Stylometry: An Overview

Stylometry

This half of the course will explore the field of stylometry – the statistical analysis of literary style, and literary texts.

Specifically we will look at the task of authorship attribution – given a particular text with an unknown author, the goal is to determine which author wrote the text.

Here ‘text’ can mean essentially any piece of writing. Classic stylometry focused on the analysis of literature and historically important documents, while more modern applications include the analysis of social media and other internet-relevant writing. An obvious reason example is identifying ChatGPT (and similar LLM) output.

Classic Example – Federalist Papers

The Federalist Papers are a historically significant set of political essays written in 1787-1788 three of the Founding Fathers of the United States of America (Alexander Hamilton, James Madison, and John Jay). They consist of 85 essays, where each was written by a single one of the above three authors

For 73 of these essays, the authorship is known and agreed on by historians. However the author of the remaining 12 essays is unknown, although it will be one of Hamilton, Madison, and Jay.

In a classic analysis in 1963 which kickstarted the modern field of stylometry, the statisticians Mosteller and Wallace used statistical methods to analyse these 12 disputed essays, and concluded that Madison was the most likely author

Modern Example – The Cuckoo's Calling

After completing the Harry Potter series of novels, the author J. K. Rowling started to write other novels under a pseudonym 'Robert Galbraith' to hide her identity and to avoid these being judged in reference to previous work. The first such novel was 'The Cuckoos Calling' and was released pseudonymously in 2013.

After being tipped off by a personal source that J. K. Rowling was the true author of this novel, The Sunday Times employed a statistician trained in stylometry to verify that she was indeed the author, by comparing the writing style of 'The Cuckoos Calling' to her previous novels

Non-English Example – Dream of the Red Chamber

Stylometric analyses have also been applied to non-English language texts. The classic Chinese novel ‘Dream of the Red Chamber’ consists of 120 chapters, and it is known that the first 80 chapters were written by a single author, Cao Xueqin.

The authorship of the final 40 chapters is disputed, and literary scholars have argued over whether Cao is also the sole author of these, or whether they come from a separate source.

In recent years, statistical techniques from stylometry have been used to analyse this problem.

Mathematical Formulation

Mathematical Formulation

In the classic authorship attribution setting, we have an unknown text $\tilde{x}$ for which we need to determine the author.

Suppose that there are A possible authors who could have written this texts (e.g. $A = 3$ for the above Federalist Papers example). For each author j, we assume we have access to n_j texts which were known to be written by this author. Denote these as $\mathbf{x}_j = (x_{j,1}, \ldots, x_{j,n_j})$. This collection of texts is known as a corpus.

For each individual text x_i in the corpus, let its (known) author be denoted by y_i. The corpus hence consists of a set of pairs (y_i, x_i)

Goal: use these known texts to learn the writing style of each of the A authors, and hence determine the unknown author of $\tilde{x}$. In machine learning terms, this is a **supervised learning** or **classification** problem.

Mathematical Formulation

The first task is representing each written text as mathematical objects. In other words, what sort of objects are the x's above?

We require a way of representing a text in a quantitative numerical form that is amenable to statistical analysis. To do this, we will extract relevant features from the text. This is known as **feature extraction**.

There are many ways in which this can be done. An intuitively simple approach is to look for the occurrence of **rare words**. Perhaps certain people use certain uncommon words more often than other people do, in a way which lets their writing by identified.

Mathematical Formulation

The general consensus in stylometry is that performing authorship attribution by looking at the occurrence of rare/uncommon words is not a good idea.

This is because the use of such words is very context dependent. For example, the Harry Potter novels contain words such as ‘Quidditch‘ and ‘Slytherin’ which are not common in the English language. But these would not be useful for predicting the author of the Cuckoos Calling, whcih is a crime novel set in the real world.

Similarly a novel set in revolutionary France would be expected to contain certain uncommon words specific to that period, which would not be contained in a novel set in modern Scotland.

Essentially, rare/uncommon words are likely to capture the idiosyncratic setting/context of the novel, rather than the actual writing style of the author.

Mathematical Formulation

Instead, most authorship attribution instead looks at the occurrence pattern of **common words**

In their classic study of the Federalist Papers, Mosteller and Wallace looked only at how often the different authors used various so-called ‘function words’.

