Krisha Sharma

CSE 13s

Winter 2022 Long

01/13

Assignment 7: Author Identification WRITEUP

Introduction:

Assignment 7: Author Identification has the overall task of creating a program that attempts to identify the most likely authors for an anonymous sample of text given a large database of texts with known authors. Modern-day stylometry usually is performed using machine learning, achieving high identification accuracies by learning over time, but implementing this from scratch would take an extraordinary effort. Instead during this assignment, we will be using an algorithm that's commonly used in machine learning to identify authors of anonymous samples of text, albeit less accurately.

This write-up will explore the ways in which the distances between authors can be calculated using Manhattan, Euclidean, or Cosine.

Computing Manhattan Distance:

This assignment introduces three possible ways to calculate the distance between authors. The first method to do so is called the Manhattan technique.

This technique requires that the 2 vector element values being compared can be normalized by their respective vector sizes. Should the Manhattan metric be chosen, each one of the two vector elements will be first normalized, and then afterwards subtracted from each other. It is then important to take the absolute value of the result. Once all of the vector elements have

been normalized and computed, the resulting sum of all the computations is the distance between the texts.

Computing Euclidean Distance:

The second method to compute the distance between authors is called the Euclidean technique. This technique, similar to Manhattan, also requires that the 2 vector element values being compared can be normalized by their respective vector sizes.
Overall, calculating the Euclidean distance is very similar to calculating the Manhattan distance. The difference between the two lies in that, instead of simply subtracting the values from each other, it is required to square the result. After all the computations have been calculated and summed up, it's necessary to take the square root of the final result. The resulting value is the distance computed using the euclidean technique.

Computing Cosine Distance:

The third method and final method to compute the distance between authors is called the Cosine technique. Cosine is also computed similar to how Euclidean and Manhattan were computers. The only change is that after the vector element values have been normalized, it is necessary to multiply the values together rather than subtract them. Then, after computing the sum, it is necessary to subtract the sum from 1. This part of the program is implemented in the text.c file, within the function text_dist.