Determining Authorship with Style

Students: Lauren M. Stuart, Saltanat Tazhibayeva, Ji Hyeon Hong, and Amy Wagoner
Advisors: Julia M. Taylor, Victor Raskin

Authors write texts and publish them, with or without bylines. How do we attribute an anonymous or pseudonymous text to a known author, discriminate between two unknown authors, or verify the authorship of a disputed text? The author may have (perhaps unknowingly) left some clues.

Stylometry is the characterization of a text’s author by capturing the style of writing as measurable features expressed in the text. There are three major applications for stylometry: authorship verification, authorship attribution, and deception detection. In authorship verification\(^1\), the styles of a text and the body of work of the supposed author are matched. In authorship attribution\(^2\), several authors’ works are evaluated for style and compared to the unknown text. Deception detection\(^3\) attempts to find instances in which an author has tried to conceal his or her style\(^4\), perhaps by imitating another author. Most stylometric studies focus on punctuation marks (“,” “!”), and other special characters, sentence and word length, word frequencies, and part-of-speech frequencies. Some look at sequences of words or part-of-speech tags, raw keystroke data, or capitalization.

Common stylometrics
- Total words: 123
- Total sentences: 8
- Average words per sentence: 16
- Word frequencies:
  - (content) author: 4; 0.0325%
  - (function) of: 6; 0.0488%
- Frequent word bigrams:
  - authorship verification
  - Frequent POS bigrams: (CC) (NN)
- Unique words: 83
- Numbers
- Readability measures
- Misspelled words
- Letter n-grams
- Topic words

Example comparison on two features

Research Questions
1. Can we use current tools\(^7\) and techniques on a wider range of corpora, such as non-fiction, Twitter, academic papers?
2. How can we include syntactic and semantic structures as stylometrics?
3. Do some stylometrics vary more between topics\(^8\) than between authors?
4. With current tools, what are the tradeoffs between accuracy and resources such as time, operations, and width of datasets?

References: