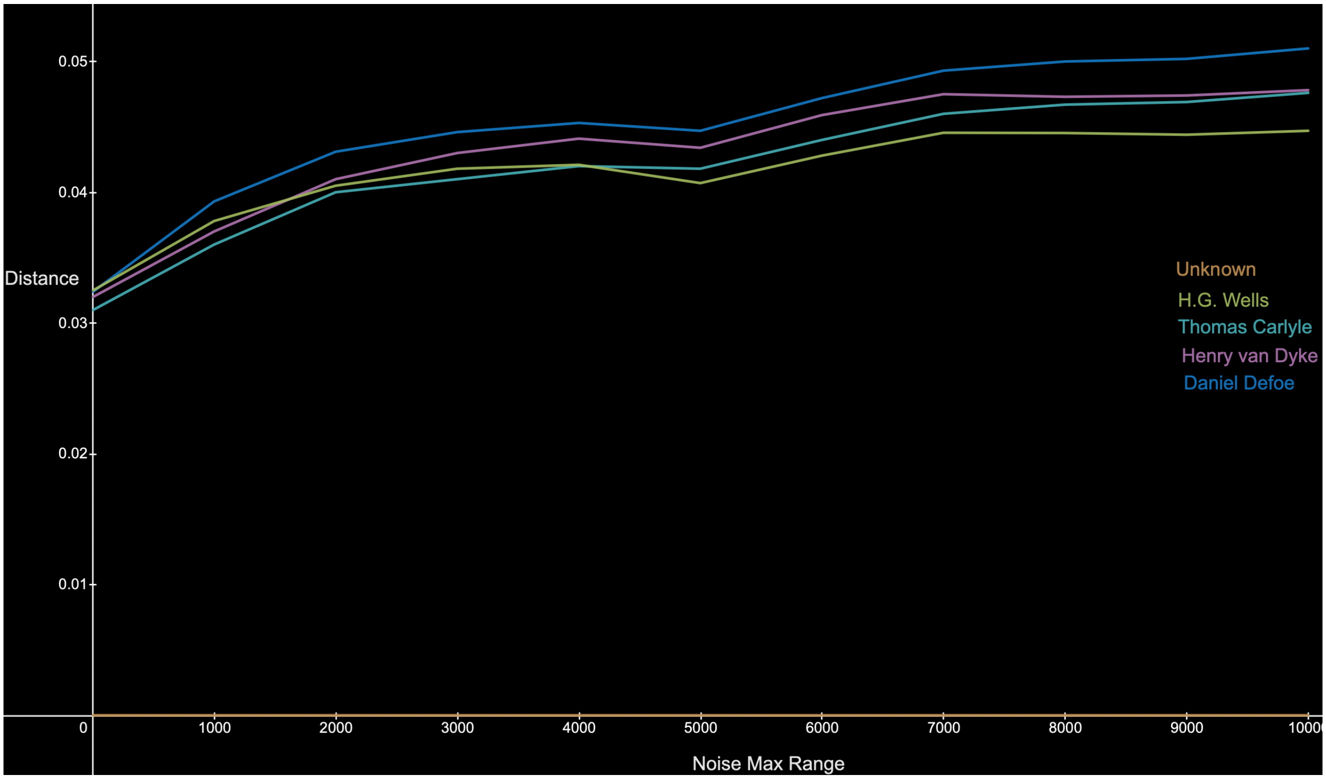


```
anpratap@skyloft:~/Desktop/cse13s/anpratap/asgn7$ ./identify -l 8000 -d small.db -e -k 10 < texts/unknown.txt
Top 10, metric: Euclidean distance, noise limit: 8000
1) Unknown [0.000000000000000]
2) H. G. Wells [0.044538808572048]
3) Saxo Grammaticus [0.045813222457731]
4) Thomas Carlyle [0.046733155663462]
5) Henry van Dyke [0.047320257998367]
6) Henry Fielding [0.047969030009522]
7) Henry James [0.048020904509994]
8) Daniel Defoe [0.050022681710365]
9) Joseph Conrad [0.051495777265962]
10) Anonymous [0.104664707262241]
```



```
anpratap@skyloft:~/Desktop/cse13s/anpratap/asgn7$ ./identify -l 10000 -d small.db -e -k 10 < texts/unknown.txt
Top 10, metric: Euclidean distance, noise limit: 10000
1) Unknown [0.000000000000000]
2) H. G. Wells [0.044765319629082]
3) Saxo Grammaticus [0.045960550274388]
4) Thomas Carlyle [0.047610181272373]
5) Henry Fielding [0.047858243914871]
6) Henry van Dyke [0.048145235515581]
7) Henry James [0.048738047388709]
8) Daniel Defoe [0.051045342546401]
9) Joseph Conrad [0.053686754698181]
10) Anonymous [0.107689483657273]
```

