Author Identifier

Discover or validate the author of documents of unproven origin

Description:

The Author Identifier is a document tool that uses decision trees to learn the style of different authors. It can then use what it has learned to make educated guesses about the authorship of document whose author is unknown (or in question). This tool has been written in python using nltk and sklearn packages for parsing and AI learning respectively. The dataset is currently about 7,500 books from the Gutenberg project (https://www.gutenberg.org/) but future releases may incorporate other media such as journals and emails.

Why:

There are several uses that one can find for the author identifier, but I will list the main ones I see:

1.) Plagiarism: the tool could determine the probability that a paper under one author is actually written from another source
2.) Validity of works: the tool could determine the legitimacy of works whose authorship is in question.
3.) Discovering authors hidden under pennames: this tool could be used to help determine the author of books that have been written under a penname
4.) Similar writing styles: the tool could help determine how close different author's writing styles are