JUOLA & ASSOCIATES

One Oxford Center
301 Grant Street, Suite 4300
Pittsburgh, Pennsylvania 15219

(412) 254 – 4089
contact@juolaassociates.com

http://www.juolaassociates.com

Stylometric Similarity Report

Generated Automatically by Envelope, An Automated Stylometric Analysis Tool by Juola & Associates

January 14, 2016

Internal Case Identifier: **20160114SAMPLE***

N.B. This is an automatically generated report and is intended for informational purposes only. Documents used to generate this report were provided, without verification or editing, by the end user. These documents, and the report generated by them, have not been examined or vetted by any expert with professional knowledge in the domain of stylometric analysis. This report should not be used as the basis for any legal action or to make any important decisions about the data provided. If you desire additional confirmation of the information found in this report, please contact Juola & Associates to discuss our expert services.

* Please cite this case identifier in any correspondence.

Results

Based on five different automated analyses of the documents presented, we have found

p = 0.0000 - Strong indications of common authorship

Known Document

Documents are processed automatically as presented by the end user. The following document was submitted as the "known document". The Known Document is a document known to have been authored by the suspected, or candidate, author. Below, find a sample of the submitted Known Document:

I am forced into speech because men of science have refused to follow my advice without knowing why. It is altogether against my will that I tell my reasons for opposing this contemplated

Questioned Document

Documents are processed automatically as presented by the end user. The following document was submitted as the "questioned document". The Questioned Document is a document of unknown authorship, suspected to have been authored by the same author as the Known Document. Below, find a sample of the submitted Questioned Document:

Popular imagination, I judge, responded actively to our wireless bulletins of Lake’s start northwestward into regions never trodden by human foot or penetrated by human imagination; though we did not mention his wild hopes of revolutionising the entire sciences of biology and geology.

Possible Contraindications

Envelope is a fully automated system, and is *not* a replacement for a full stylometric analysis performed by an expert. However, Envelope is able to identify several common issues that may occur in the preparation of the data submitted for analysis. For this experiment, Envelope identified the following possible common issues:

None

Background

Envelope is an automated Stylometric Analysis Tool, created by Juola & Associates to allow self-assisted preliminary analysis for Authorship Verification. The Envelope tool produces a preliminary report of the stylistic similarity of two documents. The goal of this analysis is to determine whether it is plausible that the Questioned Document was written by the same author as the Known Document. This tool should be considered only a *preliminary* analysis. There are many factors that can adversely influence this analysis which are not accounted for by the Envelope system. Our expert analysts can provide a forensic-quality analysis if you wish to expand upon this preliminary report.

The theory of authorship attribution is fairly simple. Using language involves a continuous set of choices, many of which are habituated by reason of regional, social, or personal variation. Two documents which share similar stylometric choices are therefore more likely to be by the same author than two documents which do not. By tracking these choices (including, but not limited to, choices in spelling, vocabulary, syntax, and punctuation), Envelope can assess which of a set of potential authors is the most similar (and hence the most likely), and thus make a judgment about the likelihood of any individual candidate author.

Hypothesis

The assumed hypothesis of the Envelope system is that the "Questioned Document" was authored by the same individual (the candidate author) that authored the "Known Document".

Analysis Method

The Questioned Document was assessed along five separate stylistic similarity measures against a standard collection of non-candidate ("distractor") authors, as well as the suspected candidate author (the "Known Document"). If the hypothesis were true (i.e. the "Questioned Document" and "Known Document" were penned by the same author), the Questioned Document should be more "similar" (according to the stylometric similarity measures, mentioned below) to the Known Document than it is to any of the other non-candidate authors' documents. The following analysis methods, curated from Juola & Associates extensive list of "best performing methods", were used by Envelope for this analysis. Please note that a full, expert-assisted stylometric analysis may include consideration of some 1,000,000 (one million) or more potential stylometric analyses in order to identify specific best performing methods for the corpus in question. Here, we have selected only five generally well-performing methods for the automated report:

1. Character N-Grams – Character grams have been shown in many instances to be a useful measure of stylometric similarity. Here, "N" is used as a placeholder for the specific size of these grams (for example, Character 3-grams of the sentence "Hello world" would include {"Hel", "ell", "llo", "lo ", "o w", " wo", "wor", "orl", "rld"}). Character N-grams allow the analysis tool to capture low-level features of a specific author's style without adding unnecessary processing complexity.
2. Word Lengths – Word Lengths have often been suggested as a potential measure of "vocabulary richness" or writing complexity. By itself, word lengths as a feature of stylometric similarity is less useful than many other "best performing" techniques (due to the relatively smaller inter-author variation inherent to the metric). It is, however, quite useful in combination with other metrics, as used in Envelope.
3. Vocabulary Overlap – Another measure of "vocabulary richness" or writing complexity, vocabulary overlap measures are based upon the theory that an author will generally use similar words in any writings. This method is somewhat more sensitive to genre, as genre-specific words may not appear in all of a given author's writings.

4. Function Words – One of the oldest and most reliable stylometric similarity metrics, function words are small words that have little lexical meaning but instead express grammatical relationships (e.g. words like "the", "a", "if", etc.). Many linguists believe that writers have unconscious patterns in their use of function words, which may serve to identify an author even when he attempts to "disguise" his writing style.
5. Punctuation – Like function words, a writer's choice of punctuation is believed to be fairly consistent (and unconscious) among a writer's works. (For example, an author may make excessive use of commas. Or, an author who uses semi-colons at all can generally be depended upon to heavily pepper his writings with the same).

The degree of overall similarity was assessed by calculating the rank sum of the ordered similarity judgments (e.g. the most similar pair for each analysis method was rank 1, the second most similar was rank 2, and so forth), and the significance of the final rank sum was in turn evaluated using a Fisher's Exact test. Fisher's Exact test is a test of statistical significance which provides a probability, *p*, of obtaining the set of observed rank sum values. Based on the results of the Fisher's Exact test, Envelope makes one of several determinations:

$p < 5\%$ - Strong indications of common authorship
$5\% < p < 10\%$ - Indications of common authorship
$10\% < p < 20\%$ - Weak indications of common authorship
$20\% < p < 80\%$ - Analysis Inconclusive
$80\% < p < 90\%$ - Weak indications of different authorship
$90\% < p < 95\%$ - Indications of different authorship
$95\% < p < 100\%$ - Strong indications of different authorship

Juola & Associates

The determination made by Envelope is merely a preliminary analysis. It should be interpreted as an indication of the likely result of a more complete analysis. Juola & Associates is a leading authority in the fields of stylometry and linguistics analysis. We have successfully applied a scientific approach to forensic text analysis, and offer a wide variety of consulting services to provide a complete analysis of the documents you have submitted to Envelope.

Our clients range from billion-dollar multinational corporations to individuals looking for political asylum. When traditional document analysis falls short with modern day text sources like blogs, emails and computer files, we are able to apply our process and deliver results. The services we provide include Authorship Attribution, Authorship Verification (as demonstrated by Envelope), Author Characteristic Profiling, Plagiarism Detection and Large Scale Document Searching. We can develop custom text solutions for your unique case needs and we can testify as an expert witness should the case go to trial and present our findings.

If the results of the Envelope analysis are such that you wish to continue with a further, more in-depth analysis, please contact us for details. We can be reached via telephone at (412) 254-4089 or via email at contact@juolaassociates.com