Authorship Attribution

Distance-based Methods

Jacques Savoy
University of Neuchâtel


Who is the author?

As possible authors, we have John F. Kennedy, Barack Obama, Abraham Lincoln. Attribute each text to its author.

**Text 1:** “Four score and seven years ago our fathers brought forth, upon this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal.”
**Text 2:** “Yes, we can.”
**Text 3:** “My fellow Americans, ask not what your country can do for you, ask what you can do for your country.”
**Text 4:** “Ich bin ein Berliner”

Different Questions

Given a sample of texts known to be written by one of a set of authors,

- **Question 1:** Closed-set. Determine the author from a set of possible authors (e.g., political tract)
- **Question 2:** Open-set. Determine, if any, the author from the set of possible authors
- **Question 3:** Verification. Determine if the given author is really the correct one. Is it really Shakespeare? (Koppel et al., 2007)
- **Question 4:** Profiling. Determine pertinent attributes of the author (sex, age, education, psychological, …) (Pennebaker, 2011)

Authorship Attribution

- The Program…
  - Problem & Context
  - Examples
  - A single measurement
  - Multivariate analysis (restricted set of terms)
  - Distance-based approaches (ad hoc)
The Output / The Data

- Only the most unlikely / probable author (or a ranked list)
- The style of the author (author’s canon, stylistic traits)
- The assignment reliability
  - Minor vs. large impact: Attribution in the court room
    Forensic Linguistics, (Olsson, 2008)
    but also Pauline Epistles, The Book of Mormon, ...

- Text sample
  - relatively large
  - balanced
  - high quality

Beyond Simple AA

- Collaborative work (with?)
- Part of a play (e.g., a scene) (Craig & Kinney, 2009)
- Analysis character by character / dialogue
  Is Hamlet really a male character?
- Historical study of language change (diachronic linguistics)
  (Juola, 2003)
- Who is behind a politician?
- Profiling the author (Pennebaker, 2011), gender studies
- Plagiarism...
- Email (spam, fraud, propaganda) authentication

How?

- Following St Jerome (347-420 AD)
  1. if one book is inferior to the others
  2. if the text contradicts the doctrine in author’s other works
  3. if the text is written in a different style, contains words and expressions not ordinarily found in the author’s production
  4. if passages quoting statements that were made or mentioning events that occurred after author’s death

  

- Comparative basis

Style

- Measurement of (aspects) of style
  "The stylometrist therefore looks for a unit of counting which translates accurately the 'style' of the text, where we may define 'style' as a set of measurable patterns which may be unique to an author"
  H. Holmes, Authorship Attribution, Computers and Humanities, 1994, p. 87

- Stylistic features (which ones?, how to select?, how many?)
  - Words, sequences of words, lemmas, n-grams, ...
  - POS, sequence, proportions, ...
  - Structural elements (e.g., layout, signature, logical structure, …)

- Hidden assumptions
  - The style is constant for an author in a given period and it differs from other authors
Variations in Style

1. The village does not have a post office.
2. The village has no post office.
3. The village doesn’t have a post office.
4. The village hasn’t got a post office.
5. The village hasn’t got no post office.
6. The village ain’t got no post office.


Style

- Style is a function of
  - Genre (novel vs. poem, prose or verse)
  - Author (social, gender, age, education, native language, …)
  - Period (same time frame)
  - Topic
  - Type (spoken vs. written, web-based)
  - Audience (official vs. informal)
  - Editors / publishers
- Data quality (J. Rudman)
  - De-editing (page number, scene description, character’s names)
  - Spelling normalization (one word = one spelling)

Why Data Quality Matters

- Authorship attribution
  - External evidence (incipits, colophon, biographical evidence, earlier attributions, social world within which the work is created, …)
  - Internal evidence (self-reference, evidence from themes, ideas, beliefs, conceptions of genre, …) (St Jerome)
  - Bibliographical evidence
  - Historical, physical evidence (ink, handwritten, watermarks, multispectral imaging)
Non-Traditional Authorship Attribution

- Stylometry (fingerprint)
  Computer Science & Statistics provide a quantitative tool
  - Single measure
  - Multivariate statistics
  - Distance-Based (similarity-based)
  - Machine Learning

“when there are very many candidate authors, similarity-based methods are more appropriate than machine-learning methods.”

Notation

- Word type: distinct forms
- Word token: number of forms (« I saw a man with a saw »)
- Lemma: headword, entry in the dictionary
- V: vocabulary used (word type)
- |V|: number of distinct word types
- V_r: vocabulary of terms appearing r times
- Hapax (hapax legomenon): word type appearing once
- |V_1|: number of hapax
- POS: Part-Of-Speech
- C: the corpus

Authorship Attribution

- Overview
  - Problem & Context
  - Examples
    - A single measurement
    - Multivariate analysis (restricted set of terms)
    - Distance-based approaches (ad hoc)

Notation

- tf_\text{j}{i}: absolute frequency of term i in text j (0, 1, ...)
- rtf_\text{j}{i}: relative frequency of term i in text j (0 ≤ rtf_\text{j}{i} ≤ 1)
- df_\text{i}: document frequency (number of texts with term i)
- a_\text{j}: j\text{th author}, j = 1, 2, ..., r
- m: number of selected features
- n: number of tokens in the corpus
- n_\text{j}: number of tokens of the j\text{th} text
  (or the size of the j\text{th} author profile)
- r: number of possible author
Classical Examples

The Federalist Papers
Set of 85 essays written by Publius

In fact three possible authors:
A. Hamilton
J. Madison
J. Jay

Who wrote what? (Mosteller & Wallace, 1964)

Federalist Papers: Zipf’s law

Vocabulary
7,860 word types (= |V|)
2,842 hapax (36%) (= |V1|)
1,176 dis legomenon (= |V2|) both 51.1%

