DOWNLOAD PDF 7. STYLOMETRY AND ATTRIBUTION STUDIES

Chapter 1 : Stylometry | Revolvy

This chapter examines various procedures for authorship attribution. In particular, it looks at how quantitative analyses based on counts can be used. Such stylometric analyses can serve two main purposes, both of which would otherwise be very difficult to tackle in any systematic way.

Tweet Of all the techniques that could be used to study ancient texts, there are a few that stand out as being both very important and largely understudied, being either ignored in practice or taken on faith due to the lack of relevant expertise or accessible tools. The ones that come to my mind right now are these: The other two mentioned may be envious of such wide respect and acceptance. Of the three, perhaps the most confusion surrounds these techniques, and a large part of it is due to the confusion and unresolved questions that still persist among the experts. I personally find this field of study fascinating and can summarize, roughly, some of the material that I have read. It turns out that this is still pretty useful even today as a description of several phenomena. It has something to do with all those hapax legomena words appearing once , dislegomena, trilegomena, and so on, and what proportion of each tend to fall in each category. They worked on the basis of word frequencies, a sort of Bayesian reasoning, and a choice between two authors The Federalist Papers. Morton appears on the stage and sets the tempo of the discussion for a while. The idea of using sets at least 50 strong of common high-frequency words and conducting what is essentially a principal components analysis on the data has been developed by Burrows and represents a landmark in the development of stylometry. Large, freely-available electronic corpora and cheap, very-powerful personal computers combine to quicken the pace of research considerably. What if we also looked at bigrams two characters in a row and trigrams three characters or other n-grams? A wide variety of mathematical techniques are employed today, and studies frequently focus on the question of which method of combining all the individual measurements provides the best final results. Most formal statistical analysis operates fairly well with adequately-sized expected populations. Even just five or so expected occurrences of something in a sample can work well with standard statistical measures and techniques. This problem is especially relevant when attempting to innovate with techniques that have reliability even with smaller sample sizes. This is often considered a difficult problem, and it was especially common to avoid broaching the issue in earlier research. But there is certainly research being done on the question of how best to determine whether the author of the tested sample text belongs to some known candidates or likely belongs to none of them. Related to this problem are all the purposes for which stylometry is used other than authorship attribution. In the domain of authorship attribution, the question is often posed: In every specific example, the question is basically this: What can be used to discriminate between many potential authors or options and pick the right one? These three issues are closely related to three general goals: In practice, achieving these goals is a compromise among them. Improvements in method often focus on making the tradeoffs involved in this compromise less severe. I will admit that I can struggle with it. There is absolutely no reason not to treat the claim that a stylometric method is reliable as a hypothesis, one which can be tested and possibly established only by an assessment of its performance with real-world data where the conclusions are already known by other means. Failing to perform any experiment of such a kind, before applying the method to a controversial or unknown case, is pure pseudoscience. The important thing is consistency. All the authors being tested as possible candidates for authorship of the text being tested should be relatively extensive, enough to get an idea of the range of their stylistic variation. The flatter graph will have more area under the curve at the extremes, while the narrower graph will have more under the curve at the center. We want these numbers for the next step, where we compare the observed frequency in the test sample of each feature to the mean and standard deviation of the feature in the individual candidates. For the purposes of the program, there will be a computed value for every candidate, feature pair that says how close that candidate is to the sample on the basis of the feature. Different candidates have different curves because they have different means and standard deviations, as computed above. On this basis, we can differentiate between the candidates. Step 5 â€" Magic Happens Or principal component analysis. Or something like that. I wanted to be aware of how likely it is that none of the candidates wrote the sample text.

So I used this one simple trick. In addition to the regular list of real candidates, I also used a second list of phony candidates. The theory is, if some more-or-less random author off the street walks in and starts providing better matches for the sample text than any of the so-called real candidates, maybe none of the guys on the initial list wrote it in the first place. So the conclusion in that case is that it seems like we have no idea who wrote it, but likely none of the candidates did. Not perfect, of course, but enough to know that the method was really getting at something that pointed to the likely authorship of texts. Let the results speak for themselves. I got a lot of valuable feedback. Both are attributed to Athenagoras of Athens. Both of the later samples from the work matched most closely to the text of the first book of it. Third Case â€" Clement of Alexandria This was much more detailed than either of the previous two. The post is here. Why Not Test the Letter to Theodore? I had the opportunity to explain a little bit about why short samples are a problem with this particular method. The root of the problem is the Poisson distribution. The frequency of a given word in a sample from an author of length N randomly occurring with some chance, call it X is actually better represented by a Poisson distribution. So the normal distribution can approximate the Poisson distribution with higher expected averages for the observed frequency of words in samples. Wikipedia has a chart: The small, useless, lumpy, scrunched-up-against-the-left Poisson distribution is shown in orange. The technique is about thirty years old. This is not set in stone. Fourth Case â€" Origen Origen provided a splendid arena in which to showcase the capabilities of the algorithm, including the ability to identify potentially dubious works, without knowing who the real author might be. All the work and all the results are documented, in extensive detail, on the forum. The results in summary, not including those texts not tested or too short to say anything meaningful.

DOWNLOAD PDF 7. STYLOMETRY AND ATTRIBUTION STUDIES

Chapter 2 : Project MUSE - Stylometry for Medieval Authorship Studies: An Application to Rhyme Words

PDF | On Jan 1, , A Brinkman and others published Musical Stylometry, Machine Learning, and Attribution Studies: A Semi-Supervised Approach to the Works of Josquin.

