Stylometry

= "measuring style"
A history of successes...

- 3 authors (A. Hamilton, J. Madison, J. Jay)
- 85 articles and essays written in 1787-1788, under the pseudonym “Publius”
- frequency of 165 words (mainly functional)

<table>
<thead>
<tr>
<th></th>
<th>enough</th>
<th>while</th>
<th>whilst</th>
<th>upon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hamilton</td>
<td>0.59</td>
<td>0.26</td>
<td>0</td>
<td>2.93</td>
</tr>
<tr>
<td>Madison</td>
<td>0</td>
<td>0</td>
<td>0.47</td>
<td>0.16</td>
</tr>
<tr>
<td>Disputed texts</td>
<td>0</td>
<td>0</td>
<td>0.34</td>
<td>0.08</td>
</tr>
<tr>
<td>Co-authored texts</td>
<td>0.18</td>
<td>0</td>
<td>0.36</td>
<td>0.36</td>
</tr>
</tbody>
</table>

Mosteller & Wallace (1964)
...and Epic Failures

- Andrew Morton in the early ‘60s adapted Cumulative Sum – CUSUM or QSUM (a method which originally was used in the industrial quality control) to be used in texts.
- BBC live show (1993)
  Documents of convicted criminals were attributed to ... the Secretary of State for Justice!!!
"Delta": a Measure of Stylistic Difference and a Guide to Likely Authorship

John Burrows
University of Newcastle, Australia

Abstract
This paper is a companion to my ‘Questions of authorship: attribution and beyond’, in which I sketched a new way of using the relative frequencies of the very common words for comparing written texts and testing their likely authorship. The main emphasis of that paper was not on the new procedure but on the broader consequences of our increasing sophistication in making such comparisons and the increasing (although never absolute) reliability of our inferences about authorship. My present objects, accordingly, are to give a more complete account of the procedure itself: to report the outcome of an extensive set of trials; and to consider the strengths and limitations of the new procedure. The procedure offers a simple but comparatively accurate addition to our current methods of distinguishing the most likely author of texts exceeding about 1,500 words in length. It is of even greater value as a method of reducing the field of likely candidates for texts of as little as 100 words in length. Not unexpectedly, it
Delta Distance

1. e
2. che
3. di
4. la
5. a
6. il
7. non
8. l
9. in
10. per
11. le
12. si
13. con
14. i
15. è
16. un
17. del
18. da
19. più
20. d
21. gli
22. ma

...
\[ \Delta(AB) = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{A_i - \mu_i - B_i - \mu_i}{\sigma_i} \right| \]
Frequent Collocations and Authorial Style

David L. Hoover
New York University, New York, USA

Abstract
This paper examines the effectiveness of multivariate analysis of the frequencies of frequent collocations in characterizing authorial style. Cluster analyses of collocations over various spans, types, and linkages are performed on groups of texts by known authors to determine how well the frequencies of those collocations correctly attribute the texts to their authors and distinguish them from texts by other authors. In each case the results are compared with those based on the frequencies of frequent words and the frequencies of frequent sequences of words. Cluster analyses based on frequent words and sequences ascribe many of the texts to their correct authors. However, analyses based on frequent collocations are more accurate for several groups of texts, sometimes producing more completely correct attributions than analyses based on either words or sequences and sometimes producing the only completely correct attributions. They also produce results for small groups of problematic novels and critical texts extracted from the larger corpora that are often superior to those based on...
Authorship Attribution

Juola vs. Rowling

Musil Project (in Verona)
Juola vs. Rowling
Soldaten, Kameraden!

Network Analysis

not only authorship attribution...
Stylometry with R: A Package for Computational Text Analysis

by Maciej Eder, Jan Rybicki and Mike Kestemont

Abstract  This software paper describes ‘Stylometry with R’ (stylo), a flexible R package for the high-level analysis of writing style in stylometry. Stylometry (computational stylistics) is concerned with the quantitative study of writing style, e.g. authorship verification, an application which has considerable potential in forensic contexts, as well as historical research. In this paper we introduce the possibilities of stylo for computational text analysis, via a number of dummy case studies from English and French literature. We demonstrate how the package is particularly useful in the exploratory statistical analysis of texts, e.g. with respect to authorial writing style. Because stylo provides an attractive graphical user interface for high-level exploratory analyses, it is especially suited for an audience of novices, without programming skills (e.g. from the Digital Humanities). More experienced users can benefit from our implementation of a series of standard pipelines for text processing, as well as a number of similarity metrics.

Introduction

Authorship is a topic which continues to attract considerable attention with the larger public. This claim is well illustrated by a number of high-profile case studies that have recently made headlines across the popular media, such as the attribution of a pseudonymously published work to acclaimed
Consensus Trees

Majority rule consensus

Numbers indicate frequency of clades in the fundamental trees

MAJORITY-RULE CONSENSUS TREE
Visualization in stylometry: Cluster analysis using networks

Maciej Eder
Pedagogical University of Kraków, Poland
Institute of Polish Language, PAS

Abstract
The aim of this article is to discuss reliability issues of a few visual techniques used in stylometry, and to introduce a new method that enhances the explanatory power of visualization with a procedure of validation inspired by advanced statistical methods. A promising way of extending cluster analysis dendograms with a self-validating procedure involves producing numerous particular ‘snapshots’, or dendograms produced using different input parameters, and combining them all into the form of a consensus tree. Significantly better results, however, can be obtained using a new visualization technique, which combines the idea of nearest neighborhood derived from cluster analysis, the idea of hammering out a clustering consensus from bootstrap consensus trees, with the idea of mapping textual similarities onto a form of a network. Additionally, network analysis seems to be a good solution for large data sets.

1 Introduction
Most of the computational methods used in stylometry have been originally introduced to solve authorship attribution problems. This fact had an algorithms, suitable for classification tasks, derived mostly from the field of biometrics, nuclear physics, or software engineering, that could be easily adopted to authorship attribution. They include naïve Bayes classification, support vector machines,
Fig. 6. Two algorithms of mapping textual relations: establishing weighted links to a nearest neighbor and two runners-up (top); producing a consensus network (bottom).