Stylometry

- https://ickevald.net/perherngren/wp-content/uploads/2017/06/LitCharts-Analitics-Hemingway-1.pdf
- https://scalar.usc.edu/works/c2c-digital-magazine-fall-winter-2016/a-light-stroll-through-computational-stylometry-and-its-early-potential
- Installing Stylo

Stylometry's claim to fame

The biography accompanying the book presented the author as a “former investigator of the British Royal Military Police." This created an aura of authenticity around the work, hinting at first-hand experience of investigative procedures and intrigue.

Despite its low initial sales, "The Cuckoo’s Calling" was extremely well-received by critics who were enthralled by this fresh voice in crime fiction. Phrases such as “major new talent” and “stellar debut” were used in reviews, indicating high praise and signaling great potential in this hitherto unknown author.

WH - India Knight, claims she was able to smell "a rat" during reading. Bullshit.

Knight: Halfway into a brilliant thriller by a shadowy new novelist, India Knight smelt a rat. **Who was this ex-soldier so acquainted with high fashion**?

Rowling said she was “disappointed” and “angry” that a partner at entertainment firm Russell’s had told his wife’s best friend the real identity of Robert Galbraith, the pseudonym she used for the publication of detective thriller The Cuckoo’s Calling, now soaring up the bestseller list.

Russells, which represents many leading names in music, sport and publishing, apologised unreservedly for the indiscretion which resulted in a Twitter post by a woman called Judith Callegari blowing the lid on publishing’s best kept secret and alerting a Sunday newspaper to the literary scoop of the year.

Ms Callegari it has now emerged is the best friend of the wife of senior lawyer Chris Gossage who revealed the scintillating morsel in a “private conversation” to someone he said he “trusted implicitly”.

Russels Law Firm: It added: “We can confirm that this leak was not part of any marketing plan and that neither JK Rowling, her agent nor publishers were in any way involved.”

Stylometry

By analyzing patterns in word usage, sentence structure, and other linguistic features, stylometry can help determine the author of a text, even when the author's identity has been deliberately obscured.

The field also includes 'stylochronometry', a technique used to estimate the date of a text's creation based on linguistic and stylistic attributes that might be characteristic of a specific period.

Additionally, stylometry can be used in genre studies, where it aids in understanding the nuances that differentiate various text types or genres.

So, how do we define style? One might say it's the "unreasonable effectiveness of counting words near other words". This quirky definition refers to the remarkable power of stylometry to reveal patterns and

connections that are otherwise invisible, simply by analyzing the frequency and proximity of word usage within a text. This seemingly simple act of counting can unearth a wealth of information about a piece of writing.

The concept of a stylome was proposed by Hans van Halteren and his colleagues in their research on computational stylometry 1. A stylome refers to the unique combination of demographic, psychological, and idiosyncratic style properties that make up an individual's idiolect or personal language form. In other words, a stylome is an individual's unique writing style that can be used to identify them as the author of a text. This concept is based on the idea that everyone has their own way of using language, influenced by their personal experiences and background 1.

Federlist Papers

The Federalist Papers is a collection of 85 articles and essays written by Alexander Hamilton, James Madison, and John Jay under the collective pseudonym "Publius" to promote the ratification of the Constitution of the United States.

In the early 1960s American statistician Frederick Mosteller and David L. Wallace conducted what was probably the most influential and widely-publicized early computer-based authorship investigation in an attempt to identify the authorship of the twelve disputed papers in the The Federalist Papers by Alexander Hamilton, James Madison, and John Jay.

Mosteller and Wallace were primarily interested in the statistical methods they employed, but they were able to show that Madison was very likely the author of the disputed papers.

Statistical analysis has been undertaken on several occasions in attempts to accurately identify the author of each individual essay. After examining word choice and writing style, studies generally agree that the disputed essays were written by James Madison.

- It applied statistical methodology to solve one of American history's most notorious questions: the disputed authorship of The Federalist papers.
- It used the once-controversial Bayesian analysis to study frequently-used words in the texts and infer the probability of each author.
- It made the cover of Time magazine and drew the attention of academics and the public alike for its innovative use of mathematics in the humanities.
- It was based on the earlier unpublished work of Frederick Williams and Frederick Mosteller, who pioneered the use of word counts for authorship attribution.

more on the federalist papers

The Federalist Papers (also known simply as the Federalist) are a collection of 85 seminal political theory articles published between October 1787 and May 1788. These papers, written as the debate over the ratification of the Constitution of the United States was raging, presented the case for the system of government that the U.S. ultimately adopted and under which it lives to this day. As such, the Federalist is sometimes described as America's greatest and most lasting contribution to the field of political philosophy.

Three of the Early Republic's most prominent men wrote the papers:

Alexander Hamilton, first Secretary of the Treasury of the United States. James Madison, fourth President of the United States and the man sometimes called the “Father of the Constitution” for his key role at the 1787 Constitutional Convention. John Jay, first Chief Justice of the United States, second governor of the State of New York, and diplomat. However, who wrote which of the papers was a matter of open debate for 150 years, and the co-authors’ behavior is to blame for the mystery.

First, the Federalist was published anonymously under the shared pseudonym “Publius”. Anonymous publication was not uncommon in the eighteenth century, especially in the case of politically sensitive material. However, in the Federalist’s case, the fact that three people shared a single pseudonym makes it difficult to determine who wrote which part of the text. Compounding the problem is the fact that the three authors wrote about closely related topics, at the same time, and using the same cultural and political references, which made their respective vocabularies hard to distinguish from each other.