Endnotes Introduction Stylometry is the quantitative study of literary style through computational distant reading methods. It is based on the observation that authors tend to write in relatively consistent, recognizable and unique ways. Each person has their own unique vocabulary, sometimes rich, sometimes limited. Although a larger vocabulary is usually associated with literary quality, this is not always the case. Ernest Hemingway is famous for using a surprisingly small number of different words in his writing, 1 which did not prevent him from winning the Nobel Prize for Literature in Some people write in short sentences, while others prefer long blocks of text consisting of many clauses. No two people use semicolons, em-dashes, and other forms of punctuation in the exact same way. The ways in which writers use small function words , such as articles, prepositions and conjunctions, has proven particularly telling. It is also very hard for a would-be forger to copy. Function words have also been identified as important markers of literary genre and of chronology. Scholars have used stylometry as a tool to study a variety of cultural questions. For example, a considerable amount of research has studied the differences between the ways in which men and women write 3 or are written about. This is what we will be doing in this lesson, using as our test case perhaps the most famous instance of disputed authorship in political writing history, that of the Federalist Papers. Learning Outcomes At the end of this lesson, we will have examined the following topics: How to apply several stylometric methods to infer authorship of an anonymous text or set of texts. How to use relatively advanced data structures, including dictionaries of strings and dictionaries of dictionaries, in Python. Prior Reading If you do not have experience with the Python programming language or are finding examples in this tutorial difficult, the author recommends you read the lessons on Working with Text Files in Python and Manipulating Strings in Python. Please note, that those lessons were written in Python version 2 whereas this one uses Python version 3. The differences in syntax between the two versions of the language can be subtle. If you are confused at any time, follow the examples as written in this lesson and use the other lessons as background material. More precisely, the code in this tutorial was written using Python 3. Required materials This tutorial uses both datasets and software that you will have to download and install. The Dataset To work through this lesson, you will need to download and unzip the archive of the Federalist Papers. The archive also contains the original Project Gutenberg ebook version of the Federalist Papers from which these 85 documents have been extracted. When you unzip the archive, it will create a directory called data. This will be your working directory and all work should be saved here while completing the lesson. The Software This lesson uses the following Python language versions and libraries: Should you encounter error messages such as: This is easiest to accomplish using the pip command. Full details are available via the Programming Historian lesson on Installing Python modules with pip. Some Notes about Language Independence This tutorial applies stylometric analysis to a set of English-language texts using a Python library called nltk. Much of the functionality provided by the nltk works with other languages. As long as a language provides a clear way to distinguish word boundaries within a word, nltk should perform well. Languages such as Chinese for which there is no clear distinction between word boundaries may be problematic. I have used nltk with French texts without any trouble; other languages that use diacritics , such as Spanish and German, should also work well with nltk. Only one of the tasks in this tutorial requires language-dependent code. This will be explained in the tutorial. Finally, note that some linguistic tasks, such as part-of-speech tagging , may not be supported by nltk in languages other than English. This tutorial does not cover part-of-speech tagging. Should you need it for your own projects, please refer to the nltk documentation for advice. The Federalist Papers - Historical Context The Federalist Papers also known simply as the Federalist are a collection of 85 seminal political theory articles published between October and May These papers, written as the debate over the ratification of the Constitution of the United States was raging, presented the case for the system of government that the U. Anonymous publication was not uncommon in the eighteenth century, especially in the case of politically

sensitive material. Compounding the problem is the fact that the three authors wrote about closely related topics, at the same time, and using the same cultural and political references, which made their respective vocabularies hard to distinguish from each other. Second, because Madison and Hamilton left conflicting testimonies regarding their roles in the project. In a famous article, historian Douglass Adair 7 explained that neither man wanted the true authorship of the Papers to become public knowledge during their lifetimes, because they had come to regret some of what they had written. The notoriously vainglorious Hamilton, however, wanted to make sure that posterity would remember him as the driving force behind the Papers. In , two days before he was to fight a duel in which he was killed , Hamilton wrote a note claiming 63 of the 85 Papers as his own work and gave it to a friend for safekeeping. Since Hamilton was long dead, it was impossible for him to respond to Madison. Frederick Mosteller and Frederick Williams calculated that, in the papers for which authorship is not in doubt, the average lengths of the sentences written by the two men are both uncommonly high and virtually identical: And as Mosteller quipped, neither man was known to use a short word when a long one would do. By comparing how often Madison and Hamilton used common words like may, also, an, his, etc. Even in the case of Federalist 55, the paper for which they said that the evidece was the least convincing, Mosteller and Wallace estimated the odds that Madison was the author at to 1. Since then, the authorship of the Federalist has remained a common test case for machine learning algorithms in the English-speaking world. Our Test Cases In this lesson, we will use the Federalist as a case study to demonstrate three different stylometric approaches. The 51 papers known to have been written by Alexander Hamilton. The 14 papers known to have been written by James Madison. Four of the five papers known to have been written by John Jay. Three papers that were probably co-written by Madison and Hamilton and for which Madison claimed principal authorship. The 12 papers disputed between Hamilton and Madison. Federalist 64 in a category of its own. The one exception is Federalist 64, which everyone agrees was written by John Jay but which we keep in a separate category for reasons that will become clear later. Our first two tests, using T. Preparing the Data for Analysis Before we can proceed with stylometric analysis, we need to load the files containing all 85 papers into convenient data structures in computer memory. The first step in this process is to assign each of the 85 papers to the proper set. The dictionary is a data type made up of an arbitrary number of key-value pairs; in this case, the names of authors will serve as keys, while the lists of paper numbers will be the values associated with these keys. For example, we can access a value by indexing the dictionary with one of its keys, we can scan the entire dictionary by looping over its list of keys, etc. We will make ample use of this functionality as we move along. This will be stored as a string. Open your chosen Python development environment. If you do not know how to do this, you should read Setting up an Integrated Development Environment Mac , Linux , Windows before continuing. How you do this depends on your Python development environment. Therefore, we should not treat the characteristic curves as a particularly trustworthy source of stylometric evidence. However, Mendenhall published his theory over one hundred and thirty years ago and made all calculations by hand. It is understandable that he would have chosen to work with a statistic that, however coarse, was at least easy to compile. The first line in the code snippet above loads the Natural Language Toolkit module nltk , which contains an enormous number of useful functions and resources for text processing. The next few lines set up data structures that will be filled by the block of code within the for loop. It plots a graph of the distribution of word lengths in the corpus, for all words up to length If you want to tokenize texts in another language, you will need to change one line in the code above to feed the proper language to the tokenizer as a parameter. The results should look like this: This is consistent with the historical observation that Madison and Hamilton had similar styles, but it does not help us much with our authorship attribution task. This is not what we are after here. The more similar the vocabularies, the likelier it is that the same author wrote the texts in both sets. Here is how to apply the statistic for authorship attribution: Take the corpora associated with two authors. Merge them into a single, larger corpus. Count the tokens for each of the words that can be found in this larger corpus. Select the n most common words in the larger corpus. Calculate how many tokens of these n most common words we would have expected to find in each of the two original corpora if they had come from the same author. Equation for the chi-squared statistic. The smaller the chi-squared value, the more similar the two corpora. Therefore, we will calculate a chi-squared for