Size
167,190 tokens (= n)
123,869 Hamilton (74%)
43,521 Madison (26%)

10 most freq. -> 35.7%
50 most freq. -> 54.9%

Federalist Papers: By Author

<table>
<thead>
<tr>
<th>Rank</th>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>the</td>
<td>14,200</td>
</tr>
<tr>
<td>2</td>
<td>,</td>
<td>10,333</td>
</tr>
<tr>
<td>3</td>
<td>of</td>
<td>9,467</td>
</tr>
<tr>
<td>4</td>
<td>to</td>
<td>5,751</td>
</tr>
<tr>
<td>5</td>
<td>,</td>
<td>4,065</td>
</tr>
<tr>
<td>6</td>
<td>and</td>
<td>3,849</td>
</tr>
<tr>
<td>7</td>
<td>in</td>
<td>3,591</td>
</tr>
<tr>
<td>8</td>
<td>a</td>
<td>3,247</td>
</tr>
<tr>
<td>9</td>
<td>be</td>
<td>3,025</td>
</tr>
<tr>
<td>10</td>
<td>that</td>
<td>2,221</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>the</td>
<td>10,293</td>
</tr>
<tr>
<td>2</td>
<td>,</td>
<td>7,508</td>
</tr>
<tr>
<td>3</td>
<td>of</td>
<td>7,149</td>
</tr>
<tr>
<td>4</td>
<td>to</td>
<td>4,498</td>
</tr>
<tr>
<td>5</td>
<td>,</td>
<td>2,998</td>
</tr>
<tr>
<td>6</td>
<td>in</td>
<td>2,782</td>
</tr>
<tr>
<td>7</td>
<td>and</td>
<td>2,681</td>
</tr>
<tr>
<td>8</td>
<td>a</td>
<td>2,476</td>
</tr>
<tr>
<td>9</td>
<td>be</td>
<td>2,270</td>
</tr>
<tr>
<td>10</td>
<td>that</td>
<td>1,679</td>
</tr>
</tbody>
</table>

The Federalist Papers (Mosteller and Wallace, 1964)

A series of newspapers articles published in 1787-88 with the aim of promoting the ratification of the new US constitution. Papers written under the pseudonym “Publius”

Some are of known (and in some cases joint) authorship but others are disputed

Written by three authors, Jay (5), Hamilton (51) and Madison (14), three by Hamilton & Madison, 12 uncertain.

Pioneering stylometric methods were famously used by Mosteller and Wallace in the early 1960s

It is now considered as settled

The Federalist Papers present a difficult but solvable test case, and are seen as a benchmark to test new ideas
Hidden Questions: Tokenization

What is a word for you? And for the computer?
- Examples
  - Richard Brown, 45-year old, is painting in New York
  - I'll send you Paul's book
  - John was prime minister to Henry VIII., permitting
    a final "take-it-or-leave-it" offer.
  - Database system in the U.S.A.
  - data base system in the US
  - data-base system in the U.S.
  - C|net, Micro$oft, and the IBM360, IBM-360, ...

Sequence of letters and digits?

Punctuation Marks?

The full stop (.) as a sentence length indicator.

“There is a strong personal element in the way people punctuate their writing. I know one novelist who puts commas in wherever possible. He writes sentences like this:

Fortunately, the bus was on time, so Sheema wasn’t late for the concert.

I know another who leaves them out whenever he can. He writes sentences like this:

Fortunately the bus was on time so Sheema wasn’t late for the concert.”


Classical Examples

- Did Shakespeare write all of his plays?
  - Various authors including Bacon and Marlowe are said to have written parts or all of several plays
  - “Shakespeare” may even be a nom-de-plume for a group of writers?
- Plays written by more than one author
  - Edward III – Shakespeare? & Kyd?
  - Two Noble Kinsmen – Shakespeare & Fletcher
  - Titus Andronicus – Shakespeare & Peele?
  - Henry VIII – Shakespeare & Fletcher?
  - Timon of Athens - Shakespeare & Fletcher?

Classical Examples

- The debate Molière vs. Corneille?
  Jean Baptiste Poquelin (1622-1673)
  Pierre Corneille (1606-1684)
- *Psyché* (1671), both are authors
- Plays (comedies) from 1658
- Corneille needs money, well-known for his dramas (but cannot write comedies, and inferior genre)
- Pierre Louys (1919) (and Voltaire) indicates that Corneille was the real author based on the rhythmus, versification.


Authorship Attribution

- Overview
  - Problem & Context
  - Examples
  - A single measurement
  - Multivariate analysis (restricted set of terms)
  - Distance-based approaches (ad hoc)

Single Measurement

- Letter counts
- Word length
- Sentence length, too obvious and easy to manipulate
- Frequencies of letter pairs, strangely successful (n-gram)
- Distribution of words of a given length (in syllables), especially relative frequencies
- And what about the vocabulary growth and richness?
- Simple, but really effective?

Vocabulary Richness

- Based on the idea that author’s vocabulary is more or less constant
- Various measures
  - Type-token ratio
  - Simpson’s index (the chance that two word arbitrarily chosen from text will be the same)
  - Yule’s K (occurrence of a given word can be modelled as a Poisson distribution)
- But not stable for AA (Hoover, 2003), (Baayen, 2008)
"What disturb me in Shakespeare's plays is the over-used of the letter "o". I can live with a lot of "e" or "i", but not a lot of "o". So, yes clearly, I prefer reading Marlowe."

T. Merriam reports "of counting the letters in the 43 plays was the implausible discovery that the letter 'o' differentiates Marlowe and Shakespeare plays to an extent well in excess of chance" (used also the letter 'a')

- Frequency less than 0.0078, 6 plays of Marlowe
- Frequency greater than 0.0078, 36 plays of Shakespeare


Authorship Attribution

- Overview
  - Problem & Context
  - Examples
  - A single measurement
  - **Multivariate analysis** (restricted set of terms)
  - Distance-based approaches (**ad hoc**)

Multivariate Analysis

- Thanks to computers it is now possible to collect large numbers of different measurements, of a variety of features
- Variants of multivariate analysis
  - Principal components analysis (PCA)
  - Correspondence analysis (CA)
  - Cluster analysis
- Tools to **visualize** the data (better than reading a lexical table)
- Variables = features = word types or lemmas
- Objects = text excerpts
PCA: explains the data using fewer variables

Explaining the max. of the variability

A cloud of birds in 3D \(\rightarrow\) 2D (\(\rightarrow\) 1D)

See (Binongo & Smith, 1999) (Craig & Kinney, 2009)

---

Principal Component Analysis

- PCA generate a smaller ordered set of new variables (principal components) uncorrelated (latent factors)
- "principal components" are computed by calculating the correlations between all the terms, then grouping them into sets that show the most correspondence

We will define a projection plane (defined by the lines \(\Delta_1\) and \(\Delta_2\), perpendicular (no correlation)) to represent the objects \((e_i, e_j)\) and conserving the real distance \(d(e_i, e_j)\). Focus: dispersion

---

PCA: Input

- Small lexical table with 4 texts (authors) and 5 words

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>15</td>
<td>30</td>
<td>5</td>
</tr>
<tr>
<td>a</td>
<td>9</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>I</td>
<td>2</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>my</td>
<td>2</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>of</td>
<td>2</td>
<td>6</td>
<td>4</td>
</tr>
</tbody>
</table>

- B is "twice" A
- A and B have more determiners "the" and "a" than other words
- C used more "I" and "my"
- D is the style of the average
- Visualize this data (apply PCA (normalize))
PCA (Federalist Papers)

- The first two components explain 9.0% + 8.3% = 17.3% of the total variance (not a lot).
- In general, Hamilton’s papers on the left, Madison paper’s on the right, and disputed papers more closer to Madison’s area (see next slide).
- In the horizontal axis, on the right, we have articles with on, by, government, and people. On the left, we can find papers with to, an, would, this, power, and if.
- In the vertical axis (up), we have more frequently be, that, it, will, government, may. In the bottom direction, we have papers using more have, with, been, and has.
Visual and real distance. Having two points \( f_i \) and \( f_k \) close together in the PC1 and PC2 plan does not mean that the corresponding \( e_i \) and \( e_k \) points are also close together.

**PCA could be useful in your context,**
- to visualize
- to synthesize your data!

**Nearest Neighbour**
- But we can imagine a simple attribution method…
  Find the text / author profile having the smallest distance with the representation of a disputed text,
- Testing instance \( Q \):
  - Compute similarity between \( Q \) and all other texts / author profiles
  - Assign \( Q \) the category of the most similar example (1-NN)
- Simple to apply. The system does not really learn the different styles.
- Nearest neighbor method depends on a distance measure

**Authorship Attribution**
- Overview
  - Problem & Context
  - Examples
  - A single measurement
  - Multivariate analysis (restricted set of terms)
- **Distance-based approaches**
  - Delta
  - Chi-square
  - Kullback-Leibler
  - Vocabulary
  - Labbé’s distance
Based on the m most (m = 50, ...) frequent words (+ POS for some types such as to, in)

"frequency-hierarchy for the most common words in a large group of suitable texts" (p. 269)

Compute a Z-score value for each word

for each word type w_i, i = 1, ..., in a text D_j
compute the relative frequency \( r_{ij} \) (in %)
\( \mu_i \) mean in the reference corpus
\( \sigma_i \) standard deviation

\[
Z_{score}(w_{ij}) = \frac{r_{ij} - \mu_i}{\sigma_i}
\]


Top 50 Most Frequent Words
(Hamilton & Madison)

the it for been other
, is not on if
of which will government at
to as with may any
. by from state than
and ; their all more
in this an power no
a would are its there
be have they but them
that or states has one

Burrows' Delta

First compute the absolute frequencies in our example.

<table>
<thead>
<tr>
<th></th>
<th>H59</th>
<th>H60</th>
<th>H61</th>
<th>H62</th>
<th>M37</th>
<th>M38</th>
<th>M47</th>
<th>M48</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>177</td>
<td>224</td>
<td>152</td>
<td>220</td>
<td>230</td>
<td>273</td>
<td>328</td>
<td>167</td>
</tr>
<tr>
<td>.</td>
<td>133</td>
<td>152</td>
<td>104</td>
<td>142</td>
<td>192</td>
<td>234</td>
<td>219</td>
<td>157</td>
</tr>
<tr>
<td>of</td>
<td>112</td>
<td>145</td>
<td>100</td>
<td>130</td>
<td>159</td>
<td>189</td>
<td>187</td>
<td>98</td>
</tr>
<tr>
<td>to</td>
<td>73</td>
<td>87</td>
<td>61</td>
<td>84</td>
<td>84</td>
<td>117</td>
<td>64</td>
<td>54</td>
</tr>
<tr>
<td>.</td>
<td>45</td>
<td>47</td>
<td>32</td>
<td>47</td>
<td>75</td>
<td>95</td>
<td>85</td>
<td>56</td>
</tr>
<tr>
<td>in</td>
<td>62</td>
<td>79</td>
<td>47</td>
<td>51</td>
<td>63</td>
<td>62</td>
<td>62</td>
<td>46</td>
</tr>
<tr>
<td>and</td>
<td>34</td>
<td>36</td>
<td>25</td>
<td>37</td>
<td>101</td>
<td>95</td>
<td>87</td>
<td>51</td>
</tr>
<tr>
<td>a</td>
<td>49</td>
<td>53</td>
<td>35</td>
<td>44</td>
<td>57</td>
<td>92</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>size</td>
<td>636</td>
<td>770</td>
<td>521</td>
<td>703</td>
<td>904</td>
<td>1065</td>
<td>1032</td>
<td>629</td>
</tr>
</tbody>
</table>

