Ellen Slocum

May 6th, 2014

**Stylometry and Authorship**

***Stylometry***- the statistical analysis of written language, most commonly used to determine authorship

Common Applications:

* Forensic Applications
* Application in music and fine art

Throughout the study of Stylometry, there is a major assumption:

* Authors have a style that…
  + Cannot be consciously manipulated
  + Can be quantified

Stylometry has had trouble gaining acceptance in the literary community because there is currently **no consensus as to a correct methodology or technique.**

Everything changed with **Computational Stylometry.**

Classic Example: Authorship of the Federalist Papers

Fun Facts about the Federalist Papers:

* Written 1787-1788
* Goal: Persuade citizens of NY to ratify the constitution
* Written by Alexander Hamilton, John Jay, and James Madison
* Of the 85 papers written, 12 were **claimed by both Madison and Hamilton**
* The analysis of these papers was originally done by American Statisticians **Mosteller** and **Wallace** in the 1960’s.

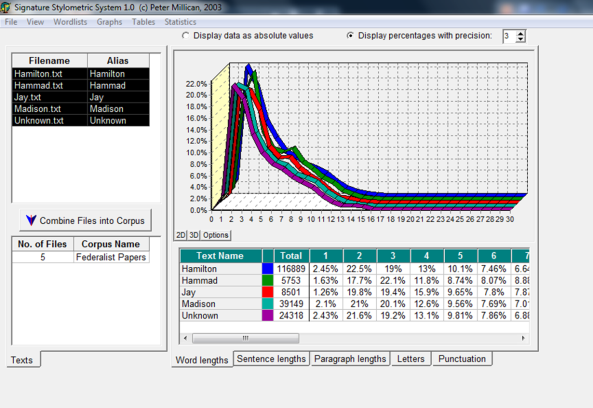
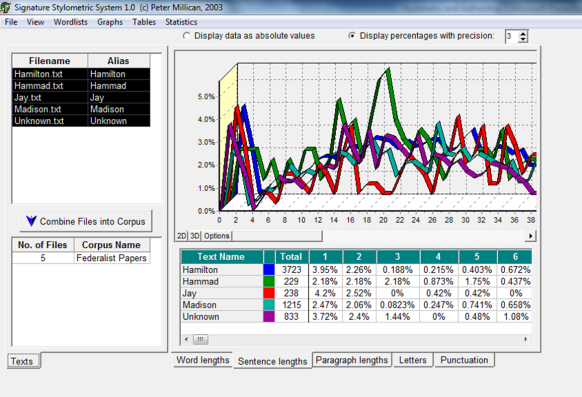
I chose to examine this data using the 2003 **Signature Stylometric System** created by Peter Millican.

**Signature Stylometric System:**

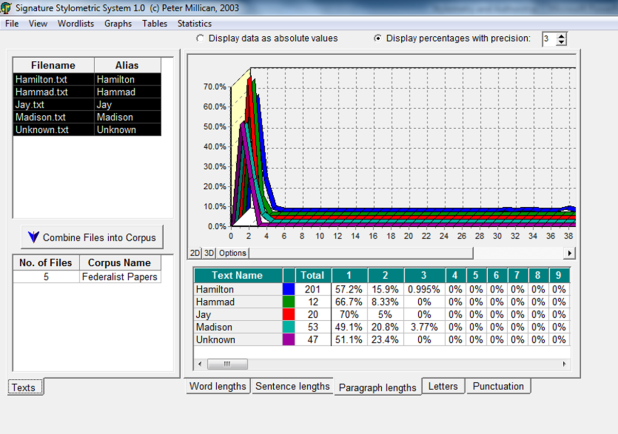
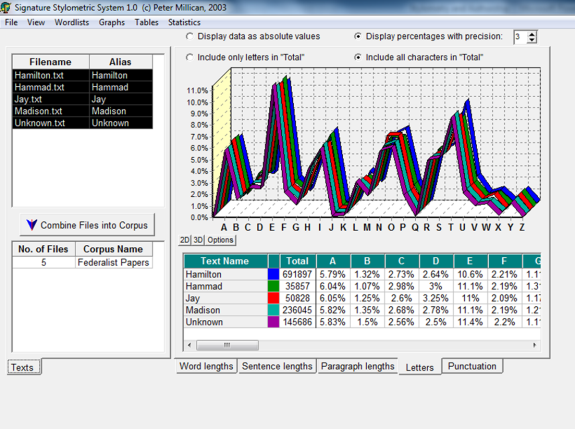
* Emphasizes author identification
* Focuses on…
  + Word lengths
  + Sentence lengths
  + Paragraph lengths
  + Letters
  + Punctuation

