

Stylometry unit

- ▶ The authorship attribution problem (**Today**)
- ▶ Lab: Stylometry techniques (Friday)
- ▶ Applied styometry (next week Monday)

Today:

- ▶ Definition of the field and main tasks
- ▶ Example
- ▶ Range of tasks and applications
- ▶ Classification approaches
- ▶ Burrow's delta
- ▶ The stylo package

Stylometry, also called **computational stylistics** is the quantitative study of writing style. It assumes the existence of a **stylome**, a “set of measurable traits of language products” [van Halteren 2005]

Authorship attribution is an application of stylometry in which stylometric techniques are used to determine the author of anonymous or disputed text.

Standard form of the problem:

Given a set of authors (classes), a collection of texts for each author (training set), and a disputed text, classify the disputed text based on similarities with the texts in the training set.

Other forms:

- ▶ Needle-in-a-haystack
- ▶ Verification
- ▶ Profiling
- ▶ Clustering
- ▶ Stylochronometry

[Koppel 2009]

Burrow's delta [Burrows 2002]

Given a set of documents, compute D features (such as word counts). For each feature, compute the mean and standard deviation accross the corpus. Let $\vec{\mu}$ and $\vec{\sigma}$ be the mean and standard deviation vectors.

Let $\vec{x}$ and $\vec{y}$ be vectors for two documents. Then the delta between $\vec{x}$ and $\vec{y}$ is

$$\begin{aligned}\Delta_B(\mathbf{x}, \mathbf{y}) &= \sum_{i=1}^n \left| \frac{x_i - \mu_i}{\sigma_i} - \frac{y_i - \mu_i}{s_i} \right| \\ &= \sum_{i=1}^n \left| \frac{x_i - y_i}{\sigma_i} \right|\end{aligned}$$

Coming up:

- ▶ Do NBC programming assignment (Mon, Nov 10)
- ▶ Read another stylometry paper (Fri, Nov 7)
- ▶ Take stylometry quiz (Tues, Nov 11)

(There will be reading for the neural nets next week.)

Bibliography

Hans van Halteren et al. "New Machine Learning Methods Demonstrate the Existence of a Human Sylome." In *Journal of Quantitative Linguistics*. 12:1. 2005

Moshe Koppel et al. "Computational Methods in Authorship Attribution." In *Journal of the ASIS&T*, 60:1. 2009.

John Burrows. "Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship." Int *Literary and Linguistic Computing*, 17:3. 2002.