Then the relative frequencies, and the mean and stdev

<table>
<thead>
<tr>
<th></th>
<th>H59</th>
<th>H60</th>
<th>H61</th>
<th>H62</th>
<th>M37</th>
<th>M38</th>
<th>M47</th>
<th>M48</th>
<th>( \mu )</th>
<th>( \sigma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>0.278</td>
<td>0.291</td>
<td>0.292</td>
<td>0.313</td>
<td>0.254</td>
<td>0.256</td>
<td>0.318</td>
<td>0.266</td>
<td>0.283</td>
<td>0.024</td>
</tr>
<tr>
<td>.</td>
<td>0.209</td>
<td>0.197</td>
<td>0.191</td>
<td>0.212</td>
<td>0.220</td>
<td>0.212</td>
<td>0.250</td>
<td>0.211</td>
<td>0.019</td>
<td></td>
</tr>
<tr>
<td>of</td>
<td>0.176</td>
<td>0.188</td>
<td>0.192</td>
<td>0.185</td>
<td>0.177</td>
<td>0.177</td>
<td>0.181</td>
<td>0.156</td>
<td>0.179</td>
<td>0.011</td>
</tr>
<tr>
<td>to</td>
<td>0.115</td>
<td>0.113</td>
<td>0.117</td>
<td>0.119</td>
<td>0.093</td>
<td>0.110</td>
<td>0.062</td>
<td>0.086</td>
<td>0.102</td>
<td>0.020</td>
</tr>
<tr>
<td>.</td>
<td>0.071</td>
<td>0.061</td>
<td>0.061</td>
<td>0.067</td>
<td>0.083</td>
<td>0.089</td>
<td>0.082</td>
<td>0.089</td>
<td>0.075</td>
<td>0.012</td>
</tr>
<tr>
<td>in</td>
<td>0.097</td>
<td>0.103</td>
<td>0.090</td>
<td>0.073</td>
<td>0.070</td>
<td>0.058</td>
<td>0.060</td>
<td>0.073</td>
<td>0.078</td>
<td>0.017</td>
</tr>
<tr>
<td>and</td>
<td>0.053</td>
<td>0.047</td>
<td>0.048</td>
<td>0.053</td>
<td>0.112</td>
<td>0.089</td>
<td>0.084</td>
<td>0.081</td>
<td>0.071</td>
<td>0.024</td>
</tr>
<tr>
<td>a</td>
<td>0.077</td>
<td>0.069</td>
<td>0.067</td>
<td>0.063</td>
<td>0.063</td>
<td>0.086</td>
<td>0.034</td>
<td>0.056</td>
<td>0.064</td>
<td>0.015</td>
</tr>
</tbody>
</table>
Burrows' Delta

Third, the author's profiles, absolute, relative and Z-score

<table>
<thead>
<tr>
<th>H</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>773</td>
</tr>
<tr>
<td>,</td>
<td>523</td>
</tr>
<tr>
<td>of</td>
<td>487</td>
</tr>
<tr>
<td>to</td>
<td>305</td>
</tr>
<tr>
<td>,</td>
<td>171</td>
</tr>
<tr>
<td>in</td>
<td>239</td>
</tr>
<tr>
<td>and</td>
<td>132</td>
</tr>
<tr>
<td>a</td>
<td>181</td>
</tr>
<tr>
<td>size</td>
<td>2630</td>
</tr>
</tbody>
</table>

Burrows' Delta

<table>
<thead>
<tr>
<th>H</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>0.294</td>
</tr>
<tr>
<td>,</td>
<td>0.199</td>
</tr>
<tr>
<td>of</td>
<td>0.185</td>
</tr>
<tr>
<td>to</td>
<td>0.116</td>
</tr>
<tr>
<td>,</td>
<td>0.065</td>
</tr>
<tr>
<td>in</td>
<td>0.091</td>
</tr>
<tr>
<td>and</td>
<td>0.050</td>
</tr>
<tr>
<td>a</td>
<td>0.069</td>
</tr>
</tbody>
</table>

Burrows' Delta

<table>
<thead>
<tr>
<th>H</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>0.430</td>
</tr>
<tr>
<td>,</td>
<td>-0.688</td>
</tr>
<tr>
<td>of</td>
<td>0.563</td>
</tr>
<tr>
<td>to</td>
<td>0.702</td>
</tr>
<tr>
<td>,</td>
<td>-0.883</td>
</tr>
<tr>
<td>in</td>
<td>0.768</td>
</tr>
<tr>
<td>and</td>
<td>-0.863</td>
</tr>
<tr>
<td>a</td>
<td>0.290</td>
</tr>
<tr>
<td>size</td>
<td>2630</td>
</tr>
</tbody>
</table>

Burrows' Delta

Four, Delta distance using the Z-score

<table>
<thead>
<tr>
<th>H</th>
<th>M</th>
<th>D54</th>
<th>D55</th>
<th>D56</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>0.430</td>
<td>-0.354</td>
<td>0.297</td>
<td>0.573</td>
</tr>
<tr>
<td>,</td>
<td>-0.688</td>
<td>0.530</td>
<td>0.211</td>
<td>0.002</td>
</tr>
<tr>
<td>of</td>
<td>0.563</td>
<td>-0.414</td>
<td>0.169</td>
<td>0.893</td>
</tr>
<tr>
<td>to</td>
<td>-0.883</td>
<td>0.865</td>
<td>0.085</td>
<td>0.769</td>
</tr>
<tr>
<td>in</td>
<td>0.768</td>
<td>-0.823</td>
<td>0.095</td>
<td>0.999</td>
</tr>
<tr>
<td>and</td>
<td>-0.863</td>
<td>0.880</td>
<td>0.055</td>
<td>0.646</td>
</tr>
<tr>
<td>a</td>
<td>0.290</td>
<td>-0.258</td>
<td>0.051</td>
<td>0.859</td>
</tr>
</tbody>
</table>