***IMPORTANT NOTE: STATISTICS ARE NOT DEFINITE***

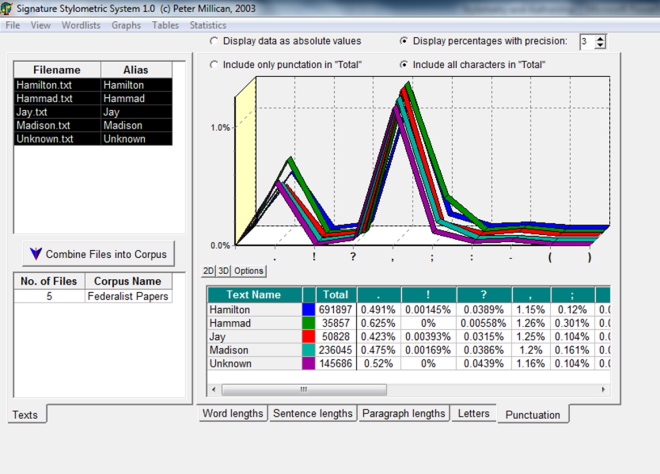
Word Length: Sentence Length:

Paragraph Length: Letters:

Punctuation:



As a result of these studies, it can be concluded that **most likely** James **Madison** is the author of the unknown papers.

Another type of Stylometry…**Visual Stylometry**

***Visual Stylometry***- the statistical analysis of visual arts, most commonly used to authenticate or attribute a piece

Big name in Visual Stylometry: **Daniel Rockmore**

* Professor of Mathematics at Dartmouth College
* New idea in Visual Stylometry: Empirical Mode Decomposition
* Examined the work of 16th century artist Pieter Bruegel the Elder

The **Bruegel Study**:

* Examined 13 landscapes that were at one time attributed to Bruegel
* Created a computer program to analyze pen strokes and characterize Bruegel’s “style”
* Begin with digital scans of each piece
  + Each piece is converted to greyscale (shades from black to white)
  + Central 2000\*2000 pixel square is extracted
  + This square is then divided again into 25, 400\*400 pixel squares
  + A grand total of 325 square samples were then run through the program
* Through this process, ***images are digitized, analyzed, searched, and plotted***🡪 process examines **brushstrokes**
  + 1. Take a 20 mp photo of the drawing or painting
  + 2. Divide image into squares, convert to greyscale, label each square 0-225 (Black-White)
  + 3. Computer program searches for patterns in the horizontal, vertical, an diagonal brushstrokes
    - This is done 7 times
  + 4. Reduce this information into a point on a 3D grid
  + 5. The grid contains the works of many other artists, works by the same artist should cluster together

This method uses ideas contained in Abstract Algebra and Advanced Statistics, through this complex math:

* End result is an algorithm that can “sort” through various images and convert the information into math that can be further analyzed
* Vectors can be created and the information can be analyzed using several ideas in statistics

Statistics involved includes:

|  |  |  |
| --- | --- | --- |
| 1. Mean 2. Standard deviation 3. Skewness | 1. Kurtosis 2. Percentage of outlier pixels greater than the mean plus the standard deviation 3. Mean of the outlier pixels | 1. Standard deviation of the outlier pixels 2. Skewness of the outlier pixels 3. Kurtosis of the outlier pixels |

**Applications in Education:**

* Lesson 1: Box and Whisker charts
  + Grades 9-12
  + Students compare 2 known writing samples and 1 unknown sample to identify the author of the unknown sample using **box and whisker charts** and a graphing calculator.
* Lesson 2: Graphing and Statistics
  + Grades 7-9
  + Students use passages from several different texts and use them to
    - **graph** the distribution of word length
    - **graph** the distribution of sentence length
    - find **average** word length and sentence length
    - **compare graphs** to help identify which passages are from the same author