the difference between the Madison and Disputed corpora, and another for the difference between the Hamilton and Disputed corpora; the smaller value will indicate which of Madison and Hamilton is the most similar to Disputed. No matter which stylometric method we use, the choice of n, the number of words to take into consideration, is something of a dark art. In the literature surveyed by Stamatatos 2 , scholars have suggested between and 1, of the most common words; one project even used every word that appeared in the corpus at least twice. As a guideline, the larger the corpus, the larger the number of words that can be used as features without running the risk of giving undue importance to a word that occurs only a handful of times. In this lesson, we will use a relatively large n for the chi-squared method and a smaller one for the next method. Changing the value of n will certainly change the numeric results a little; however, if a small modification of n causes a change in authorship attribution, this is a sign that the test you are performing is unable to provide meaningful evidence regarding your test case. Who are the authors we are analyzing? How often do we really see this common word?

DOWNLOAD PDF 7. STYLOMETRY AND ATTRIBUTION STUDIES

Chapter 3 : Satoshi Nakamoto unmasked? UK based non-profit claims, stylometry is proof

*Musical Stylometry, Machine Learning, and Attribution Studies: A Semi-Supervised Approach to the Works of Josquin Andrew Brinkman*1 Daniel Shanahan, *2 Craig Sapp#3 *School of Music, Louisiana State University, Baton Rouge, Louisiana, USA.*

University of the West of England david. This session proposes to look at the history of the field, identify many of the more major problems, and offer some solutions that will go a long way towards giving the field credibility and validity. Every area of authorship attribution studies has this problem -- research, experimental set-up, linguistic methods, statistical methods It seems that for every paper announcing an authorship attribution method that "works" or a variation of one of these methods, there is a counter paper pointing out crucial flaws: This widespread disagreement has not only kept authorship attribution studies out of most United States court proceedings, but it also threatens to undermine even the legitimate studies in the court of public and professional opinion. The time has come to sit back, review, digest, and then present a theoretical framework to guide future authorship attribution studies. The first paper, by David Holmes, will give the necessary history, scope, and present direction of authorship attribution studies with particular emphasis on recent trends. The second paper, by Harald Baayen and Fiona Tweedie, will focus on one problem: The third paper, by Joseph Rudman, will point out some of the problems that are keeping authorship attribution studies from being universally accepted and will offer suggestions on how these problems can be overcome. Its Origins, Development and Aspirations. Holmes Introduction This paper is the opening paper in the session on stylometry and aims to review the historical development of stylometry up to and including its current standing as a statistical tool within the humanities. Most stylometric studies employ items of language and most of these items are lexically based. A sound exposition of the rationale behind such studies has been provided by Laan The main assumption underlying stylometric studies is that authors have an unconscious as well as a conscious aspect to their style. The two primary applications are attributional studies and chronological problems, yet a difference in date or author is not the only possible explanation for stylistic peculiarities. Variation in style can be caused by differences of genre or content, and similarity by literary processes such as imitation. Word-length and sentence-length The origins of stylometry may be traced back to the work of Mendenhall on word-lengths and the idea of counting features of a text was extended by Yule to include sentence-lengths. Morton used sentence-lengths for tests of authorship of Greek prose, but we now know that neither of these measures are wholly reliable indicators of authorship. Function words Word-usage offers a great many opportunities for discrimination. Some words vary considerably in their rate of use from one work to another by the same author, others show remarkable stability within an author. Morton developed techniques of studying the position and immediate context of individual word-occurrences but his method has, however, come under much criticism and Smith has demonstrated that it cannot reliably distinguish between the works of Elizabethan and Jacobean playwrights. The idea of using sets at least 50 strong of common high-frequency words and conducting what is essentially a principal components analysis on the data has been developed by Burrows and represents a landmark in the development of stylometry. Examples of the technique will be displayed. The best fitting model appears to be that attributed to Sichel and this paper will cover the Sichel model in addition to looking at the behaviour of the once-occurring words hapax legomena and twice-occurring words hapax dislegomena as useful stylometric tools. Content analysis Content analysis refers to tabulating the frequency of types of words in a text, the aim being to reach the denotative or connotative meaning of the text. Neural networks Stylometry is essentially a case of pattern recognition. Neural networks have the ability to recognise the underlying organisation of data which is of vital importance for any pattern recognition problem, so their application in stylometry is both inevitable and welcome. The results achieved by Merriam and Matthews and by Lowe and Matthews will be discussed. Automated feature finders will be developed Forsyth and Holmes, to let the computer take over the task of finding the features that best discriminate between two candidate authors for a disputed text. There will be theoretical advances too, as in the change from lexically based techniques to syntactic annotation proposed by Baayen, Van

