

# 01 – Stylometry

## IA161 Advanced Techniques of Natural Language Processing

Jan Rygl

rygl@fi.muni.cz

NLP Centre, FI MU, Brno

September 18, 2017

- 1 Stylometry
  - Motivation
  - Definition
  - History
  - Author information
  - Field demarcation
- 2 Stylometry techniques
  - Stylometric-technique categories
  - Examples of stylometric techniques
- 3 Feature extraction and machine learning
  - Feature extraction
  - Machine learning

# Computational stylometry

Example: Dating services automated user content control

*Has a user managed to select a correct gender?*

\_\_\_\_\_ Female author \_\_\_\_\_  
LÁSKA

(contains love  $\Rightarrow$  female & doesn't contain money  $\Rightarrow$  female)  $\rightarrow$  60%  
FEMALE

\_\_\_\_\_ Female author \_\_\_\_\_  
Hledám blízkého člověka pro spokojený a harmonický rodinný  
život...Možná, že se objevíš v téhle specifické virtuální sféře..

(contains family  $\Rightarrow$  female & contains harmony  $\Rightarrow$  female & contains  
virtual world  $\Rightarrow$  male)  $\rightarrow$  60% FEMALE

\_\_\_\_\_ Female author \_\_\_\_\_  
Přečtete si profil a snad to napoví víc...

(is short  $\Rightarrow$  male)  $\rightarrow$  CANNOT DECIDE

# Computational stylometry

## Definition

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

## Application

- 1 basic research on the linguistic properties of text determining style<sup>a</sup>
- 2 literary research (resolving disputed authorship)
- 3 forensic applications (disputed authorship of suicide notes, blackmail letters etc.)
- 4 human resources profiling (describe and explain the causal relations between psychological and sociological properties of authors on the one hand, and their writing style on the other)[Daelemans, 2013]

---

<sup>a</sup><http://www.clips.ua.ac.be/~walter/educational/stylometry.html>

# History

*Mendenhall, T. C. 1887.*

*The Characteristic Curves of Composition. Science Vol 9: 237–49.*

- The first algorithmic analysis
- Calculating and comparing histograms of word lengths
- Authorship verification of Shakespeare's plays



Oxford, Bacon  
Derby, Marlowe

# Information about author

Stylometry techniques can reveal following information:

- 1 gender,
- 2 region of origin,
- 3 age,
- 4 personality (extraverted or introverted),
- 5 education level,
- 6 indication of the identity of the author:
  - ▶ authorship attribution,
  - ▶ machine generated text detection:
    - ★ spam detection,
    - ★ automatic translation detection,
- 7 etc.

# Demarcation against other fields

## Topic recognition

Topic recognition features	Stylometry features
repeated non-stop-words repeated phrases (rare in corpus)	(repeated) stop-words repeated phrases (common in corpus)
usually based on entity detection	mostly without entity detection

## Plagiarism detection

Plagiarism detection features	Stylometry features
word n-grams rare character n-grams based on word substitution based on word reordering in sentence	POS tags n-grams frequent character n-grams word choice is important word order is important

# Stylometric-technique categories

## Categories

- 1 morphological
- 2 syntactic
- 3 lexical
- 4 other

## Assumptions

Author has:

- 1 unique active vocabulary
- 2 favourite phrases and word n-grams
- 3 a certain level of knowledge of grammar (mistakes)
- 4 personalized handling of typography



# Examples: lexical stylometric features

## Word length

- use lemmas instead of tokens
- active vocabulary contains long words, preference of longer/shorter words, ...

## Vocabulary richness

- Yule, 1944:

$$K = \frac{10^4(\sum i^2 V_i - N)}{N^2} \quad (1)$$

where  $V_i$  denotes the number of words with frequency  $i$  and  $N$  is the number of words in the text.

- Simpson, 1949:

$$D = \sum (V_i \cdot \frac{i}{N} \cdot \frac{i-1}{N-1}) \quad (2)$$

where  $V_i$  denotes the number of words with frequency  $i$  and  $N$  is the number of words in the text.

