Texts as networks: How many words are sufficient to identify an author?

12 April 2019

The author of an unsigned text can be identified by analyzing the relationship between just a few words of the text, as shown by physicist-statisticians from the Institute of Nuclear Physics of the Polish Academy of Sciences in Cracow. (Source: IFJ PAN) Credit: IFJ PAN

People are more original than they think—this is suggested by a literary text analysis method of stylometry proposed by scientists from the Institute of Nuclear Physics Polish Academy of Sciences. The author's individuality can be seen in the connections between no more than a dozen words in an English text. It turns out that in Slavic languages, authorship identification requires even fewer words, and is more certain.

The researchers sought a solution to the problem of verifying the authorship of historical texts known only from fragments, the identification of plagiarism, and similar problems. In many cases, traditional stylometric methods fail or do not lead to sufficiently reliable conclusions. In *Information Sciences*, scientists from the Institute of Nuclear Physics of the Polish Academy of Sciences (IFJ PAN) in Cracow now present their own statistical tool for stylometric analysis. Constructed with the use of graphs, it analyzes the structure of texts in a qualitatively new way.

"The conclusions of our research are, on the one hand, encouraging. They indicate that the individuality of any person manifests itself clearly in the way they use a surprisingly small number of words. But there is also a dark side. Since it turns out that people are so original, it will be easier to identify individuals by their statements," says Professor Stanislaw Drozdz of Cracow University of Technology.

Stylometry, the science dealing with the statistical characteristics of the style of texts, is based on the observation that each person uses the same language in slightly different ways. Some have a broader vocabulary, others narrower, some prefer certain phrases and make mistakes, others avoid repetition and are linguistic purists. And in written text, they also differ in the way they use punctuation. In the typical stylometric approach, the basic features of a text are usually examined, including the frequency of occurrence of individual words, while punctuation is ignored. Analyses are carried out for the studied text and for texts written by potentially well-known authors. The creator is deemed to be the person whose works have parameters with the values closest to those obtained for the material being identified.

"We suggested that the characteristic features of the style could be represented in a network representation of the text, using graphs," explains Tomasz Stanisz, Ph.D. student at the IFJ PAN and the first author of the publication. "The graph is a collection of points or vertices on the graph, connected by lines, i.e. the edges of the graph. In the simplest case—in the so-called unweighted network—the vertices correspond to individual words and are connected by edges if and only if two given words have occurred adjacent to each other at

least once in the text. For example, for the sentence 'Jane is hungry,' the graph would have three vertices, one for each word, but there would only be two edges, one between 'Jane' and 'is,' the other between 'is' and 'hungry.'"

While constructing their stylometric tools, the researchers tested different types of graphs. The best results were obtained for weighted graphs, that is, those in which each edge carries information about the number of occurrences of its corresponding connection between words. Two parameters turned out to be the most useful in such networks: the node degree and the clustering coefficient. The first describes the number of edges coming from a given node and is directly related to the number of occurrences of a given word in the text. In turn, the clustering coefficient describes the probability that two words connected by an edge with a given word are also connected with an edge between themselves.

Using statistical tools prepared in this way, the Cracow-based physicists looked at 96 books: six novels by eight well-known English authors (Austen, Conrad, Defoe, Dickens, Doyle, Eliot, Orwell and Twain) and eight Polish authors (Korczak, Kraszewski, Lam, Orzeszkowa, Prus, Reymont, Sienkiewicz and Zeromski). The authors included two winners of the Nobel Prize for Literature (Wladyslaw Reymont and Henryk Sienkiewicz). All the texts were obtained from internet resources: Project Gutenberg, Wikisource and Wolne Lektury. The group from the IFJ PAN then checked the reliability with which the authorship of 12 randomly selected works in one language could be determined, treating the rest of the pool of works as comparative material.

"In the case of English texts, we identified the authors correctly in almost 90 percent of cases. In addition, in order to achieve success, it was necessary to trace the connections between only 10 to 12 words of the examined text. Contrary to naive intuition, a further increase in the number of words studied did not significantly increase the effectiveness of the method," says Stanisz.

In Polish, the determination of authorship turned out to be even simpler: analyzing only five to six words was required. Notably, despite the fact that the pool of significant words was half as many as in English, the probability of correct identification was increased by up to 95 percent. Such high diagnostic accuracy, however, was only achieved when punctuation marks were also treated as separate words. In both languages, omitting punctuation resulted in a significant reduction in the number of correct guesses. The observed role of punctuation is another confirmation of the conclusions from a 2017 publication of the group of Prof. Drozdz, in which it was shown that punctuation plays a role in language equally important as the words themselves.

"In comparison with English, Polish seems to give greater possibilities of revealing the style of the author. We think that the other Slavic languages are characterised by similar features. English is a positional language, which means that the order of the words in a sentence is important. This sort of language leaves less room for an individual style of expression than the Slavic languages, in which inflection, or variation, determines the role of a word or phrase in a sentence. This allows for greater freedom to organize the order of words in a sentence, while its meaning remains unchanged," says Prof. Drozdz.

More information: Andrzej Kulig et al, In narrative texts punctuation marks obey the same statistics as words, *Information Sciences* (2016). DOI: 10.1016/j.ins.2016.09.051

Provided by Polish Academy of Sciences