Digging Deeper Reaching Further

Libraries Empowering Users to Mine the HathiTrust Digital Library Resources

Text Analysis Research Examples

"Rowling and Galbraith": an authorial analysis

Patrick Juola, 2013

http://languagelog.ldc.upenn.edu/nll/?p=5315

Research question:
Did J.K. Rowling write *The Cuckoo's Calling* under the pen name Robert Galbraith?

Project description:

- In 2013, scholars Patrick Juola and Peter Millican were approached by a reporter, Cal Flyn (from the *Sunday Times)*, to conduct an authorship analysis to see if the *The Cuckoo's Calling* by Robert Galbraith was actually written by J.K. Rowling. Juola decided to compare *The Cuckoo's Calling* with Rowling's own writing, as well as with samples written by three other authors for comparison. Thus, Flyn provided him with machine-readable texts of *The Cuckoo's Calling,* Rowling's previous novel *The Casual Vacancy,* and three other novels written by British women who specialize in crime fiction.
- Juola conducted a Stylometric analysis to complete the computational comparison of diction. He conducted four analyses focusing on different linguistic variables that indicate style, including the distribution of word lengths, the 100 most common words in the text, the distribution of character 4-grams (groups of four adjacent characters) and word bigrams (pairs of adjacent words).
- Based on the overall patterns of these linguistic variables for each author, the results suggested that Rowling's writing style is the closest to the author of the *The Cuckoo's Calling*. This provided a kind of statistical "proof" of authorial fingerprint, and Rowling admitted to writing the novel under the pen name Galbraith shortly after.

Significant Themes in 19th Century Literature

Matthew Jockers and David Mimno, 2012

http://digitalcommons.unl.edu/englishfacpubs/105

Research question:
What themes are common in 19th century literature?

Project description:

- To explore what themes are common in 19th century literature, Jockers and Mimno applied statistical methods to identify and extract hundreds of "topics" from a corpus of 3,346 works of 19th-century British, Irish, and American fiction collected by the Stanford Literary Lab. Topic modeling is based on the idea that words that co-occur are likely to be about the same thing, so co-occurring words are represented as topics. Their computational analysis yielded 500 topics, which they labeled as "female fashion", "animals and beasts", "eyes and expressions", "machine and industry", and so on. For the female fashion topic, for example, analysis showed that words such as "gown", "silk", "dress", "lace", and "ribbons" tended to co-occur across their corpus of nineteenth century text, so Jockers and Mimno are able to argue through these results that authors from this time period wrote about what women wore.

The Emergence of Literary Diction

Ted Underwood and Jordan Sellers, 2012

http://journalofdigitalhumanities.org/1-2/the-emergence-of-literary-diction-by-ted-underwood-and-jordan-sellers/

Research question:
What textual characteristics constitute 'literary language'?

Project description:

- In their project, Ted Underwood and Jordan Sellers used classification algorithms to study what characteristics constitute "literary language". Their approach was to show the difference in words used in poetry, drama, and fiction with those used in nonfiction to demonstrate how "literary language" developed over time. They used the relative use of newer words to older words to investigate differences, because based on known developments in literature history, older words tend to be more informal and newer words were more learned or literate language. Therefore, the changing ratio of older and newer words over time may reveal patterns in the development of literary language.
- Their dataset consisted of a collection of 4,275 mostly book-length documents. It includes eighteenth-century documents from ECCO-TCP (Eighteenth Century Collections Online), documents in the period 1700-1850 from the Brown Women Writers Project, and a set of nineteenth-century books selected by Sellers.

- In their analysis, first, they trained a computational model to identify literary genres. Then, they compared which words are most frequently used over time in non-fiction prose versus “literary” genres. Their results demonstrated tendency for poetry, drama, and fiction to use older English words. Therefore, this analysis reveals some patterns in the development of literary language.