Halteren and Tweedie Stylometry, though, presents no threat to traditional scholarship. In the context of authorship attribution, stylometric evidence must be weighed in the balance along with that provided by more conventional studies made by literary scholars. Harald Baayen Introduction Various measures of lexical richness have been employed in stylometry and authorship attribution see, e. These measures have been advanced as characteristic constants whose value is not influenced by the text size. This study investigates in detail to what extent these measures are truly constant, how well they are suited for discriminating authors, and to what extent the values assumed by these measures are influenced by discourse structure see Baayen, Text constants have been developed because the simplest measure of lexical richness, the vocabulary size V N , varies with the number of tokens in the text, N. In order to remove this dependency, constants have been proposed that are supposed to be independent of N. Combinations of these constants have been used to investigate problems of authorship see for example Holmes, and Baayen et al, The latter discriminated two authors at lexical and syntactic levels using analyses of function words and lexical richness. They found that function words performed better than the constants at both levels, and that the inclusion of syntactic information improved the discrimination. Nevertheless, the constants also tap into stylistic properties of texts at a fairly abstract level. In order to properly evaluate the discriminatory potential of the text constants, we must clarify whether and how effectively they capture similarities and differences between authors, and to what extent they are truly constant. Validity - Are the Constants Constant? Measurements have been taken at 20 equally-spaced points in the text. One hundred such randomisations were carried out.. Crucially, only the mean values for K indicate that its value is theoretically truly constant for randomised text; those for W and H increase and decrease with text size, while S rises then decreases. It is clear from these graphs, both of the actual and randomised texts, that far from being stable, the constants are as variable as V, the variable that they were intended to replace. In sum, with the exception of K and possibly the Zipf size, constants are not constant in theory, and, without exception, none are constant in practice. The empirical values of the constants are co-determined by the way in which the randomness assumption is violated in running text, namely by coherence in lexical use at the discourse level see Baayen, Developmental Profiles Thus far we have considered a single text. It is possible that the variability we have observed is very small when compared to other texts and that discrimination is still possible between authors. In order to investigate this we analysed a total of fifteen texts, detailed in Table 1. Examining the first plot, it can be seen that, while the value of W varies with N, texts by the same author vary in the same way; the Carroll texts are coincident, as are the James texts and two of the Conan Doyle texts. It is also clear that this is not necessarily the case; the Baum texts are widely separated, as is the third Conan Doyle text from the other pair. A similar structure is found in the graph of H, with slightly different orderings. Turning to S, however, we find that the constant is so variable that it is impossible to separate authors, even at larger text sizes. The plot of K again yields a pattern in which texts are fairly well separated. The different ordering of the texts in this graph indicates that K is measuring a different facet of the lexical structure of these texts. The Conan Doyle texts now group together, as do the Baum texts, but now the Carroll texts diverge. We have calculated the values for fifteen lexical richness constants and found that the resulting profiles could be classified into four families, exemplified in the graphs above. The largest family of constants is that to which W belongs. S comprises the family of constants that are of no discriminatory value. K makes up a family with D, variables that are theoretically constant given the urn model of word distribution within text. Some texts that are separated in the other families are coincident in this family, others are more divergent. It is clear from the above that several constants measure the same facet of the vocabulary structure. Thus, only those constants with the greatest discriminatory sensitivity within a given family need to be considered. The developmental profiles of the constants show sensitivity to authorship, although this is not absolute in that texts written by the same author may diverge. We have also developed techniques for evaluating the statistical significance of patterns of similarity and dissimilarity in the developmental curves. While the variance of most constants is not known, so that comparisons on the basis of constants for full texts remain impressionistic, we can now evaluate in a more precise way whether or not the developmental profile of a constant differentiates between texts. Conclusions Almost all textual constants in our survey are highly variable, and assume values that change systematically as the text size is increased.

Some constants are inherently variable, others are truly constant in theory. All constants are substantially influenced by the non-random way in which word usage is governed by discourse cohesion. This variability indicates that the constants cannot be relied on to compare texts of different lengths. Crucially, however, the developmental profiles of the majority of constants have an interesting discriminatory potential, in that they reveal consistent and interpretable patterns that pick up author-specific aspects of word use. For authorship attribution studies, we strongly recommend the use of the developmental profiles of selected constants, rather than the isolated values of the constants for complete texts. The discourse structure of texts by the same author can be quite different, and the same holds for the kind of vocabulary an author exploits for a given text. Compared to the use of syntax, word use is more easily influenced by choices which are under the conscious control of authors. Consequently, the developmental profiles of constants are less reliable than syntax-based measures for the purpose of authorship attribution. At the same time, the developmental profiles capture essential differences in word use and discourse structure. From this perspective, we would like to defend their usefulness in the domain of quantitative stylistics. Oxford University Press, Using syntactic annotation to enhance authorship attribution. Literary and Linguistic Computing: Texts used in this study Author Title Key Baum, L. There are major problems in the science of "non-traditional" authorship attribution studies those using statistics and the computer. This paper will show that the problems exist, will list and explain some of the more major problems, and will offer some suggestions on how these problems can be resolved. Non-traditional authorship attribution research has had enough time and effort -- well over studies and 30 years -- to pass through the "shake-down" phase and enter one marked by steady, solid, and scientific studies that force a consensus among its practitioners. A major indication that there are problems in a field is when there is no consensus as to correct methodology or technique. Every area of authorship attribution studies has this problem -- e. It seems that for every paper announcing an authorship attribution method that "works" or a variation of one of these methods, there is a counter paper pointing out crucial flaws, e. Most authorship attribution studies have been governed by expediency, e. The text is not what should be used but it was available. There is a lack of experimental memory. Researchers working in the same "area" of authorship attribution fail to cite and make use of pertinent previous efforts. Too many researchers are led into the swampy quicksand of statistical studies by the ignis fatuus of a "more sophisticated statistical technique". Problems and suggested solutions: The "umbrella" problem is that most non-traditional authorship attribution researchers do not understand what constitutes a valid study. They do not understand that it is a scientific experiment and must be approached and carried out as such. The corrections for many of the specific problems become apparent once the problem is pointed out and there is a consensus that there is a problem.

