WRITEUP.pdf
Assignment 2: Numerical Integration

By Santosh Shrestha
CSE 13s Winter 2022
Professor Darrell Long

Introduction:

In this write up we will focus on how the identity programs Manhattan and Euclidean distance formulas are impacted when changing the number of noise words that are filtered out. There are a total of 4 graphs that represent the top 5 authors that have the closest text word patterns to the inputted text of William Shakespeare with the noises of 450 and 50. The program will also use the medium database of texts and authors. My current hypothesis is that author's position won't change that much since the program is comparing the word occurrences causing the distances to be more heavily impacted rather than the authors' position.

Graph 1(Manhattan noise - 450)

Author	Distance(Manhattan)
William Shakespeare	[0.000000000000000]
Christopher Marlowe	[1.289386848510580]
Dante Alighieri	[1.350955977371455]
Charles Dickens	[1.379645015976581]
Various	[1.389629352516162]

Graph 2(Manhattan noise - 50)

Author	Distance(Manhattan)
William Shakespeare	[0.000000000000000]
Christopher Marlowe	[0.983836177579127]
Dante Alighieri	[1.114722020208530]
Charles Dickens	[1.126522499415406]
Various	[1.146450500374275]

Graphs 1 and 2 above quite evidently show that the authors positioning didn't change whatsoever. However, if we compare the Manhattans distance of the authors with a noise of 450 to the one with the noise of 50 we can see that the distances do change quite a bit. The distance increases as the noise filter increases and we are able to see that some of the values either get closer or farther when in the two graphs. For example, Charles Dicken's difference between Various when noise is 50 is 0.01992800095 whereas when it's 450 you get 0.00998433653. For Christopher Marlowe and Dante Alighieri, the difference goes from 0.13088584262 at 50 to 0.06156912886 at 450. This shows a trend that as the noise increased the difference in the authors distances began to decrease. This can be caused due to the cause of the word count of the text being decreased narrowing down the area of difference between the authors'

Graph 4(Euclidean noise - 450)

Author	Distance(Euclidean)
William Shakespeare	[0.000000000000000]
Various	[0.025916875434978]
Thomas Carlyle	[0.026466476408452]
Charles Dickens	[0.027213738410507]
Dante Alighieri	[0.027492301978303]

Graph 3(Euclidean noise - 50)

Author	Distance(Euclidean)
William Shakespeare	[0.000000000000000]
Various	[0.030529997494576]
Dante Alighieri	[0.030613290446841]
Thomas Carlyle	[0.030624648123271]
Edgar Allan Poe	[0.030688141910685]

In the Euclidean distance graphs authors actually swapped around. This might be caused by the fact that the integers are getting decreased which is similar to the trend that we found in the Manhattan Graphs however in the reverse since it is decreasing as the noise increases. All the distance values even if they are different authors are greatly downsized when we increase the noise. This contradicts the idea of how the text word count will make the distance smaller.

Conclusion:

From my analysis of the graphs I see that changing the amount of noise greatly impacts the distance formulas in opposite the exact opposite ways. The Manhattan formula doesn't change the authors' positions but rather creates a larger gap between the authors' distances. However, in the Euclidean distance, the authors do change and the gap between the authors' distances shrink making them closer to the value of 0. This makes it seem that increasing the noise will make the authors' text more susceptible to being similar for the Euclidean distance. The exact opposite is done once again for the Manhattan distance strays away from 0 as we increase the noise, personally, this makes it seem that the Manhattan distance is the better option when comparing texts since you don't want to be able to just increase the noise to get a higher likelihood of the author matching. Overall the Manhattan distance did prove my hypothesis to be correct in that the Authors position wouldn't move that much and that the distance would be more heavily impacted. However, the Euclidean distance formula completely disproved my theory because it did the exact opposite which leads me to believe that all the distance formulas including the Cosine formula are all impacted differently when the noise is altered.