Function words are the context free words which make up the basic grammar of the English language. The 70 function words considered were

a, all, also, an, and, any, are, as, at, be, been, but, by, can, do, down, even, every, for, from, had, has, have, her, his, if, in, into, is, it, its, may, more, must, my, no, not, now, of, on, one, only, or, our, shall, should, so, some, such, than, that, the, their, then, there, things, this, to, up, upon, was, were, what, when, which, who, will, with, would, your

Mathematical Formulation

Remarkably, different people will exhibit small variations in how often they use these function words when writing, in a way which allows their writing to be identified

This is quite surprising – the frequency with which you use words such as ‘the’ and ‘it’ will vary enough from your classmates to allow your writing to be distinguished from theirs with near-perfect accuracy, given a large enough sample of your writing.

Unlike rare/uncommon words, the frequency with which authors use function words tends to be fairly context-invariant, and is hard to deliberately mask.

Mathematical Formulation

Basic idea: given a text with a known author, we count how many times each of these function words has appeared in the text. For the 70 function words above, this means that each text is represented mathematically as a length 70 vector, where the first component is the number of times the word ‘a’ appears in the text, the second component is the number of times the word ‘all’ appears, and so on.

By convention, we include a 71^{st} component to count the number of non-function words (i.e. every word in the text other than the 70 function words above)

Example

Consider the following sample text from the Federalist Paper:

> There are again two methods of removing the causes of faction: the one, by destroying the liberty which is essential to its existence; the other, by giving to every citizen the same opinions, the same passions, and the same interests.
>
> It could never be more truly said than of the first remedy, that it was worse than the disease. Liberty is to faction what air is to fire, an aliment without which it instantly expires. But it could not be less folly to abolish liberty, which is essential to political life, because it nourishes faction, than it would be to wish the annihilation of air, which is essential to animal life, because it imparts to fire its destructive agency. The second expedient is as impracticable as the first would be unwise.

Example

To convert this to a vector, we count the number of times the word ‘a’ occurs, and make this the first element. This is repeated for each of the above 70 function words. You can check that the resulting vector for this text is

0, 0, 0, 1, 1, 0, 1, 2, 0, 4, 0, 1, 2, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
6, 7, 2, 0, 1, 0, 0, 0, 1, 0, 4, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 3, 1, 12, 0, 0, 1,
0, 0, 9, 0, 0, 1, 0, 1, 0, 4, 0, 0, 0, 2, 0, 62

So ‘a’, ‘all’, ‘also’ never occurred, ‘an’ and ‘and’ each occurred once, ‘the’ occurred 12 times, and so on.

Example

Our corpus will consist of numerous texts, each repersented as length 71 vectors. However the texts will generally be of different lengths, which makes it difficult to directly compare the counts of how often each function word occurs.

As such, we typically **normalise** the texts so they sum to 1, by defining a new vector where the elements are:

$$x_i' = \frac{x_i}{\sum_{i=1}^{71} x_i}$$

This means that rather than (e.g.) the first element being a **count** of the number of times the word ‘a’ occurred, it is now the **proportion** of the words which are ‘a’

Other Feature Sets

The choice of 70 function words from the Federalist Papers study has been influential, but many other different choices of features can also be considered.

In general, the goal is to capture the **common** features of language.

Given a corpus of texts, we could instead find the 100 (or more) most common words in that particular corpus, and then represent each text as a length 101 (or more) vector containing the counts or proportions of these words. We must use the same words for each text in the corpus, of course.

Note that this easily allows stylometry to be deployed on non-English language texts even when you are not familiar enough with the language to know the function words. We can instead use the most common words from the text, which are likely to be very similar to the function words

Example

To illustrate this, here are the 100 most commonly occurring words on English-language Twitter, in order (starting with the most common, and with all punctuation removed).

the, i, to, and, a, of, in, that, it, my, is, you, for, was, on, me, but, so, this, with, have, be, we, at, not, all, he, im, as, like, are, just, its, out, up, about, they, what, or, one, if, do, from, had, get, when, will, there, her, dont, she, time, know, now, can, some, then, by, his, really, no, well, an, your, go, more, were, am, think, would, who, people, good, been, how, has, got, him, them, going, because, back, day, see, much, our, only, which, want, their, love, even, other, too, after, today, went, over, way, here

Note this list contains most of the above function words

Summary

Regardless of which feature set we choose, the end result is that the corpus consists of a number of known texts (y_i, x_i) where y_i is a categorical variable denoting the author of the text, and x_i is a vector which represents the occurrence proportion of the chosen function words in the text.

We are then presented with a new text for which the author is unknown. We first convert this into a vector $\tilde{x}$ using the same function words, and then try to predict the most likely author (i.e. $\tilde{y}$).

This is a supervised learning problem, and can be attacked using standard techniques that we will discuss in the next lecture