Chapter 4 : Basic Stylometry Â» Peter Kirby

Stylometry uses statistical methods to analyze style in order to determine authorship. Here is an overview.. Style + Measurement = Stylometry. It is largely based in Attribution Studies and Computational Linguistics, but it can also be used for Forensic Analysis.

This KnoWhy is the first in a series which discusses stylometry and its relevance to questions of Book of Mormon authorship. This first article explains what stylometry is and gives readers a short history of stylometric studies performed on the Book of Mormon. Building on this foundation, subsequent KnoWhys will discuss some of the exciting new results from more recent stylometric research. In order to help shed light on this issue, several studies have relied upon a type of analysis called stylometry. Hilton Study In , 19 John Hilton and a team of researchers from Berkeley most of whom were not LDS 20 conducted a study using word pattern ratios 21 and a new method of differentiation based on what Hilton called rejections. Using a slightly different method, researchers from Utah State University essentially reproduced the results of the Hilton study in This means that NSC analysis will always deliver positive results for one of the candidate authors in the set, even if his or her style happens to be very different from the Book of Mormon. The Fields study also included Joseph Smith as a candidate author. Thus, the Fields research team offers a third stylometric study on the Book of Mormon which independently contradicts theories of 19th century authorship. In contrast the Larsen, Hilton, and Fields studies each produced sound results. Their mutually supporting conclusions should therefore be taken seriously by anyone assessing questions of Book of Mormon authorship. It should be understood that stylometry cannot prove that the Book of Mormon was written by multiple ancient American prophets. What it can reliably demonstrate, and what valid data from the above studies collectively argue, is that 1 the Book of Mormon was written in multiple, distinct authorship styles, 2 these distinct styles are consistent with the authors designated within the text itself, and 3 none of the proposed 19th century authorsâ€”including Joseph Smith himselfâ€”have writing styles that are similar to those found in the Book of Mormon. Not only do these conclusions strongly refute the most popular alternative theories for 19th century authorship, but they can also strengthen faith that the Book of Mormon is what it claims to be. Bruce Schaalje, John L. Hilton, and John B. The Evidence for Ancient Origins, ed. FARMS, , â€" New Light on Ancient Origins, ed. Reynolds and Charles D. For an overview of this field of study, see Michael P. See Frederick Mosteller and David L. Wallace, Inference and Disputed Authorship: Addison-Wesley, ; David I. Churchill, Shakespeare and His Betters: Cambridge University Press, These findings are important for Book of Mormon authorship studies since it is claimed to be an English translation of a text written in an ancient language. See also, John L. For a brief overview of these methods, see Paul J. CA is a method that can identify which items are closest to each other among all items compared. LDA is a method for determining a set of mathematical functions discriminant functions that can be used to classify items into categories based on their characteristics. Non-contextual words are ideal for statistical analysis because they show up frequently and most authors have very little conscious awareness of their own unique patterns of using them. As explained by Fields et al. Phelps, Oliver Cowdery, and Parley P. The Lectures on Faith and two sections from the Doctrine and Covenants were also included p. For example, see D. In response to these concerns, see Wayne A. Larsen and Alvin C. The basic assumption that underlies it is false. For the summarized findings of the Hilton study, see John L. A Decade of New Research, ed. The 65 noncontextual word-pattern ratios which Hilton relied upon in his study were derived from A. The total of the rejections measured when the two texts are tested for a large number of word patterns is identified as the number of rejections. The larger the number of rejections, the more likely the disputed text was not written by the author of the other compared text. Thus, testing a contested document against comparable texts from all possible candidate-authors will identify the most likely writer by eliminating authors whose texts generate high numbers of rejections. This study used a generalized discriminant analysis which is an extension of the linear discriminant analysis used in the Larsen study. Moon, Peg Howland, and Jacob H. Witten, and Craig S. While delta analysis was already being used in stylometric studies, the application of NSC, which was originally developed for genomic testing, was unique. A

foundational premise of the Jockers study was that the once prominent, though long discarded, Spaulding-Ridgon theory of Book of Mormon authorship might actually be valid. For several historical-based arguments against this well-known theory, see Matthew Roper and Paul J. Bruce Schaalje, Paul J. Fields, Matthew Roper, Gregory L. The conclusions of these findings were summarized and adapted for LDS audiences in two different articles: In total, four different stylometric methods were used in the Larsen and Hilton studies. Each of these methods independently detected evidence of multiple authorship and ruled out commonly proposed 19th century candidates. Thus, when viewed collectively, there is plenty of corroborating data to confirm the most fundamental conclusions of these studies. Charles Scribner and Sons,

Chapter 5 : Stylometry and Attribution Studies - Oxford Scholarship

Their contribution comes to answer some dissenting voices arguing that the idiosyncrasies of characters' speech have a direct effect on the so-called singularity of the individual in language, an argument which in the last decade or so has called into question the validity of attribution studies, literary compositions in particular. 7 7.

