CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)

Saida M. Woolf

Project Number

J0818

Project Title

Using Stylometry Combined with an In-Class Writing Sample to Detect Plagiarism

Abstract

Objectives/Goals

Classroom teachers sometimes put student essays through a plagiarism detector to search the web for any matching sentences and paragraphs. But this doesn#t detect if the paper was written by ghostwriters. I came up with a system where the teacher uses a one-time in-class writing sample and compares it to future essays with stylometry. This will detect unusual changes in style or vocabulary. The writing sample could be an autobiography, which could be quickly written by most students on the first day of class. The purpose of my experiment is to test my system.

Methods/Materials

I did two preliminary experiments to test a JavaScript tool I made to count letters per word and words per sentence to determine authorship. Preliminary Experiment 1 tested two chapters by the same author. Preliminary Experiment 2 tested two chapters by different authors. The success rates for two chapters by the same author was 10% and the success rate for two chapters by different authors was 50%. My conclusion was that counting letters per word and words per sentence were unreliable for determining authorship, so I tried using Most Frequent Word analysis with Stylo, an open-source library for R. While my JavaScript tool had a success rate that was worse than guessing, Stylo succeeded 70% of the time. I decided to use Stylo in my Final Experiments. In my Final Experiments, I wrote an autobiography and compared it with 10 writing samples I wrote in the past using Stylo to perform Most Frequent Word Analysis.

Results

My Final Experiments had a success rate of 40%. Counting words per sentence and letters per word was not accurate for determining authorship. Analyzing most frequent words worked 70% of the time when testing English authors against each other, but worked only 40% of the time when testing on my own writing. My hypothesis was partly incorrect.

Conclusions/Discussion

I came up with a system where the teacher uses a one-time in-class writing sample and compares it to future essays with stylometry. This will detect unusual changes in style or vocabulary that could indicate plagiarism or ghostwriting. It was this system that my experiment was designed to test. Although the stylometry methods that I tested did not give me the desired accuracy, I believe that further testing with other stylometric methods could make this a useful classroom tool.

Summary Statement

I developed and tested a system to detect plagiarism in the classroom setting using an in-class writing sample and stylometry.

Help Received

My father, Reagan Woolf, an aerospace engineer, helped me learn JavaScript, data analysis, statistics, and stylometry.