# CALIFORNIA STATE SCIENCE FAIR
# 2016 PROJECT SUMMARY

**Name(s)**

**Cameron N. Cage**

**Project Number**

36487

**Project Title**

## A Sense of Stylometry

**Abstract**

**Objectives/Goals**

Stylometry, which has been around since at least the 19th century, is the analysis of linguistic style in writing, generally used to determine the author of an unknown piece, but could produce patterns about genre, gender, and relationships between pieces of the same author. With the Internet becoming even more prevalent in our lives, stylometry has growing potential in the field of deanonymizing authors for criminal investigation. My question was whether stylometric analysis is correct more often than random chance in attributing authorship.

**Methods/Materials**

My materials were a self written Python script that automates the procedures running on a computer with Ubuntu Linux. The Python script utilized 34 writing sample sets, each containing one 4,000 word excerpt from C. S. Lewis, Nevil Shute and George Orwell, and another 4,000 excerpt from one of the authors that is #unknown# to the computer. These writing sample sets were generated by another computer program from a selection of out of copyright books.

1. Determine average characters per word in the writing sample.
2.Repeat 1 for each writing sample.
3. Compare each writing samples# average characters per word to the unidentified piece of writing and choose the piece with closest average characters per word to the unidentified piece.
4. Record whether the closest average characters per word to the unidentified piece is correct.
5. Repeat 1 - 4 individually comparing word occurrences, frequency of pronouns, frequency of word pairs, frequency of word triples, frequency of word quadruples and weighted combination metric to the unknown text instead of average characters per word.
6. Repeat 1 - 5 with each writing sample set.

**Results**

In each sample set or trial there were three authors making random chance 1/3. All of the metrics were more effective than random chance. The correct prediction rate of frequency of pronouns, average characters per word, frequency of word quadruples, word occurrences, word triples, word quadruples, and weighted combination was 35.29%, 38.24%, 50.00%, 52.94%, 55.88%, 61.76% and 61.76%, respectively.

**Conclusions/Discussion**

My project provided clear evidence that stylometric analysis is more effective than random chance, given that all seven metrics were more effective than random chance and that my mean correct prediction rate was 50.84% in addition to the fact that my median prediction rate being 52.84%.

**Summary Statement**

I created a program that could attribute authorship from text samples more effectively than random chance.

**Help Received**

None. I designed and programmed my Python script all by myself, with only the Python documentation to assist me.

Preliminary