Authorship and Stylometry

0930 Wednesday 18 April
marc.alexander@glasgow.ac.uk

Stylometry

The "measurement" of style
Sometimes called computational stylistics or computational text analysis

Used for:
- genre classification
- diachronic linguistics
- literary analysis
- authorship attribution

Discriminators

Criteria which make a style distinguishable
And, for our purposes, that computers can count

Need closed questions, not open ones

‘If X wrote A, did X also write B?’
‘Who of X, Y, or Z is more likely to have written A?’

NOT ‘Who wrote X?’

Style

Must begin with an understanding of style
Authors’ styles change over time and with their purpose:

- style of author
- style of author at a particular time
- style of author in a particular way
- style of author when writing for a particular person
- style of author when writing in a particular genre
Style

Style can be:
- subconscious ("real" style)
- conscious (pastiche and parody)

So questions are better if they don’t ask:
- Is this like Jane Austen’s style?

And instead ask:
- Is this like Jane Austen’s style of writing the speech of female characters in her middle novels?

Stylometry

Stylometry examples

Federalist Papers
Mosteller and Wallace, 1964; closed set of authors (Hamilton, Madison, Jay)

Primary Colors
Roman à clef about US politics and a Bill-Clinton-like governor

The Book of Mormon, Jockers et al.
The style of Henry James; David Hoover

Douglas Biber et al. on register analysis

Stylometry

Word/sentence length
Frequencies of letter pairs
Vocabulary richness
Word frequency
Selected sets of words
- Mosteller and Wallace: upon, whilst, there, on, while, by, consequently, would, etc

Most frequent words (MFW)

MFW

Best technique
Focuses on function words
Very successful (often very good in 500-4000 range)
Usually needs culling
Once you have your data...

You have multivariate data
  data with lots of variables
You then need the ‘distance’ between each variable
  We use Burrows’ Delta, the most often used measure
Don’t really need to understand the inner workings of Delta, but there are references at the end of the slides…

Analysing Multivariate Data

Principal Component Analysis (PCA)
Cluster Analysis
Bootstrap consensus

PCA

PCA

PCA

PCA

PCA

Tomoji Tabata
Hugh Craig

Cluster analysis of forty-eight novels:
- comedy
- tragedy
- history
- romance
- Roman play

First principal component

Second principal component

Hugh Craig

Principal Components Analysis

Graph 1. Jane Austen's major characters.

John Burrows
Newman looked at her a moment; he saw that she was pretty, but he was not in the least dazzled. He remembered poor M. Nioche's solicitude for her 'innocence,' and he laughed out again as his eyes met hers. Her face was the oddest mixture of youth and maturity, and beneath her candid brow her searching little smile seemed to contain a world of ambiguous intentions. She was pretty enough, certainly, to make her father nervous; but, as regards her innocence, Newman felt ready on the spot to affirm that she had never parted with it. She had simply never had any; she had been looking at the world since she was ten years old, and he would have been a wise man who could tell her any secrets.

That brought back to Maisie--it was a roundabout way--the beauty and antiquity of her connexion with the flower of the Overmores as well as that lady's own grace and charm, her peculiar prettiness and cleverness and even her peculiar tribulations. A hundred things hummed at the back of her head, but two of these were simple enough. Mrs. Beale was by the way, after all, just her stepmother and her relative. She was just--and partly for that very reason--Sir Claude's greatest intimate ('lady-intimate' was Maisie's term) so that what together they were on Mrs. Wix's prescription to give up and break short off with was for one of them his particular favourite and for the other her father's wife.

David Hoover
Bibliography

Highly recommended:

Classic Federalist Papers work:

Extended Bibliography