Stylometry Save Stylometry is the application of the study of linguistic style , usually to written language, but it has successfully been applied to music[1] and to fine-art paintings[2] as well. History Stylometry grew out of earlier techniques of analyzing texts for evidence of authenticity, author identity, and other questions. The modern practice of the discipline received major impetus from the study of authorship problems in English Renaissance drama. Researchers and readers observed that some playwrights of the era had distinctive patterns of language preferences, and attempted to use those patterns to identify authors in uncertain or collaborative works. Early efforts were not always successful: The development of computers and their capacities for analyzing large quantities of data enhanced this type of effort by orders of magnitude. The great capacity of computers for data analysis, however, did not guarantee quality output. In the early s, Rev. Morton produced a computer analysis of the fourteen Epistles of the New Testament attributed to St. Paul, which showed that six different authors had written that body of work. One notable early success was the resolution of disputed authorship in twelve of The Federalist Papers by Frederick Mosteller and David Wallace. Indeed, this was apparent even before the advent of computers: Applications Applications of stylometry include literary studies, historical studies, social studies, gender studies, and many forensic cases and studies. Notable cases Stylometry has been used in a number of high-profile cases. Matthew Jockers used stylometric techniques to analyse the Book of Mormon , concluding that Sidney Rigdon , and not Joseph Smith , was its author. The closed-set method leads to nonsensical results when applied to other authorship attribution problems, as the Schaalje et al. Schaalje and colleagues produced results from an open-set NSC method showing that Sidney Rigdon was a very unlikely authorâ€”a result that comports with documented history that Rigdon and Smith did not even know each other in or early , when the Book of Mormon was dictated and published, making it manifestly impossible for Rigdon to have been the author. Schaalje and colleagues also demonstrated that both Joseph Smith and the men who acted as scribes were unlikely authors, while also revealing multiple distinct voices aligning with the main purported authors of the text. PAN formulates shared challenge tasks for plagiarism detection,[25] authorship identification,[26] author gender identification,[27] author profiling ,[28] vandalism detection,[29] and other related text analysis tasks, many of which hinge on stylometry. Case studies of interest Around to BC, as recorded in the Book of Judges , one tribe identified members of another tribe in order to kill them by asking them to say the word Shibboleth which in the dialect of the intended victims sounded like "sibboleth". Helander was first convicted of writing the letters and lost his position as bishop but later partially exonerated. The letters were studied using a number of stylometric measures and also typewriter characteristics and the various court cases and further examinations, many contracted by Helander himself during the years up to his death in discussed stylometric methodology and its value as evidence in some detail. After his personal notes were made public on his 90th birthday in , a study to determine which of those talks were written by him and which were written by various aides used stylostatistical methods. This case was only resolved after a handwriting analysis confirmed the authorship. In , stylometric methods were used to compare the Unabomber manifesto with letters written by one of the suspects, Theodor Kaczynski to his brother, which led to his apprehension and later conviction. In , a group of linguists, computer scientists, and scholars analysed the authoship of Elena Ferrante. They were able to compare her writing style with 39 other novelists using, for example, stylo. Domenico Starnone is the secret hand behind Elena Ferrante. Most methods are statistical in nature, such as cluster analysis and discriminant analysis , are typically based on philological data and features, and are fruitful application domains for modern machine learning approaches. Whereas in the past, stylometry emphasized the rarest or most striking elements of a text, contemporary techniques can isolate identifying patterns even in common parts of speech. Most systems are based on lexical statistics, i. In this context, unlike in information retrieval , the observed occurrence patterns of the most

common words are more interesting than the topical terms which are less frequent. An example of a writer invariant is frequency of function words used by the writer. In one such method, the text is analyzed to find the 50 most common words. The text is then broken into 5, word chunks and each of the chunks is analyzed to find the frequency of those 50 words in that chunk. This generates a unique number identifier for each chunk. These numbers place each chunk of text into a point in a dimensional space. This dimensional space is flattened into a plane using principal components analysis PCA. If two literary works are placed on the same plane, the resulting pattern may show if both works were by the same author or different authors. Neural networks Neural networks , a special case of statistical machine learning methods, have been used to analyze authorship of texts. Text of undisputed authorship are used to train the neural network through processes such as backpropagation , where training error is calculated and used to update the process to increase accuracy. Through a process akin to non-linear regression, the network gains the ability to generalize its recognition ability to new texts to which it has not yet been exposed, classifying them to a stated degree of confidence. Such techniques were applied to the long-standing claims of collaboration of Shakespeare with his contemporaries Fletcher and Christopher Marlowe ,[41][42] and confirmed the view, based on more conventional scholarship, that such collaboration had indeed taken place. This study from Vrije Universiteit examined identification of poems by three Dutch authors using only letter sequences such as "den". This involves a method that starts out with a set of rules. An example rule might be, "If but appears more than 1. The program is presented with text and uses the rules to determine authorship. The rules are tested against a set of known texts and each rule is given a fitness score. The 50 rules with the lowest scores are thrown out. The remaining 50 rules are given small changes and 50 new rules are introduced. This is repeated until the evolved rules correctly attribute the texts. Rare pairs One method for identifying style is called "rare pairs", and relies upon individual habits of collocation. The use of certain words may, for a particular author, idiosyncratically entail the use of other, predictable words. Authorship attribution in instant messaging The diffusion of Internet has shifted the authorship attribution attention towards online texts web pages, blogs, etc. Efforts to take into account such aspects at the level of both structure and syntax were reported in. Furthermore, the similarity between spoken conversations and chat interactions has been neglected while being a key difference between chat data and any other type of written information.

DOWNLOAD PDF 7. STYLOMETRY AND ATTRIBUTION STUDIES

Chapter 6 : Introduction to stylometry with Python | Programming Historian

Stylometry is the application of the study of linguistic style, usually to written language, but it has successfully been applied to music and to fine-art paintings as well. [3] Stylometry is often used to attribute authorship to anonymous or disputed documents.