Burrows' Delta

The Delta Distance

<table>
<thead>
<tr>
<th>H</th>
<th>D54</th>
<th>D55</th>
<th>D56</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>0.870</td>
<td>0.876</td>
<td>1.770</td>
</tr>
<tr>
<td>,</td>
<td>0.750</td>
<td>0.822</td>
<td>0.989</td>
</tr>
</tbody>
</table>

Burrows' Delta

- Distance between two texts / profiles D (doubtful) and D’ (known)
- If Δ is small, D and D’ are written by the same author.

\[ \Delta (D, D') = \frac{1}{m} \sum_{i} |Z(w_{ij}) - Z(w_{ij}')| \]

- Modification suggested (Hoover, 2004)
  - n must be greater than 150 (e.g., 800 – 4,000)
  - ignoring personal pronouns
  - culling at 70% (words for which a single text supplies more than 70% of the occurrences)

Selection in Delta

Selection the $k$ most frequent forms, we can find
1. Referential: articles and pronouns
2. Temporal / modal: auxiliary verbs and some adverbs
3. Connective: conjunctions, prepositions, relative pronouns
4. Modificatory: adjectives, adverbs

Chi-Square Approach

- How many terms do we need to take account?
- A set of terms (very frequent) defined a priori?
- Define $m$ number of features, as a $k$-limit, meaning that the selected terms must appear in at least $k$ documents writing by each author (df-based)
- Low $k$ value means a larger number of terms
- Large $k$ value implies a smaller set of features (limit: the total number of articles writing by a single author)
- Guarantee that each cell is not empty (smoothing is not needed)

Kullback-Leibler Divergence

- We can define a priori a set of very frequent words appearing in a given language
- Stopword list in information retrieval (search engines)
  - Zhao & Zobel: 369 terms (e.g., the, we, is, not, became, …)
  - Italian language: 399 terms (e.g., del, essi, non, volta,…)
- For each word, we can estimate the occurrence probability $q(t)$ and $a_j(t)$ for term $t$, in Q or $A_j$
- Compute the distance between the distribution in the query text Q and the distribution obtained from each author profile


Kullback-Leibler Divergence

- Distance between Q query text and $A_j$ author profile of author $j$

$$KLD(Q \mid \mid A_j) = \sum_{i=1}^{m} q(t_i) \cdot \log_2 \frac{q(t_i)}{a_j(t_i)}$$

where $m$ number of features, $q(t_i)$ and $a_j(t_i)$ the occurrence probability for term $t_i$ in Q or $A_j$.

- We assume that $0 \log_2(0/a) = 0$, and $q(0) = \infty$.

- Low KLD value indicates probable author


Kullback-Leibler Divergence

- How to estimate $q(t_i)$ (similar for $a_j(t_i)$) ?

$$q(t_i) = \frac{tf_{iq}}{n_q}$$

$$q(t_i) = \frac{tf_{iq} + 1}{n_q + \lambda \cdot |V|} \quad \text{or} \quad q(t_i) = \frac{tf_{iq} + \lambda}{n_q + \lambda \cdot |V|}$$

$$q(t_i) = \frac{tf_{iq}}{n_q + \mu} + \frac{\mu}{\mu + n_q} \cdot q_B(t_i)$$

With $q_B(t_i)$ the probability of term $t_i$ in the background model

Z Score

Why limited ourselves to functional words?
The vocabulary could be different between two authors (personal, genre, social, region).

1. have a bath, bike bicycle, luncheon, sick, England, Scotch, sofa.
2. take a bath, cycle, dinner, ill, Britain, Scottish, settee.

Two authors may use the same words with different intensity, one may over-used a set of forms while the second may under-used them.

Idea: Define the vocabulary specific to an author (genre, type, …) (Ssvoy, 2012)

Variant: see (Pauli & Tuzzi, 2009)
Z Score

Example
Splitting the whole corpus into two parts.

- Size of the corpus: \( n = 15 \).
- Subcorpus: 3 (or \( \frac{1}{5} = 0.2 \)).

Number of G in total: 4. In the subcorpus: 2.

Expected frequency in the subcorpus: \( 0.2 \times 4 = 0.8 \).

Observed frequency in the subcorpus: 2. Thus G is over-used in the subcorpus!

Contingency Table

The word “\( \omega \)” in the sub-corpus and in the rest \( C- \) (\( C = C- \cup \text{Sub-corpus} \))

<table>
<thead>
<tr>
<th>Sub-corpus</th>
<th>C-</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \omega )</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>not “( \omega )”</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td>( a+c )</td>
<td>( b+d )</td>
<td>( n = a+b+c+d )</td>
</tr>
</tbody>
</table>

- \( \eta_{\text{sub-corpus}} = a + c \)
- \( \text{Prob}[u] = \frac{(a+b)}{n} \)
- \( \text{Prob[\text{word in Sub-corpus}]} = \frac{(a+c)}{n} \)

Z Score

- We have a Z score for each term \( t_i \) in a subcorpus \( D_j \)

\[
Z \text{ score}(t_{ij}) = \frac{a - \text{Prob}[t_i] \cdot n_j)}{\sqrt{n_j \cdot \text{Prob}[t_i] \cdot (1 - \text{Prob}[t_i])}}
\]

- When comparing two texts, considering all Z scores

\[
\text{Dist}(D_j, D_k) = \frac{1}{m} \sum_{i} (Z \text{ score}(t_{ij}) - Z \text{ score}(t_{ik}))^2
\]