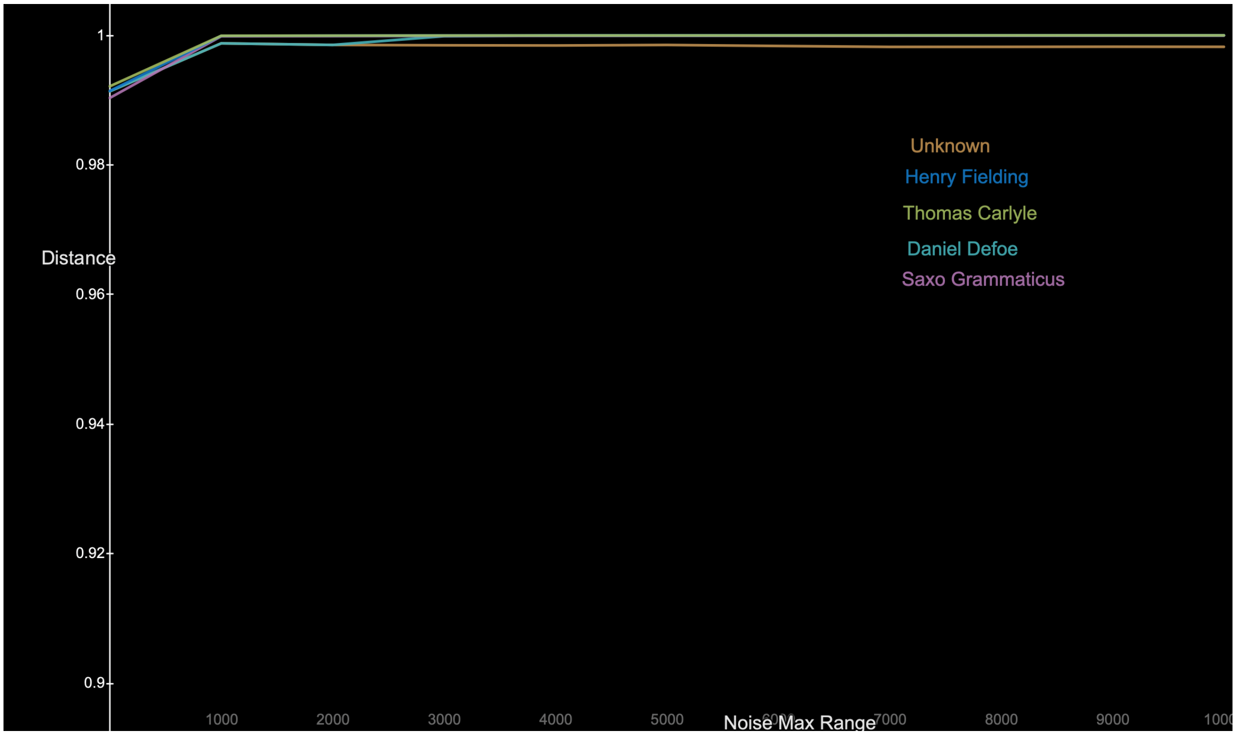
When finally graphing a Noise vs Distance graph, this can be seen:



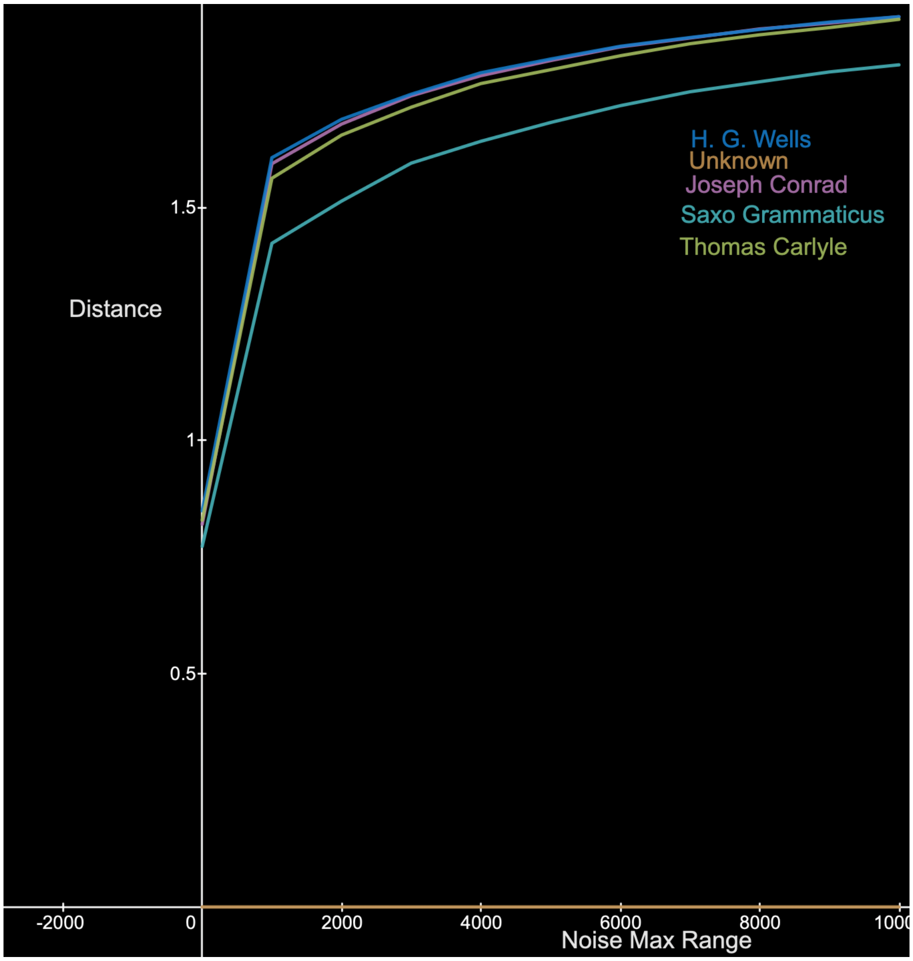
When the noise starts at 5, the top 5 authors are given. One would think that these are the most accurate guesses for this author but as the noise limit increases, interesting things happen. The distance for each likely author begin relatively close to each other. This could possibly be due to the fact that the authors use other noise words that weren't caught in the range. The "uniqueness" of the author is hard to decode if only a few noise words are selected——resulting in a more focused distance. As soon as the noise increases, the authors begin diverging and some more prominently than others. This is due to the fact that when more noise words are used, only the author's unique words are taken into consideration——resulting in more accurate data. Daniel Defoe appeared to be accurate at the start, but he began drifting away since his "unique words" didn't match that of the unknown.txt. H.G Wells, on the other hand, began farther than van Dyke and Carlyle, but seen to be the closest to the unknown text as it shares more common words.

3 Comparing different metrics and tuning noise

Based on the previous graph, it can be seen that the more noise words there are, the more accurate the guess becomes. But how does this same text compare with other distance metrics? Will H.G. Wells always be the best guess for unknown.txt——besides for unknown of course. The cosine graph is of the following:



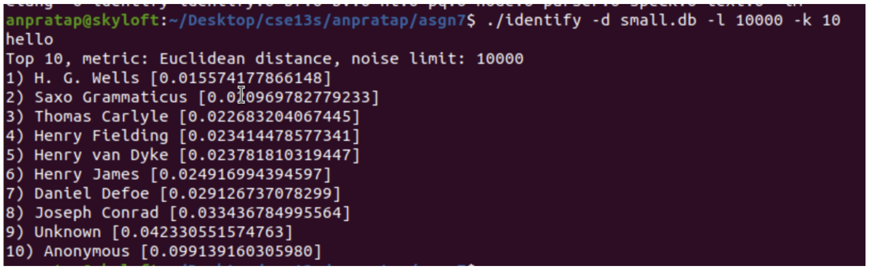
All the guesses seem relatively close together and although I am tracking the top 5 points that appeared with a noise of 5, H. G Wells appears to be missing from the list entirely. How is it that H. G Wells appears using euclidean but not cosine? This could be because this metric system utilizes the idea of angles while euclidean focuses on the "hypotenuse" per se. Unknown is still the least as it should be, but the other authors are very close as they overlap each other on the graph. The Manhattan graph has similar results to the Euclidean formula since Manhattan is a step of Euclidean:



Once again, Unknown has a 0 distance while it seems Saxo Grammaticus is the next closest. When compared to H. G. Wells from the Euclidean metric, it can be seen that metric does have an affect on the final result. Euclidean serves as a Pythagorean theorem while Manhattan just calculates the magnitude difference, resulting in a different answer.

