Project Proposal: Authorship Attribution

Kareem Ahmed
kareem.ahmed@tum.de
imat: 03658722

December 14, 2014

Background

Authorship attribution is the process of determining the writer of a piece of a given piece of literature, given a collection of documents whose authorship is known. Applications of authorship attribution include plagiarism detection, deducing the writer of inappropriate communications used for harassment, as well as resolving historical questions of unclear or disputed authorship.

Although closely linked to the problem of text classification, this problem differs in that it not only depends upon the textual content but also upon the stylistic attributes of the author. Stylometry - the statistical analysis of literary style - is therefore used in capturing the elusive character of an author by quantifying some of the features of his/her writing.

Motivation

I have always been fascinated by the human subconscious, and its astounding deductive capabilities. An ability, one might argue, is shared by Artificial Neural Networks (ANNs), which are capable of deriving extremely complex outputs of rather simple inputs, to the extent that till this point no one really knows how or why they work so well. Given the nature of the problem, I believe it would not suffice to only reason about explicit features but also implicit ones in order to reach a good solution, an aspect which I believe ANNs excel at. I therefore feel that such a problem would allow me to explore ANNs more vividly, and to greater depths.

Data

My main source for data is Project Gutenberg, which offers over 46,000 free ebooks. I have read that using books as training data requires deploying some techniques such as shallow text analysis, and the like. An alternative would be to crawl the website of a given newspaper, and using the available articles as training data.