Z Score: Example

The word “upon” in Hamilton’s papers

<table>
<thead>
<tr>
<th>Hamilton</th>
<th>rest</th>
<th>Federalist Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>“upon”</td>
<td>370</td>
<td>10</td>
</tr>
<tr>
<td>not “upon”</td>
<td>73,475</td>
<td>41,882</td>
</tr>
</tbody>
</table>
|          | 73,845 | 41,892          | 115,357

- \( \text{Prob}[t_i] = \text{Prob[“upon”]} = 380 / 115,737 = 0.003283 \).
- \( \eta_i = 73,845 \quad a = 370 \)
- We expect in Hamilton’s subcorpus: \( \eta_i \cdot \text{Prob}[t_i] = 242.46 \)
- Z score (“upon” in Hamilton) = 8.2046
Z Score: Example

The word "on" in Hamilton's articles

<table>
<thead>
<tr>
<th></th>
<th>Hamilton</th>
<th>rest</th>
<th>Federalist Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;on&quot;</td>
<td>374</td>
<td>485</td>
<td>859</td>
</tr>
<tr>
<td>not &quot;on&quot;</td>
<td>73,471</td>
<td>41,407</td>
<td>114,878</td>
</tr>
<tr>
<td></td>
<td>73,845</td>
<td>41,892</td>
<td>115,737</td>
</tr>
</tbody>
</table>

- $\text{Prob}[t] = \text{Prob}[^{\text{"on"}}] = 859 / 115,737 = 0.007422.$
- $\eta = 73,845 \quad a = 374$
- We expect in Hamilton's subcorpus: $\eta \cdot \text{Prob}[t] = 548.08$
- Z score ("on" in Hamilton) $\approx -7.46$

Intertextual Distance (Labbé, 2007)

- Based on the vocabulary, how can we select part of it?
- Most frequent: Burrows
- Used by every author, every time: Grieve
- Specific vocabulary: Savoy
- Why not all words? Labbé
- The vocabulary choice depends on the subject, genre, epoch, and author
- Define an intertextual distance based on the word types used and their frequencies between two texts.
  - But texts with the same genre and epoch.


Z Score

<table>
<thead>
<tr>
<th>Over-used terms</th>
<th>Hamilton</th>
<th>Madison</th>
</tr>
</thead>
<tbody>
<tr>
<td>upon powers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>would confession</td>
<td></td>
<td></td>
</tr>
<tr>
<td>to department</td>
<td></td>
<td></td>
</tr>
<tr>
<td>there on</td>
<td></td>
<td></td>
</tr>
<tr>
<td>courts congress</td>
<td></td>
<td></td>
</tr>
<tr>
<td>kind and</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Under-used terms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>on upon</td>
<td></td>
<td></td>
</tr>
<tr>
<td>representatives there</td>
<td></td>
<td></td>
</tr>
<tr>
<td>by would</td>
<td></td>
<td></td>
</tr>
<tr>
<td>department to</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Intertextual Distance (Labbé, 2007)

- We define $\text{tf}^i = \text{frequency of word type } i \text{ in text } A$
  - $n_A = \text{size (number of tokens) of text } A$
  - $V_A = \text{vocabulary of text } A$
  - $n_A = \sum_{i \in V_A} \text{tf}_i A$

- Distance $D(A, B)$ between Text A and Text B (similar size)
  - $D(A, B) = \sum_{i \in (V_A \cup V_B)} |\text{tf}_i A - \text{tf}_i B|$ with $n_A = n_B$

- $D(A, B) = 0$ both texts use the same words with the same frequencies
- Otherwise $D(A, B) > 0$ (lim: $n_A + n_B$)
  - the number of tokens that differ
Intertextual Distance (Labbé, 2007)

- When both sizes differ (assuming \( n_A < n_B \)) we reduced the \( tf \) of term \( i \) in \( B \) as

\[
\text{tf}^*_{iB} = \frac{\text{tf}_{iB}}{n_B}
\]

- Problem when the two corpora have different size

\[
D_{rel}(A, B) = \sum_{i \in (V_A \cup V_B)} \frac{|\text{tf}_{iA} - \text{tf}^*_{iB}|}{2 \cdot n_A}
\]

Example: Two texts with the same size \((n_A = n_B = 7)\)

<table>
<thead>
<tr>
<th>Text A</th>
<th>Text B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes, we can, and yes we can.</td>
<td>Yes, we still can.</td>
</tr>
</tbody>
</table>

\[
D(A, B) = (0 + 0 + 1 + 1 + 1 + 1 + 1) = 4
\]

\[
D_{rel}(A, B) = \frac{4}{14} = 0.286
\]

Example: Two texts with the different sizes \((n_A = 4, n_B = 8)\)

<table>
<thead>
<tr>
<th>Text A</th>
<th>Text B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes, we can scan.</td>
<td>Yes, we can, and we can do more.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \text{tf} ) ( \text{tf}^* )</th>
<th>( \text{tf} ) ( \text{tf}^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes: 1</td>
<td>yes: 1</td>
</tr>
<tr>
<td>we: 1</td>
<td>we: 1</td>
</tr>
<tr>
<td>can: 1</td>
<td>can: 1</td>
</tr>
<tr>
<td>scan: 1</td>
<td>and: 1</td>
</tr>
</tbody>
</table>

\[
D_{rel}(A, B) = \frac{3}{8} = 0.375
\]

Intertextual distance take account of all word types with their frequencies

- Largest impact is coming from word types with low frequencies \(< 5\)
- Difference in text size max: 1:8
- Min number of tokens: 10,000
- Can be used to generate a matrix distance, then a clustering or tree
- Variant: See (Cortelazzo et al., 2013)
Intertextual Distance (Labbé, 2007)