Advanced Search 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. 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. This fact had an immense influence on the further development of the whole discipline. The seminal study by Mosteller and Wallace [] showed in a very convincing way that authorship attribution based on statistical analysis of style is ultimately the problem of classification. Even if one deals with an open-set attribution caseâ€”where the list of possible candidates cannot be reliably establishedâ€”the general idea does not differ substantially from other classification problems. Exact science has developed a number of well-performing, sophisticated machine-learning 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. Independently, a ground-breaking monograph on Jane Austen published by Burrows ushered stylometry into literary criticism. The methods adopted or introduced by Burrows, Hoover, Craig, and others Burrows, , , ; Hoover, a , b ; Craig and Kinney, were very intuitive and easily-applicable to literary studies. These include principal components analysis, multidimensional scaling, cluster analysis, Delta, Zeta, and Iota. Despite their limitations the lack of validation of the obtained results being the most obvious , they are still widely used. The reason of their popularity is that they meet the needs of literary scholars, also because they offer convincing visualizations. Needless to say, visualization has an undeniable explanatory power. Scatterplots, maps, trees, and diagrams provide an insight into the whole corpus at one glance. Moreover, they allow to draw conclusions about literature from a distant-reading perspective, through a visual interpretation of groupings and separations of several samples. Certainly, this is particularly desired in stylometry beyond authorship attribution. The attractiveness of visualization in computational literary criticism is confirmed not only by the aforementioned studies by Burrows or Hoover, but also by immense popularity of beautiful yet relatively simple plots presented by Moretti, Jockers, Posavec, and others Morretti, ; Posavec, ; Jockers, ; Sinclair and Rockwell, The aim of this article is to discuss reliability issues of a few visual techniques, and to enhance the explanatory power of visualization with a procedure of validation inspired by advanced statistical methods. Sophisticated machine-learning methods of classification routinely try to estimate the amount of potential error that may be due to inconsistencies in the analyzed corpus. A standard solution here is a fold cross-validation, or 10 random swaps between two parts of a corpus: Most unsupervised methods used in stylometry, such as principal components analysis, multidimensional scaling, or cluster analysis, lack this important feature. Also, given a tree-like graphical representation of similarities between particular samples, one can easily interpret the results in terms of finding out the group of texts to which a disputed sample belongs. Hierarchical cluster analysisâ€”as discussed in the present studyâ€”is a technique which tries to find the most similar samples e. What makes this method attractive is the very intuitive way of graphical representation of the obtained results: However, despite obvious advantages, some problems still remain unresolved. The final shape of a dendrogram highly depends on many factors, the most important being 1 the particular distance measure applied to the data, 2 the algorithm of grouping the samples into clusters, and 3 the number of variables e. These factors will be briefly discussed below. The distance used by Burrows is a widely accepted solution in the field of computational stylistics; there are no studies, however, that would satisfactorily explain the principles of using this particular measure. If this is true, the distance used here is in fact equivalent to the Linear Delta measure introduced by Argamon , p. There is no denying that

Delta, and ipso facto the distance measure embedded in it, proved to be very effectiveâ€”a fact confirmed by numerous stylometric studies; thus, it should be also applicable to hierarchical cluster analysis procedure. Even if convincing at first glance, however, the choice of this particular measure needs to be theoretically justified and confirmed by empirical comparisons with other distances. Another factor affecting the final shape of a dendrogram is the method of linkage used. In the above-cited statement, Burrows favors the complete linkage algorithm as the most effective one. We do not know, however, which were the other algorithms considered by Burrows, and we do not know what method of comparison was used to test their effectiveness. Although it seems to be accurate indeed, there is no awareness, however, that this method has been designed for large-scale tests of more than samples: However, the same cannot be said about the third factor cluster analysis depends on, which is the number of features e. Color versions of all figures are available online. The question how many features should be used for stylometric tests has been approached in many studies, but no consensus has been achieved: Although all these solutions are reasonable and theoretically justified, the final choice of the number of features to analyze is a priori arbitrary. Awareness of this issue, followed by partial solution, can be observed in the studies by Hoover a , b , who assesses a given corpus with a few discrete cluster analyses for different most frequent word MFW values. Even if still subject to arbitrary choices, this approach gives a fairly good insight into variability of the input data. This way of dealing with uncertainty will be discussed below in detail, with its possible extension to other visualization techniques. Without deciding which of the three factors discussed in the previous sectionâ€”linkage algorithm, distance measure, and the number of words analyzedâ€”is more likely to affect the final shape of a dendrogram, one must admit that the first two are related to the method of clustering, while the third factor is inherently linked to certain linguistic features of analyzed texts. Endless discussions of how many frequent words or n-grams should be taken into account e. Mosteller and Wallace, ; Hoover, a ; Burrows, ; Koppel et al. Just the opposite, it seems that the authorial signal is spread throughout the whole frequent and not-so-frequent words spectrum, but at the same time it may become obscured by additional and unpredictable signals, which are considered noise in classical approaches to attribution. Why are some authors misclassified? Which texts are wrongly attributed to a given author, and why are they linked to this very author and not to others? Obviously, the problem is not new. Cross-genre authorship attribution, for one, has always been a major challenge Kestemont et al. Also, there have been a few attempts to extract particular signals hidden in texts: On theoretical grounds, function words should be responsible for authorial recognition, while content words should be more topic- and genre-related. The abovementioned empirical studies, however, do not really confirm this assumption. There is no clear rule here, and the same words are sometimes claimed to reveal different signals. The difficulties with separating one specific signal suggest that a text written or spoken is a multi-layer phenomenon, in which particular layers are correlated. These layers include authorship, chronology, personality, gender, topic, education, literary quality, translation if applicable , intertextuality, literary tradition e. Arguably, literary quality somehow depends on education, genre depends on topic, authorial voice is affected by chronology, gender affects personality, and so on. Some layers might be barely noticeable, and some others might become surprisingly strong. In authorship attribution, this complex system of uncontrollable layers is a problem of unwanted noise, and in literary-oriented computational stylistics, an opportunity to see more. Different combinations of linkage algorithms, number of MFWs, and distance measures applied, one obtains a convincing example of how unstable the final results might be. What does the main division into two large clusters mean? Figures 1â€”4 might support many contradictory hypotheses. The problems do not end here: An example of this behavior is shown in Figs 3 and 4. What is more important here is the side-effect: Such abrupt changes seem to be a rule rather than the exception, at least for textual data sets. The decision which of the dendrograms presented above reveal the actual separation of the samples and which show fake similarities is not trivial at all. Generating hundreds of dendrograms covering the whole spectrum of MFWs, a variety of linkage algorithms, and a number of distance measures, would make this choice even more difficult. At this point, a stylometrist inescapably faces the abovementioned cherry-picking problem Rudman, This technique has been developed in phylogenetics Paradis et al. It has been also introduced into stylometry Eder, b and applied in a number of stylometric studies Rybicki, ; Rybicki and Heydel, ; van Dalen-Oskam, The goal, then,