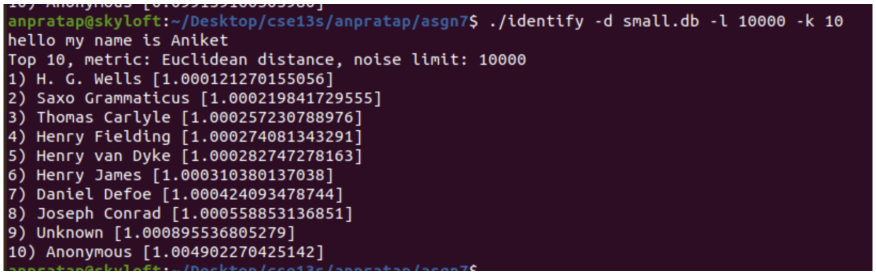
4 text size

How does text size relate to accuracy. Well, the short answer is heavily. The smaller the sample size, the less accurate something is going to be. For example, I wrote the word "hello" in the terminal to see who I would write like and this was the output (using defaults):



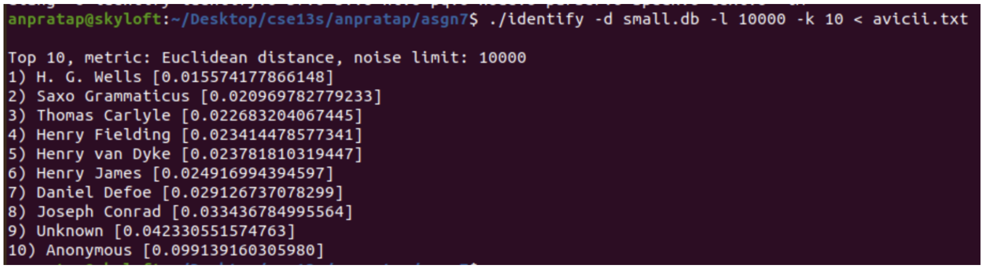
```
anpratap@skyloft:~/Desktop/cse13s/anpratap/asgn7$ ./identify -d small.db -l 10000 -k 10
hello
Top 10, metric: Euclidean distance, noise limit: 10000
1) H. G. Wells [0.015574177866148]
2) Saxo Grammaticus [0.020969782779233]
3) Thomas Carlyle [0.022683204067445]
4) Henry Fielding [0.023414478577341]
5) Henry van Dyke [0.023781810319447]
6) Henry James [0.024916994394597]
7) Daniel Defoe [0.029126737078299]
8) Joseph Conrad [0.033436784995564]
9) Unknown [0.042330551574763]
10) Anonymous [0.099139160305980]
```

Notice how all the distances are very small when compared to the Euclidean data showed previously? This may seem "accurate" but "hello" is in the noise.txt file. Meaning there isn't really mush to go on in order to hone in who wrote the message. Adding a unique word, like my name, shows far weirder results.



```
anpratap@skyloft:~/Desktop/cse13s/anpratap/asgn7$ ./identify -d small.db -l 10000 -k 10
hello my name is Aniket
Top 10, metric: Euclidean distance, noise limit: 10000
1) H. G. Wells [1.000121270155056]
2) Saxo Grammaticus [1.000219841729555]
3) Thomas Carlyle [1.000257230788976]
4) Henry Fielding [1.000274081343291]
5) Henry van Dyke [1.000282747278163]
6) Henry James [1.000310380137038]
7) Daniel Defoe [1.000424093478744]
8) Joseph Conrad [1.000558853136851]
9) Unknown [1.000895536805279]
10) Anonymous [1.004902270425142]
```

Notice how the distances are miles apart——even with the noise set to 10000. The only real word we are checking is Aniket and meaning we still don't have much to go of, further decreasing our accuracy. Now, How about something larger? I tried my program with the lyrics of Avicii's song "Wake me up." This is what it outputted:



```
anpratap@skyloft:~/Desktop/cse13s/anpratap/asgn7$ ./identify -d small.db -l 10000 -k 10 < avicii.txt

Top 10, metric: Euclidean distance, noise limit: 10000
1) H. G. Wells [0.015574177866148]
2) Saxo Grammaticus [0.020969782779233]
3) Thomas Carlyle [0.022683204067445]
4) Henry Fielding [0.023414478577341]
5) Henry van Dyke [0.023781810319447]
6) Henry James [0.024916994394597]
7) Daniel Defoe [0.029126737078299]
8) Joseph Conrad [0.033436784995564]
9) Unknown [0.042330551574763]
10) Anonymous [0.099139160305980]
```

The distances are much smaller now when compared to the previous experiment. Why is this so? Well, the only thing we changed was the text size——the noise was constant. Based on these results, it can be seen that text size of a file determines how accurate something is. More text means closer and more accurate results. This will not however work for files that only contain noise text, like the "hello" I showed earlier.

5 Conclusion

It can be seen that the metric chosen to calculate the distance has a profound effect on the final predicted results. Each metric has a unique results based on their traits. Changing the noise also effects the accuracy because you are filtering out more "useless" words——allowing you to focus more on the unique words that define an author. Finally, the text size also has an influence because you have more words to parse and compare to. The more words you have, the more accurate the data is due to more comparisons. Overall this assignment showed me how Machine Learning works in a broad sense and how certain data inputs can effect outputs.