- Decision according to the $D_{rel}(Q,A_j)$

same author for $Q$ and $A_j$

mean - $\sigma$ mean

mean + $\sigma$

different authors for $Q$ and $A_j$

Intertextual Distance (Labbé, 2007)

Federalist Papers

$D_{rel}(A,B)$ & Clustering

Newspapers Corpora

Glasgow Herald (1995)

La Stampa (Italy) (1994)
Newspapers Corpora

We have selected 20 authors (journalists) from
Glasgow Herald (5,408 articles)
La Stampa (Italy) (4,326 articles)

- From the GH, we have between 30 to 433 articles from each possible author
  word tokens mean: 724.9 (min: 44, max: 4,414), median: 668, standard deviation: 393.2
- In La Stampa, we can find between 52 to a maximum of 434 articles from each author
  word tokens mean: 777.1 (min: 60; max: 2,935), median: 721; standard deviation: 332.6

Evaluation

Micro-averaging (each author has the same weight) over 20 authors
Appropriate parameter values is important!

<table>
<thead>
<tr>
<th></th>
<th>Glasgow</th>
<th>La Stampa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delta, 40 word types</td>
<td>43.53%</td>
<td>43.44%</td>
</tr>
<tr>
<td>Delta, 150 word types</td>
<td>58.54%</td>
<td>63.62%</td>
</tr>
<tr>
<td>Delta, 200 word types</td>
<td>59.91%</td>
<td>68.70%</td>
</tr>
<tr>
<td>Delta, 400 word types</td>
<td>63.70%</td>
<td>76.07%</td>
</tr>
<tr>
<td>Delta, 800 word types</td>
<td>54.81%</td>
<td>66.30%</td>
</tr>
<tr>
<td>Delta, 400 word types - PP</td>
<td>60.63%</td>
<td>74.90%</td>
</tr>
<tr>
<td>Delta, 600 word types - PP</td>
<td>61.32%</td>
<td>74.78%</td>
</tr>
<tr>
<td>Delta, 800 word types - PP</td>
<td>53.92%</td>
<td>67.73%</td>
</tr>
</tbody>
</table>

Evaluation

Micro-averaging over 20 possible authors
Z score: Terms having a frequency (in C) $tf_c > 10$
appearing in $df_i > 2$, and used by more than one author $df_A > 1$
GH: 2,511 terms, La Stampa: 9,825 terms

<table>
<thead>
<tr>
<th></th>
<th>Glasgow</th>
<th>La Stampa</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$, 2-limit, 653/720 terms</td>
<td>65.26%</td>
<td>68.28%</td>
</tr>
<tr>
<td>$\chi^2$, 5-limit, 289/333 terms</td>
<td>62.39%</td>
<td>63.49%</td>
</tr>
<tr>
<td>$\chi^2$, 10-limit, 149/203 terms</td>
<td>59.39%</td>
<td>66.07%</td>
</tr>
<tr>
<td>$\chi^2$, 20-limit, 52/106 terms</td>
<td>52.27%</td>
<td>62.83%</td>
</tr>
<tr>
<td>$\chi^2$, 30-limit, 15/71 terms</td>
<td>40.03%</td>
<td>62.51%</td>
</tr>
<tr>
<td>$\chi^2$, 40-limit, -/42 terms</td>
<td>n/a</td>
<td>59.78%</td>
</tr>
<tr>
<td>$\chi^2$, 50-limit, -/30 terms</td>
<td>n/a</td>
<td>56.26%</td>
</tr>
<tr>
<td>$\chi^2$, 52-limit, -/20 terms</td>
<td>n/a</td>
<td>49.24%</td>
</tr>
</tbody>
</table>
Hidden Questions / Problems

- Split clearly between a training set and a test set
- Each model has its own limits
- Size of the (disputed / training) texts
  100 tokens to 10,000 tokens
  Better performance
  with long texts, long profiles, few authors
- (Un)Balanced set in generating the author’s profiles
- Type of text (e.g., dialogue, descriptive, narrative)
- The style may change during the author’s life
- Style related to a given character (or set of characters)
  for a given author

- Der Teufel liegt im Detail

Conclusion

- Authorship attribution
  - The result of computational linguistics are always
    matters of probability, not certainty. …After all, we are
    dealing with writers who are at liberty to imitate each
    others, to try new styles, and to write differently for a
    particular occasion or in a new genre, …”
  (Craig & Kinney, 2009, p. 24-25)
- L’Aquila quake: Italian scientists guilty
  - Explain the decision with stylistic elements

Next steps
- Consider other representation than isolated words
  n-gram of characters, n-gram of words,
  POS, n-gram of POS
- Other languages
- Other paradigm (machine learning) to promote better
  classifier(?)
- Author profiling
- Other medium

Being in Padova...

Use your eyes…
Knowing the author of
this three paintings,
Are the last two painted
by the same author?
References

Books, overview

Article, overview

Articles of Delta
References
Articles of various approaches

Articles on Labbé’s methods

Other articles

Forensic Linguistics

DEAR BILL,
I SUPPOSE YOU THOUGHT I WOULD FORGET BUT YOU ARE WRONG HOW COULD I FORGET A RAT LIKE YOU. I HAVE SENT A LETTER WITH ALL YOUR PAST DETAILS TO THE PRESIDENT. ALL YOUR DEBTS AND PAST MISDEMEANOURS. IF YOU DON’T RESIGN FROM THE COUNCIL IMMEDIATELY THE PRESS WILL PRINT A LIST OF ALL YOUR DEBTS BOTH LOCALLY AND NATIONALLY… YOU MIGHT BE ABLE TO FOOL SOME PEOPLE BUT NOT ME. YOUR FORGET I HAVE KNOWN YOU FOR ALL OF YOUR LIFE.
Admission in Court (US)

1. Knowledge and stature: the exert must have sufficient knowledge of the subject.
2. Testing: the technique must be empirically tested.
3. Peer review: subjected to a peer review.
4. Scientific method: the error rate is known.
5. Straightforwardness: the technique can be explained with clarity and simplicity.