# Implemented morphological stylometric features

## Overview

### Distribution of word lengths

- Naive word length distribution
- Improved word length distribution
- Word trigram length distributions

### Distribution of sentence length

- Naive sentence length distribution
- Improved sentence length distribution
- Sentence-trigram length distributions

### Word repetition

- Naive counting word repetition
- Bag of words repetition
- Wordclass repetition
- Distance between repeated words
- Sentence positions of repeated words

### Word class n-grams

### Morphological tags n-grams

- Morphological tags n-grams
- Relative freq. of simplified morphological tags

### Presence of letter-casing in sentences

- Presence of casing sequences
- Presence of indexed casing sequences

### Word suffixes

- Stemmer based word suffixes
- Parameter based word suffixes

### Word richness

### Dynamic stopwords

### Punctuation

- Punctuation rel. frequency
- Punctuation position rel. freq.
- Punctuation n-grams in a sentence

### Dynamic Typography

### Distribution of character sequences

### Emoticons

- Presence of emoticon n-grams
- Emoticon categories n-grams

### Character n-grams

### Syntactic analysis

# Examples: morphological stylometric features

## N-grams of part-of-speech tags

- $N > 1$  is better in free word order languages
- overcomes topic dependency of token N-grams

## Majka tags n-grams

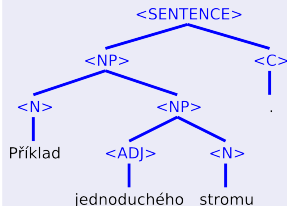
- Word class bigrams (“myslel bych”  $\Rightarrow$  [k5][kY])
- Most frequent reduced tags N-grams
  - 1 Filter out rare morphological information ([kYmCp1nS]  $\Rightarrow$  [kYp1nS])
  - 2 Find the most common tag sets on test data
  - 3 Use K the most frequent tag sets as features

# Examples: syntactic stylometric features

Requires syntactic analyser

SET

(<http://nlp.fi.muni.cz/trac/set>):



N-grams of clause types

- 1 unigrams:  $2 \times [N]$  and  $1 \times [ADJ]$
- 2 bigrams:  $1 \times [N][ADJ]$  and  $1 \times [ADJ][N]$
- 3 trigrams:  $1 \times [N][ADJ][N]$

Syntactic tree features

- 1 branching factor
- 2 depth
- 3 maximal width

# Feature extraction process

## Build train corpus

- 1 consists of texts similar to examined data
- 2 used to find the most common N-grams, stop words, ...
- 3 bigger is better

## Text normalization (same for train corpus and analysed data)

- 1 remove automatically generated tags (HTML, XML) and decode encoded entites
- 2 remove automatic text repetition, quotations (e-mails)
- 3 replace URLs, images, keys, ... by custom tag

# Feature extraction process

## Text preprocessing

- 1 annotate document (tokenization, morphological and syntactic analysis, entity and collocation detection, date and time recognition, ...)
- 2 save documents as object consisting of original text (needed for extending features and debugging) and all analysis outputs

## Training: Feature extraction, normalization and selection

- Given  $F$  features, generate feature vector  $\{f_{f1}, f_{f2}, \dots, f_{fF}\}$  for each document.
- Normalize each feature  $f_i$  (linear function  $S_{f_i}$  with target domain  $\langle 0, 1 \rangle$  or  $\langle -1, 1 \rangle$ )
- Feature selection  $F \Rightarrow F'$ .

# Feature extraction process

## Analysis

- Use  $F'$  features, generate feature vector for each document.
- Scale each feature  $f_i$  using function  $S_{f_i}$

## Process of document analysis

Pipeline consisting of:

- 1 Text normalization function: raw text  $\Rightarrow$  clean text
- 2 Text annotation functions: clean text  $\Rightarrow$  support objects containing morphological, syntactic and semantic information about text
- 3 Feature extraction: support objects  $\Rightarrow$  feature vector
- 4 Feature scaling (normalization): feature vector  $\Rightarrow$  scaled feature vector

# Machine learning

## Machine learning notes

- If using linear models, discretize or divide features (e.g. feature avg. world length convert into short, average and long words relative frequency features)
- Think if you analyse:
  - ① seen classes (for authorship attribution, we know all candidates, for gender prediction, there is only fixed number of genres) or
  - ② unseen classes (unknown authors, age wasn't present in train data): more difficult, requires tricks using features if data domain
- Think about your target audience:
  - ① Is important only result (automatic data classification)? Experiment with feature combinations and all possible features.
  - ② Do people want to examine results and evidence (court experties)? Feature need be comprehensible (add explanations of tags, don't use too complicated features). Be prepared to explain why was feature selected (linguistic background).



Thank you

# References I



Daelemans, W. (2013).

Explanation in computational stylometry.

In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 7817 of *Lecture Notes in Computer Science*, pages 451–462. Springer Berlin Heidelberg.



Kestemont, M. (2014).

Function words in authorship attribution from black magic to theory?

*EACL 2014*, pages 59–66.



Stamatatos, E. (2009).

A survey of modern authorship attribution methods.

*Journal of the American Society for Information Science and Technology*, 60(3):538–556.