Revelation of the author’s identity using machine learning and stylometry

Jan Rygl
rygl@fi.muni.cz
Mar 12, 2015

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
History and motivation

History

Motivation

- Antiquity: Homer, Demosthenes vs Anaximenes
- Jewish and Christian Bibles: Pentateuch
- England 1694 (end of pre-publication censorship): pseudonyms
- England 1887: the first algorithmic method
- England 1976: evidence in court
- present: analysis of anonymous documents in the Internet, mobiles, . . .

Revelation of the author’s identity using machine learning and stylometry
Authorship recognition methods

1. **ideological and thematic analysis**
   historical documents, literature

2. **documentary and factual evidence**
   inquisition in the Middle Ages, libraries

3. **language and stylistic analysis** – **stylometry**
   present
History and motivation

Authorship Verification

Definition
- decide if two documents were written by the same author (1v1)
- decide if a document was written by the signed author (1vN)

Examples
- The Shakespeare authorship question
- The verification of wills
History and motivation

Authorship Verification

The Shakespeare authorship question


- The first algorithmic analysis
- Calculating and comparing histograms of word lengths

Oxford, Bacon, Derby, Marlowe

Definition

- find out an author of a document
- candidate authors can be known

Examples

- False reviews
- Anonymous e-mails
The police falsify testimonies


Evidence in courts of law in Britain, U.S., Australia
**Definition**

Cluster documents or text paragraphs according to the authors.

**Examples**

- The Bible
- Analysis of anonymous documents
The Bible

K. Grayston and G. Herdan.  

Gustav Herdan, statistician and linguist:
- born 1897 in Brno
- author of *Quantitative linguistics*
- mathematical language laws, e.g. the dependence of the number of *distinct words* in a document as a function of the *document length*
History and motivation

Related fields

Compututional stylometry

- Online social networks: predicting age and gender
- Plagiarism: co-authorship
- Supportive authentication, biometrics (e.g. in e-learning)
- Native language prediction
- . . .
History and motivation

Public security
- Anonymous documents, threats, ...
- Ministry of the Interior of CR within the project VF20102014003

Research for Ministry of the Interior of CR
- authorship detection for Czech
- new author’s characteristics and adaptation of existing for flective free-word-order languages
- new techniques for “Internet documents”
- software Authorship Recognition Tool (ART)
1 History and motivation

2 Techniques

3 Results
Stylometry

**Definition**

Computational stylometry techniques that allow us to find out information about the authors of texts on the basis of an automatic linguistic analysis.

**Motivation**

Stylometry analysis is used for:
- Linguistic expertise
- Stylome: set of characteristic author’s features
- Machine learning: stylometric features ~ attributes for machine learning
Stylometry

Preprocessing

- document crawling
- text and meta data extraction (detect author’s label)
- text cleaning
  - deduplication
  - boilerplate removal
  - remove markup tags
- language and encoding detection
- tokenize

Revelation of the author’s identity using machine learning and stylometry
Stylometry

Preprocessing

- morphological analysis

je       byt       k5eAaImIp3nS
spor     spor      k1gInSc1
mezi     mezi      k7c7
Severem  sever     k1gInSc7

- syntactic analysis

15       ekonomiky  43      p
16       .          44      p
17       <CP>      20      p
18       <CLAUSE>  20      p
19       <CLAUSE>  20      p
Authorship recognition through stylometry

For each text:

1. preprocess text
2. count values of stylometric features (text is represented by a vector of feature values)

Depending on the task:

1. compare two documents, subtract one feature-value vector from the second one
2. characterize label (author), analyze feature-value vectors with the same label (author)
Stylometry-feature categories

**Categories**

- Morphological
- Syntactic
- Vocabulary
  - semantic words
  - stop-words
- Technical (text formatting, publishing time)
- Other
Techniques

Author’s characteristic features

**Word length statistics**
- Count and normalize frequencies of selected word lengths (e.g. 1–15 characters)
- Modification: word-length frequencies are influenced by adjacent frequencies in histogram, e.g.: 1: 30%, 2: 70%, 3: 0% is more similar to 1: 70%, 2: 30%, 3: 0% than 1: 0%, 2: 60%, 3: 40%

**Sentence length statistics**
- Count and normalize frequencies of
  - word per sentence length
  - character per sentence length
Techniques

Author’s characteristic features

Author gender

- Detect sentences written in the first person
- Extract author’s gender if possible
- \( \text{včera jsem byla v Brně a viděla} \)

Wordclass (bigrams) statistics

- Count and normalize frequencies of wordclasses (wordclass bigrams)
- \( \text{verb is followed by noun with the same frequency in selected five texts of Karel Čapek} \)
### Morphological tags statistics
- Count and normalize frequencies of selected morphological tags
- *the most consistent frequency has the genus for family and archaic freq in selected five texts of Karel Čapek*

### Word repetition
- Analyse which words or wordclasses are frequently repeated through the sentence
- *nouns, verbs and pronouns are the most repetetive in selected five texts of Karel Čapek*
Techniques

Author’s characteristic features

Stopwords

- Count normalized frequency for each word from stopword list
- Stopword ~ general word, semantic meaning is not important, e.g. prepositions, conjunctions, ...
- stopwords ten, by, člověk, že are the most frequent in selected five texts of Karel Čapek
Techniques

