Determining Authorship of Several Texts
Taylor Smith
Level of interest: 8/10

Project Description:

It is theorized that through a complex analysis process one could determine if two texts were written by the same author, even if the author was trying to write two distinct characters, such as in a fictional work. This is often known as stylometry or "wordprint" studies. Notably, this was used in several studies concerning Book of Mormon authorship[1]. Comparing two texts often requires a deep understanding of language structure by experts. I want to know if I can train an AI model to determine the important features when comparing two texts to determine authorship. If successful, the AI would be able to be given two different texts, and return an accurate prediction regarding if the two texts were written by the same author.

Data Features:

Common stylometry studies include statistics on words and word groups, especially those that are non-contextual. Since this model will be comparing two texts, many of the features will be measuring the difference in word groups between the texts. Each sample will contain the same number of words. Here is an example of what data could be gathered:

Same Author? (Label)	Difference in Number of Unique words	Difference in number of uses of "a/an"	Difference in number of uses of "the"	Difference in Hapax legomena[2]	Difference in Readability Index	etc.
Yes	19	25	17	5	1.2	…
No	58	24	89	78	9.1	…

Gathering the Data:

There are large databases of public domain works including Project Gutenberg, Internet Archive etc.[3] After gathering the texts, we can use a python script to parse through and separate the texts into blocks to use for our data collection. From here we can manually write a script to compare different blocks (NLTK and SpaCy are both very capable natural language toolkits for python) or we can use a stylometric analysis software such as WebSty to generate data[4].

1. https://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=1492&context=jbms
2. https://serhack.me/articles/unveiling-anonymous-author-stylometry-techniques/
3. https://guides.lib.uw.edu/research/openresources/text
4. https://www.clarin.eu/blog/tour-de-clarin-clarin-pl-presents-websty-open-web-based-system-stylometric-analysis