Second, because Madison and Hamilton left conflicting testimonies regarding their roles in the project. In a famous 1944 article, historian Douglass Adair7 explained that neither man wanted the true authorship of the Papers to become public knowledge during their lifetimes, because they had come to regret some of what they had written. The notoriously vainglorious Hamilton, however, wanted to make sure that posterity would remember him as the driving force behind the Papers. In 1804, two days before he was to fight a duel (in which he was killed), Hamilton wrote a note claiming 63 of the 85 Papers as his own work and gave it to a friend for safekeeping. Ten years later, Madison refuted some of Hamilton’s claims, stating that he was the author of 12 of the papers on Hamilton’s list and that he had done most of the work on three more for which Hamilton claimed equal credit. Since Hamilton was long dead, it was impossible for him to respond to Madison.

Third, because in the words of David Holmes and Richard Forsyth,8 Madison and Hamilton had “unusually similar” writing styles. Frederick Mosteller and Frederick Williams calculated that, in the papers for which authorship is not in doubt, the average lengths of the sentences written by the two men are both uncommonly high and virtually identical: 34.59 and 34.55 words respectively.9 The standard deviations in the lengths of the two men’s sentences are also nearly identical. And as Mosteller quipped, neither man was known to use a short word when a long one would do. Thus, there was no easy way to pinpoint any given paper as clearly marked with Hamilton’s or Madison’s stylistic signature.

It wasn’t until 1964 that Mosteller and David Lee Wallace10, using word usage statistics, came up with a relatively satisfactory solution to the mystery. By comparing how often Madison and Hamilton used common words like may, also, an, his, etc., they concluded that the disputed papers had all been written by Madison. Even in the case of Federalist 55, the paper for which they said that the evidece was the least convincing, Mosteller and Wallace estimated the odds that Madison was the author at 100 to 1.

Since then, the authorship of the Federalist has remained a common test case for machine learning algorithms in the English-speaking world.11 Stylometric analysis has also continued to use the Federalist to refine its methods, for example as a test case while looking for signs of hidden collaborations between multiple authors in a single text.12 Interestingly, some of the results of this research suggest that the answer to the Federalist mystery may not be quite as clear-cut as Mosteller and Wallace thought, and that Hamilton and Madison may have co-written more of the Federalist than we ever suspected.

- 51 papers known to have been written by Alexander Hamilton.
- 14 papers known to have been written by James Madison.

- Four of the five papers known to have been written by John Jay. Three papers that were probably co-written by Madison and Hamilton and for which Madison claimed principal authorship. The 12 papers disputed between Hamilton and Madison. Federalist 64 in a category of its own.

underlying assumptions

The research by Adam Drewnowski and Alice F. Healy entitled “Detection errors on the and and: Evidence for reading units larger than the word” was published in 1977 in the journal Memory & Cognition 1. In this study, the authors conducted five experiments in which subjects read 100-word passages and circled instances of a given target letter, letter group, or word. They found that subjects made a disproportionate number of detection errors on the common function words “the” and “and” 2. The predominance of errors on these two words was reduced for passages in which the words were placed in an inappropriate syntactic context and for passages in which word-group identification was disturbed by the use of mixed typecases or a list, rather than a paragraph, format 2. These results were taken as evidence that familiar word sequences may be read in units larger than the word, probably short syntactic phrases or word frames

“The Secret Life of Pronouns: What Our Words Say About Us” is a book by social psychologist and language expert James W. Pennebaker 1. In this book, Pennebaker uses his groundbreaking research in computational linguistics to show that our language carries secrets about our feelings, our self-concept, and our social intelligence 1. He argues that the smallest, most commonly used, and most forgettable words, such as pronouns, articles, and prepositions, can serve as windows into our thoughts, emotions, and behaviors 2. By analyzing the ways people use these function words, Pennebaker suggests that we can learn about their personality, honesty, social skills, and intentions.

parallell in art history

Giovanni Morelli (1816-1891) was an Italian art critic and historian who is known for developing an innovative method of attributing artwork to specific artists. His approach, often referred to as the "Morellian" method, revolutionized the field of art history.

Before Morelli, the attribution of art often relied on subjective impressions and historical documentation. Morelli argued for a more scientific approach. He proposed that artists, like all people, have unconscious habits that manifest in their work. These habits, he believed, are most evident in the details that artists pay least conscious attention to, such as the depiction of minor characters, ears, hands, or the folds in clothing.

Morelli's method involves closely examining these seemingly insignificant details. He suggested that by comparing these details across different works of art, one could identify patterns or 'signatures' that could help determine the authorship of artworks.

In many ways, Morelli's approach resembles modern forensic methods or even stylometry in literature, where minor details can reveal the identity of an individual. His technique, despite being occasionally criticized, is still in use today, and is recognized as a significant contribution to the discipline of art history. Morelli's method has helped attribute numerous artworks to their rightful creators, resolving many debates in art history.

The desire to transform the authentication process through science—to supplant a subjective eye with objective tools—was not new. During the late nineteenth century, the Italian art critic Giovanni Morelli, dismissing many traditional connoisseurs as “charlatans,” proposed a new “scientific” method based on

“indisputable and practical facts.” Rather than search a painting for its creator’s intangible essence, he argued, connoisseurs should focus on minor details such as fingernails, toes, and earlobes, which an artist tended to render almost unconsciously.