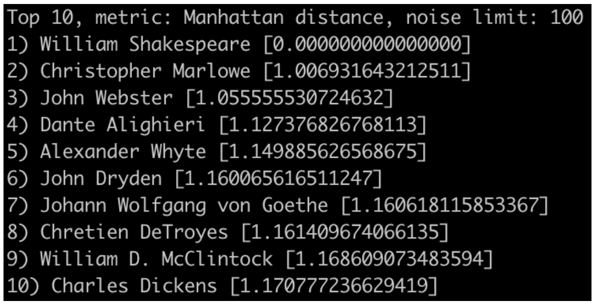
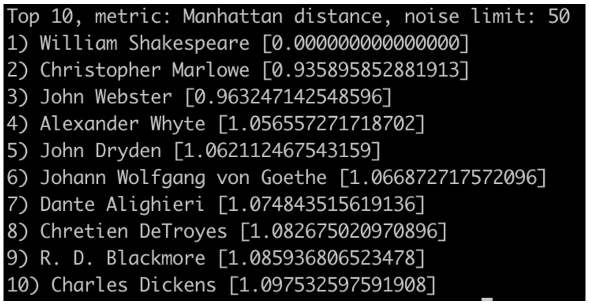
Guillermo Alfaro
CSE13S Winter 2022

Assignment 7: Author Identification

My author identification program works as intended. I am able to properly identify the correct author of a text. If I decrease the number of noise words the program still can identify the correct author. It has does increase the likelihood of other texts but they are still very far off from the original.

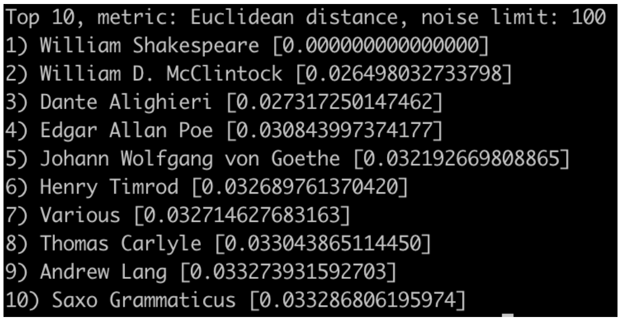


```
Top 10, metric: Manhattan distance, noise limit: 100
1) William Shakespeare [0.000000000000000]
2) Christopher Marlowe [1.006931643212511]
3) John Webster [1.055555530724632]
4) Dante Alighieri [1.127376826768113]
5) Alexander Whyte [1.149885626568675]
6) John Dryden [1.160065616511247]
7) Johann Wolfgang von Goethe [1.160618115853367]
8) Chretien DeTroyes [1.161409674066135]
9) William D. McClintock [1.168609073483594]
10) Charles Dickens [1.170777236629419]
```

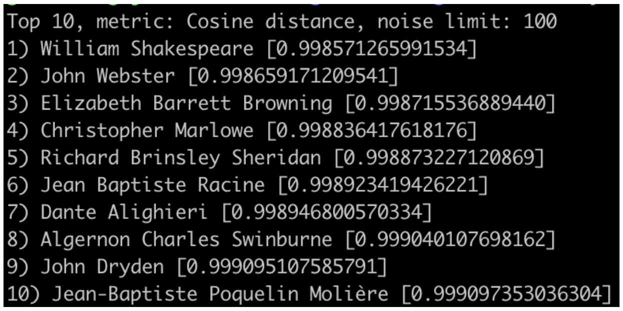


```
Top 10, metric: Manhattan distance, noise limit: 50
1) William Shakespeare [0.000000000000000]
2) Christopher Marlowe [0.935895852881913]
3) John Webster [0.963247142548596]
4) Alexander Whyte [1.056557271718702]
5) John Dryden [1.062112467543159]
6) Johann Wolfgang von Goethe [1.066872717572096]
7) Dante Alighieri [1.074843515619136]
8) Chretien DeTroyes [1.082675020970896]
9) R. D. Blackmore [1.085936806523478]
10) Charles Dickens [1.097532597591908]
```

As for the different KNN methods, Euclidean places other similar texts many times closer to the actual text imputed than Manhattan distance. Cosine does the same as Euclidean.

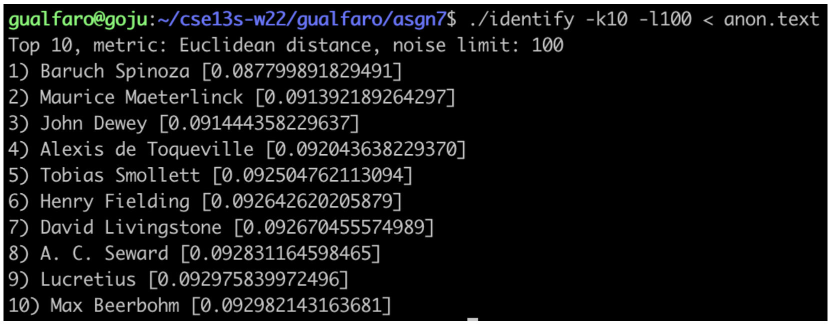


```
Top 10, metric: Euclidean distance, noise limit: 100
1) William Shakespeare [0.000000000000000]
2) William D. McClintock [0.026498032733798]
3) Dante Alighieri [0.027317250147462]
4) Edgar Allan Poe [0.030843997374177]
5) Johann Wolfgang von Goethe [0.032192669808865]
6) Henry Timrod [0.032689761370420]
7) Various [0.032714627683163]
8) Thomas Carlyle [0.033043865114450]
9) Andrew Lang [0.033273931592703]
10) Saxo Grammaticus [0.033286806195974]
```



```
Top 10, metric: Cosine distance, noise limit: 100
1) William Shakespeare [0.998571265991534]
2) John Webster [0.998659171209541]
3) Elizabeth Barrett Browning [0.998715536889440]
4) Christopher Marlowe [0.998836417618176]
5) Richard Brinsley Sheridan [0.998873227120869]
6) Jean Baptiste Racine [0.998923419426221]
7) Dante Alighieri [0.998946800570334]
8) Algernon Charles Swinburne [0.999040107698162]
9) John Dryden [0.999095107585791]
10) Jean-Baptiste Poquelin Molière [0.999097353036304]
```

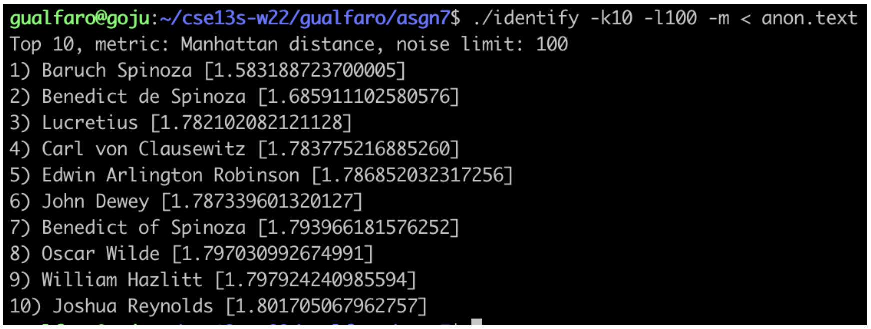
Running different tests with different amounts of text still leads to the correct author identification. I copied three paragrams worth of text from Baruch Spinoza and my program was able to output them as the author.



```
gualfaro@goju:~/cse13s-w22/gualfaro/asgn7$ ./identify -k10 -l100 < anon.text
Top 10, metric: Euclidean distance, noise limit: 100
1) Baruch Spinoza [0.087799891829491]
2) Maurice Maeterlinck [0.091392189264297]
3) John Dewey [0.091444358229637]
4) Alexis de Toqueville [0.092043638229370]
5) Tobias Smollett [0.092504762113094]
6) Henry Fielding [0.092642620205879]
7) David Livingstone [0.092670455574989]
8) A. C. Seward [0.092831164598465]
9) Lucretius [0.092975839972496]
10) Max Beerbohm [0.092982143163681]
```

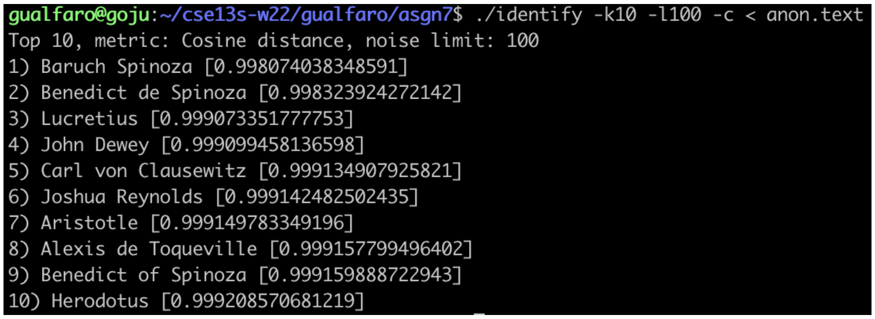
Guillermo Alfaro
CSE13S Winter 2022

Identifying the correct author isn’t that interesting, it’s just doing its job. What is interesting though is the authors that come in after the correct one. They change based on the KNN method used. So far from my testing, I can see that Manhattan has a wider distribution from the correct identification and the second. While using Euclidean the distance between first and second is very much quite smaller.



```
gualfaro@goju:~/cse13s-w22/gualfaro/asgn7$ ./identify -k10 -l100 -m < anon.text
Top 10, metric: Manhattan distance, noise limit: 100
1) Baruch Spinoza [1.583188723700005]
2) Benedict de Spinoza [1.685911102580576]
3) Lucretius [1.782102082121128]
4) Carl von Clausewitz [1.783775216885260]
5) Edwin Arlington Robinson [1.786852032317256]
6) John Dewey [1.787339601320127]
7) Benedict of Spinoza [1.793966181576252]
8) Oscar Wilde [1.797030992674991]
9) William Hazlitt [1.797924240985594]
10) Joshua Reynolds [1.801705067962757]
```

The same can be said about Cosine, it is able to give the correct answer but the other potential answers are very similar in distances while Manhatten has a much wider gap.



```
gualfaro@goju:~/cse13s-w22/gualfaro/asgn7$ ./identify -k10 -l100 -c < anon.text
Top 10, metric: Cosine distance, noise limit: 100
1) Baruch Spinoza [0.998074038348591]
2) Benedict de Spinoza [0.998323924272142]
3) Lucretius [0.999073351777753]
4) John Dewey [0.999099458136598]
5) Carl von Clausewitz [0.999134907925821]
6) Joshua Reynolds [0.999142482502435]
7) Aristotle [0.999149783349196]
8) Alexis de Toqueville [0.999157799496402]
9) Benedict of Spinoza [0.999159888722943]
10) Herodotus [0.999208570681219]
```

Based on the information above I see that Manhatten better identifies the potential authors than Cosine and Euclidean. All three are able to get the right answer but when the amount of text is continuously decreased and the noise words are increased that could change and Manhatten seems the safer method to use.