is to capture the robust patterns across a set of generated snapshots. The procedure is aimed at producing a number of virtual dendrograms, and then at evaluating robustness of groupings across these dendrograms. Unlike typical dendrograms, however, the established links do not represent stylometric distances between samples.

Chapter 7 : Shakespeare attribution studies - Wikipedia

Lexical `constants' in stylometry and authorship studies Fiona J. Tweedie R. Harald Baayen Introduction Various measures of lexical richness have been employed in stylometry and authorship attribution (see, e.g., Holmes, , for a review).

PAN formulates shared challenge tasks for plagiarism detection, [25] authorship identification, [26] author gender identification, [27] author profiling , [28] vandalism detection, [29] and other related text analysis tasks, many of which hinge on stylometry. Case studies of interest[edit] Around to BC, as recorded in the Book of Judges , one tribe identified members of another tribe in order to kill them by asking them to say the word Shibboleth which in the dialect of the intended victims sounded like "sibboleth". Helander was first convicted of writing the letters and lost his position as bishop but later partially exonerated. The letters were studied using a number of stylometric measures and also typewriter characteristics and the various court cases and further examinations, many contracted by Helander himself during the years up to his death in discussed stylometric methodology and its value as evidence in some detail. After his personal notes were made public on his 90th birthday in , a study to determine which of those talks were written by him and which were written by various aides used stylostatistical methods. This case was only resolved after a handwriting analysis confirmed the authorship. In , stylometric methods were used to compare the Unabomber manifesto with letters written by one of the suspects, Theodor Kaczynski to his brother, which led to his apprehension and later conviction. In , a group of linguists, computer scientists, and scholars analysed the authoship of Elena Ferrante. They were able to compare her writing style with 39 other novelists using, for example, stylo. Domenico Starnone is the secret hand behind Elena Ferrante. Most methods are statistical in nature, such as cluster analysis and discriminant analysis , are typically based on philological data and features, and are fruitful application domains for modern machine learning approaches. Whereas in the past, stylometry emphasized the rarest or most striking elements of a text, contemporary techniques can isolate identifying patterns even in common parts of speech. Most systems are based on lexical statistics, i. In this context, unlike in information retrieval , the observed occurrence patterns of the most common words are more interesting than the topical terms which are less frequent. An example of a writer invariant is frequency of function words used by the writer. In one such method, the text is analyzed to find the 50 most common words. The text is then broken into 5, word chunks and each of the chunks is analyzed to find the frequency of those 50 words in that chunk. This generates a unique number identifier for each chunk. These numbers place each chunk of text into a point in a dimensional space. This dimensional space is flattened into a plane using principal components analysis PCA. If two literary works are placed on the same plane, the resulting pattern may show if both works were by the same author or different authors. Neural networks[edit] Neural networks , a special case of statistical machine learning methods, have been used to analyze authorship of texts. Text of undisputed authorship are used to train the neural network through processes such as backpropagation , where training error is calculated and used to update the process to increase accuracy. Through a process akin to non-linear regression, the network gains the ability to generalize its recognition ability to new texts to which it has not yet been exposed, classifying them to a stated degree of confidence. Such techniques were applied to the long-standing claims of collaboration of Shakespeare with his contemporaries Fletcher and Christopher Marlowe , [41] [42] and confirmed the view, based on more conventional scholarship, that such collaboration had indeed taken place. This study from Vrije Universiteit examined identification of poems by three Dutch authors using only letter sequences such as "den". This involves a method that starts out with a set of rules. An example rule might be, "If but appears more than 1. The program is presented with text and uses the rules to determine authorship. The rules are tested against a set of known texts and each rule is given a fitness score. The 50 rules with the lowest scores are thrown out. The remaining 50 rules are given small changes and 50 new rules are introduced. This is repeated until the evolved rules correctly attribute the texts. Rare pairs[edit] One method for identifying style is called "rare pairs", and relies upon individual habits of collocation. The use of certain words may, for a particular author, idiosyncratically entail the use of other, predictable words.

DOWNLOAD PDF 7. STYLOMETRY AND ATTRIBUTION STUDIES

Authorship attribution in instant messaging[edit] The diffusion of Internet has shifted the authorship attribution attention towards online texts web pages, blogs, etc. Efforts to take into account such aspects at the level of both structure and syntax were reported in. Furthermore, the similarity between spoken conversations and chat interactions has been neglected while being a key difference between chat data and any other type of written information.

Chapter 8 : Stylometry - Wikipedia

Middle Dutch literature, Stylometry, Authorship attribution, Stylistics, Digital Humanities When dealing with literary statistics it is best to err on the side of caution.

Chapter 9 : What Can Stylometry Tell Us about Book of Mormon Authorship? | Book of Mormon Central

Literary studies has arguably been the most active branch of the humanities in computational text analysis, most notably in the use of stylometry, a form of inquiry that combines literary theory with linguistics in order to examine the underlying structure of texts.