Author’s characteristic features

Syntactic Analysis

- Extract features using SET (Syntactic Engineering Tool)

```
<SENTENCE>
  <CLAUSE>
    <VP>
      <V>
      Verifikujeme
    <NP>
      autorství
      <PREP>
      <NP>
      se
      <ADJ>
      syntaktickou
      <N>
      analýzou
```

- **syntactic trees have similar depth in selected five texts of Karel Čapek**
Author’s characteristic features

Other stylometric features

- typography
- formatting richness
- emoticons
- errors
- vocabulary richness
 Techniques

Author’s characteristic features

Document comparison

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>global</td>
<td>0.0183</td>
</tr>
<tr>
<td>Wordclass bigrams</td>
<td>0.0003</td>
</tr>
<tr>
<td>Wordclasses</td>
<td>0.04</td>
</tr>
<tr>
<td>Morph. tags</td>
<td>0.0735</td>
</tr>
<tr>
<td>Word repetition</td>
<td>0.1105</td>
</tr>
<tr>
<td>Syntactic analysis</td>
<td>0.1358</td>
</tr>
<tr>
<td>Stopword</td>
<td>0.3538</td>
</tr>
<tr>
<td>Punctuation</td>
<td>0.3929</td>
</tr>
<tr>
<td>Chars per sentence</td>
<td>0.7667</td>
</tr>
</tbody>
</table>

Example: comparison between two different authors

Revelation of the author’s identity using machine learning and stylometry
Techniques

Author writeprint/stylome

Collection of author’s documents

Author analysis:

1. Range: typical feature values for that author
2. Consistency (deviation): which features are most important
3. Corpus similarity: which features are uncommon in corpus
Techniques

Machine learning approach

Automatic parameter tuning

- use models with probability estimation only if necessary
- try different techniques (Support vector machines, Nearest neighbors, Naive Bayes)
- try different kernels for SVM
- parameter grid search
- each problem and data type uses different ML model
**Single-layer ML technique (two-class: same vs different authorship)**

1. Extract document features for each author characteristic

2. Compare documents to obtain a similarity vector

3. ML classifier predicts probability of the same authorship

<table>
<thead>
<tr>
<th>vocabulary richness</th>
<th>word-length similarity</th>
<th>sentence-length sim.</th>
<th>stopword similarity</th>
<th>word repetition</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7</td>
<td>0.4</td>
<td>0.2</td>
<td>0.6</td>
<td>0.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A</th>
<th>0.7</th>
<th>0.7</th>
<th>0.2</th>
<th>0.6</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>0.3</td>
<td>0.1</td>
<td>0.3</td>
<td>0.1</td>
<td>0.9</td>
</tr>
<tr>
<td>1-</td>
<td>A-B</td>
<td></td>
<td>0.6</td>
<td>0.4</td>
<td>0.9</td>
</tr>
</tbody>
</table>

0.65
### Results

#### Similarity ranking

<table>
<thead>
<tr>
<th>Document Type</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Book:</strong></td>
<td>long coherent text</td>
</tr>
<tr>
<td><strong>Blog:</strong></td>
<td>medium-length text</td>
</tr>
<tr>
<td><strong>E-mail:</strong></td>
<td>short noisy text</td>
</tr>
</tbody>
</table>

- Different “document conditions” are considered
- **Attribution:** replace similarity by ranking of the author against other authors
- **Verification:** select random similar documents from corpus and replace similarity by ranking of the document against these selected documents
Double-layer machine learning

Replace heuristics by 2nd machine learning layer

- **Heuristic (proposed by linguist):**

\[
sim = 1 - \frac{1}{7} \cdot \sum_{s \in \{1..7\}} \frac{|A_s| - |B_s|}{|A| - |B|}
\]

- **New ML layer (replace linguist’s heuristic by empirical evidence):**

\[
\text{vector} = \left\langle \frac{A_s}{|A|} - \frac{B_s}{|B|} \right\rangle \text{ for } s \in \{1..7\}
\]

\[
sim = \text{classifier} (\text{vector})
\]
Balanced accuracy:

Verification:
- books, essays: 90% → 99%
- blogs, articles: 70% → 99%
- tweets: 70% → 99% (given enough tweets)

Attribution (depends on the number of candidates, comparison on blogs):
- up to 4 candidates: 80% → 95%
- up to 100 candidates: 40% → 60%

Revelation of the author’s identity using machine learning and stylometry
Current work

- Machine translation detection
  - Recognize texts translated by Google, Bing and other machine translators
  - Remove translations from corpora
  - Detect texts falsely submitted as translated by a human expert
  - \url{http://nlp.fi.muni.cz/sir}

Revelation of the author’s identity using machine learning and stylometry
Results

Current work

- Web structure detection
  - Create stylometric corpora
  - Detect web structure and download documents with meta-data (author, gender, age, title, topic)
Current work

- Gender detection
  - Use data from dating services
  - Detect advertisements with a falsely submitted gender

- Authorship detection consultations
Results

Thank you for your attention

Revelation of the author’s identity using machine learning and stylometry