Forensic Linguistic (Olssen, 2008)

Example with the *Federalist*

- Spelling variation: *while* (Hamilton) vs. *whilst* (Madison)
- In the vocabulary used only by one: Hamilton: destruction, offensive, defensive, contribute
  Madison: violence, fortune, although
- Vocabulary used more frequently by one: *considerable* (13 Hamilton, 4 Madison)
  *voice* (1, 8)
  *again* (1, 7)
  *language* (2, 10)

Evaluation: Federalist Papers

<table>
<thead>
<tr>
<th></th>
<th>Default</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delta, 40 word types</td>
<td>10 / 12</td>
<td>#55, #56</td>
</tr>
<tr>
<td>Delta, 50 word types</td>
<td>9 / 12</td>
<td>#55, #56, #63</td>
</tr>
<tr>
<td>Delta, 100 word types</td>
<td>10 / 12</td>
<td>#55, #56</td>
</tr>
<tr>
<td>Delta, 150 word types</td>
<td>11 / 12</td>
<td>#56</td>
</tr>
<tr>
<td>Delta, 200 word types</td>
<td>9 / 12</td>
<td>#50, #56, #57</td>
</tr>
<tr>
<td>KLD, Zhao</td>
<td>9 / 12</td>
<td>#49, #55, #57</td>
</tr>
<tr>
<td>KLD, Hughes</td>
<td>12 / 12</td>
<td></td>
</tr>
<tr>
<td>Intertextual distance</td>
<td>No assign.</td>
<td></td>
</tr>
<tr>
<td>Intertextual &amp; Clustering</td>
<td>12 / 12</td>
<td></td>
</tr>
</tbody>
</table>

Z Score: Example

The word “Bush” in McCain’s speeches in 2008 (= D_j) vs. all other US electoral speeches

<table>
<thead>
<tr>
<th></th>
<th>McCain’08</th>
<th>rest</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Bush&quot;</td>
<td></td>
<td>26</td>
<td>396</td>
</tr>
<tr>
<td>not &quot;Bush&quot;</td>
<td>154,339</td>
<td>474,331</td>
<td>628,670</td>
</tr>
<tr>
<td></td>
<td>154,365</td>
<td>474,729</td>
<td>629,094</td>
</tr>
</tbody>
</table>

- $\Pr(t) = \Pr(\text{"Bush" in C}) = 424 / 629,094 = 0.000674$.
- $\eta_i = 154,365 \quad a = 26$
- We expect in McCain’08 (= D_j): $\eta_i \Pr(t) = 104.04$
- Z score ("Bush" in McCain’08) = -7.65
I and Obama

« I » (and me) is the prototypical stealth word. When a person uses a lot of « I », arrogant, self-confident
G. Will (Washington Post, June 7, 2009)
S. Fish (New York Times, June 7, 2009)
pointed out the Obama’s frequent use of « I »
(Pennebacker, 2011)

<table>
<thead>
<tr>
<th>State of the Union</th>
<th>Obama</th>
<th>Bush</th>
<th>Clinton</th>
</tr>
</thead>
<tbody>
<tr>
<td>“I”</td>
<td>1.13%</td>
<td>0.62%</td>
<td>5.76%</td>
</tr>
<tr>
<td>“we”</td>
<td>3.73%</td>
<td>2.81%</td>
<td>17.21%</td>
</tr>
</tbody>
</table>

Psychological Profile

We can establish the psychological profile of the writer according to four dimensions (MBTI indicator):
1. Extroversion vs. Introversion
   (social interaction vs. solitude)
2. Intuition – Sensing
   (prefers theoretical info vs. perceiving the info)
3. Thinking – Feeling
   (logical decision vs. decision according to subjective values)
4. Judgment – Perception
   (judgement according to my perceptions vs. don’t quickly jump to a conclusion)
   (Noecker, Ryan, Juola, LLC, 2013), (Pennebacker, 2011)

Age

- Older people tend to use more future tense
- Young people tend to use more past tense

- Upper class: more « we »
- Lower class: more « I »
(Pennebacker, 2011)

Why data quality matters
Why data quality matters

Zeta: Less Frequently Used Words

- Instead of focussing on very frequent words, focus on words used more frequently by a given author. E.g., Shakespeare uses more gentle, answer but less frequently brave, sure, hopes, or beseech.
- Split the texts into blocks (20,000 tokens), form a set of texts written by A, and a counter-set written by others (-A)
- Select word types having \( df^A \geq \delta \) (e.g., in 3 blocks) and word types must have \( df^-A \geq \delta \) (e.g., 3)
- Binary view: term present or not


First solution: for a given term \( t_i \), we can count:
- the number of texts (blocks) in which the term \( t_i \) appears (or \( df^i \))
- the number of texts (blocks) where it does not appear (or \( df^-i \))
- the ratio (\( df^i / df^-i \))

But we can include the fact that the author was A or B (\( df^A \) number of blocks written by A with term \( t_i \), \( T^A \) denotes the number of texts written by A)

\[
index(t_i, A, B) = \frac{|df^A_i|}{|T^A|} + \frac{|df^-B_i|}{|T^B|}
\]

If a term \( t_i \) appears in all and only in texts written by A, the index will be \( 1 + 1 = 2 \)
If a term \( t_i \) is used by both writers, and in all of their texts, the index will be \( 1 + 0 = 1 \)
If a term \( t_i \) is used by both writers in the same proportion (e.g., 30%), the index will be \( 0.3 + 0.7 = 1 \)
If a term \( t_i \) is used only by B (with a proportion of 20%), the index will be \( 0 + 0.8 = 0.8 \)
