James Yim

CSE13s Winter 2022

Assignment 7 Writeup.pdf

quick note about this writeup
My assignment was incomplete in that I was not able to finish identify.c. Everything else though should work as intended. My writeup will be based on what I expect to happen in identify.c if it were to be finished.

Introduction

This writeup will talk about how as we tune our tools, we will be able to get more accurate results. As we tune out more words that are common between authors, we will be able to get a better understanding of how an author writes due to them being different from each other. We will also talk about how the more data you have, the more accurate your final results will be.

Cleaning out Distractions

People have their own different quirks and things that make them different and that are able to be identifying factors for that person. Texts are the same way. If we have words such as “hath” and phrases such as “wherefore art thou” we can point to those identifying words to describe Shakespeare’s work. If we however look at words such as “the” and “me”, these could be used by anyone and do not do a good job in identifying an author.

As we filter out these common words, it is easier for us to identify the author using what can be called their “signature words”. We immediately recognize the old English to be Shakespeare. We want to focus on these words when we look to identify an unknown piece of text. By doing this we are better able to figure out who wrote a piece of text, as only those authors are using these “signature words”.

You can never have enough

The more you see someone eat mac and cheese the more you can conclude that they love mac and cheese. The same thing can be said about data points. The more data points you have the more confident you can be in your final conclusion. If we have a bigger sample of text, we are able to pick out more of these “signature words” from the text and get a better and stronger conclusion on who authored the text.

As we have bigger samples of text, we are able to get more examples of the “signature words” in the text and thus conclude more confidently the author. One of Shakespeare’s

“signature words” can be “Romeo” from Romeo and Juliet. But, lets say that there was a text with a story about a completely different Romeo from a completely different author. There is a chance that if we just look for “Romeo” we get a high percentage chance of matching this unrelated text to Shakespeare. This is why it is important to have multiple data points. We want to have a confident conclusion and having a bigger sample of text, helps us achieve this.

Different tools for the same job

You can use a magnifying glass to look at an ant, or you can use a microscope to look at an ant, or you can use a macro photography lens to look at an ant. There are different tools that we can use for the same job. In our program we have three different ways of calculating the distance between texts. Some ways are more accurate than others, and some do a just good enough job for our purposes.

To start off this section, I would like to note that the results of my own program may contradict what I say here. The program is incomplete, but I will try my best to explain how I believe it should be. When looking at distances in real life, there are different ways of calculating distances. We can look using triangles, we can look at measuring straight lines, etc. In the program when we use Manhattan, it is like looking at a grid and we have to use the total distance of the path we traveled. This makes it so that the distance using the Manhattan distance is usually larger than the distance actually is. When we use Euclidean, this is more like using triangles to find the distance. We find the hypotenuse of the data points, and use that as our distance. This gets us that straight path we want and is more accurate than the Manhattan method. The cosine distance from what I learned, is the most accurate. It uses the vectors and compares the angle between the two vectors.