

03 – Stylometry

IA161 Advanced Techniques of Natural Language Processing

Jan Rygl, Aleš Horák

NLP Centre, FI MU, Brno

September 29, 2021

- 1 Stylometry
 - Motivation
 - Definition
 - History
 - Author information
- 2 Stylometry techniques
 - Stylometry techniques
 - Feature extraction
 - Stylometric-technique categories
 - Examples of stylometric techniques
- 3 Authorship recognition results
- 4 Propaganda detection

Computational stylometry

Example: Dating services automated user content control

Has a user managed to select a correct gender?

_____ Female author _____

LÁSKA (LOVE)

(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..

(I am looking for a close person for a happy and harmonious family
life... Maybe you'll show up in this particular virtual realm...)

(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...

(Read the profile and hopefully it will tell you more...)

(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 forensic analysis (plagiarism, disputed authorship of suicide notes, blackmail letters etc.)
- 2 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)
- 3 supportive authentication (biometrics, e.g. in e-learning)
- 4 propaganda detection (manipulative style recognition)
- 5 literary research (resolving disputed authorship)
- 6 basic research on the linguistic properties of text determining style

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.

Stylometry techniques

Computational stylometry

- transform **text** → **vector** of characteristics/features (based on linguistic analysis)
- learn **weights** of each feature from **labelled documents**
- **analyze** features of **new/unknown** document to find its label

Authorship recognition through stylometry

- clean text (deduplication, boilerplate removal, remove markup tags)

| | |
|---|---|
| 1 | doc_id JM002 |
| 2 | Praví se v ní , že status quo nemůže pokračovat . |
| 3 | V nejbližší době je spíše pravděpodobné , že Řecko opustí eurozónu . |
| 4 | Odchod Řecka bude divoký a způsobí volatilitu , ale měnová unie s menším počtem členů přežije . |
| 5 | Aby mohla fungovat , bude potřebovat silnější fiskální unii , větší podporu bankovnímu systému a větší vzájemnost , provázanost (mutualization) dluhů , aby se zabránilo přeshraničním úprkům kapitálu Hugo Dixon má na Reuters zajímavý pohled na krizi eurozóny . |
| 6 | Podle něj existují dvě linie přetahování a sporů , první je spor mezi Severem a Jihem . |

Authorship recognition through stylometry

- morphological analysis

| | | |
|------------|-------|--------------|
| je | byt | k5eAaImIp3nS |
| spor | spor | k1gInSc1 |
| mezi | mezi | k7c7 |
| Severem | sever | k1gInSc7 |
| a | a | k8xC |
| Jihem | jih | k1gInSc7 |
| <g/> | | |
| . | . | kIx. |
| </s> | | |
| <s id="2"> | | |
| Jde | jit | k5eAaImIp3nS |

Authorship recognition through stylometry

- syntactic analysis

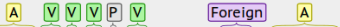
| | | | |
|----|------------|----|---|
| 13 | reformovat | 41 | p |
| 14 | svoje | 42 | p |
| 15 | ekonomiky | 43 | p |
| 16 | . | 44 | p |
| 17 | <CP> | 20 | p |
| 18 | <CLAUSE> | 20 | p |
| 19 | <CLAUSE> | 20 | p |
| 20 | <CLAUSE> | 22 | p |
| 21 | <SENTENCE> | -1 | p |
| 22 | <VP> | 21 | p |


Stylometry

Authorship recognition through stylometry

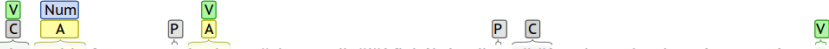
- extraction of the set of **stylometric features**

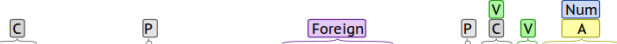
1 doc_id JM002

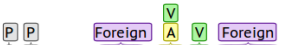
2  Práví se v ní , že status quo nemůže pokračovat .

3  V nejbližší době je spíše pravděpodobné , že Řecko opustí eurozónu .

4  Odchod Řecka bude divoký a způsobí volatilitu , ale měnová unie s menším počtem členů přežije .

5  Aby mohla fungovat , bude potřebovat silnější fiskální unii , větší podporu bankovnímu systému a

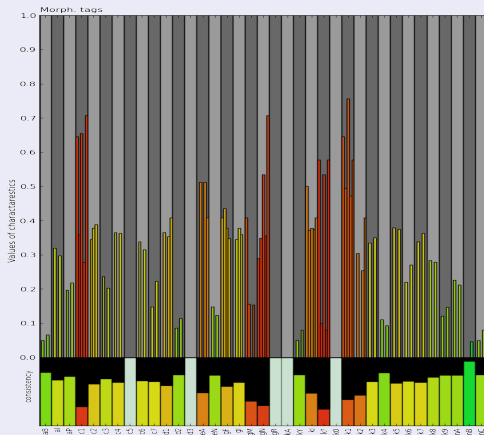
 větší vzájemnost , provázanost (mutualization) dluhů , aby se zabránilo přeshraničním úprkům

 kapitálu Hugo Dixon má na Reuters zajímavý pohled na krizi eurozóny .

Stylometry

Authorship recognition through stylometry

Stylome – author writeprint (from 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

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 markup** tags (HTML, XML) and decode encoded entites
- ② **remove** automatic text **headers**, 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 analyses** 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_{fi} 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

Stylometric-technique categories

Categories

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

Assumptions

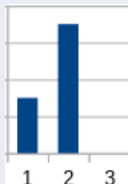
Author has:

- 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

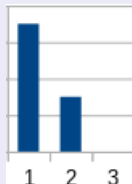
Author's characteristic features

Word/Sentence length statistics

- Count and normalize **frequencies** of:
 - selected **word** lengths (eg. 1–15 characters)
 - word per sentence** length
 - character per sentence** length
- Modification: word-length frequencies are influenced by **adjacent frequencies** in **histogram**, e.g.:



1:30% 2:70% 3:0%



1:70% 2:30% 3:0%



1:0% 2:60% 3:40%

Author's characteristic features

Author gender

- Detect sentences written in the **first person**
- Extract **author's gender** if possible
- *včera jsem byla v Brně a viděla*

Wordclass (bigrams) statistics

- Count and normalize frequencies of **word classes**/word class bigrams
- *verb is followed by noun with the same frequency in selected five texts of Karel Čapek*

Author's characteristic features

Morphological tags statistics

- Count and normalize frequencies of selected morphological tags
- Karel Čapek: family gender and archaic words have the most consistent frequencies

| Rod: Rodina (příjmení) | |
|------------------------|----------------------------------|
| Pád | Plurál |
| 1 | Novákovi |
| 2 | Novákových, Nováků |
| 3 | Novákům, Novákovým, Novákovům |
| 4 | Novákovy |
| 5 | Novákovi |
| 6 | Novákových |
| 7 | Novákovými |

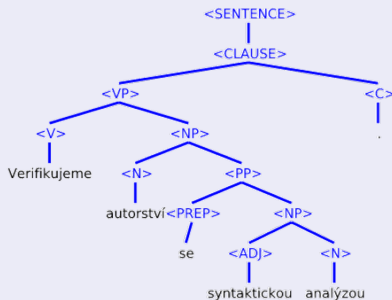
Word repetition

- Analyse which words or wordclasses are frequently repeated through the sentence
- Karel Čapek: nouns, verbs and pronouns are the most repetitive

Author's characteristic features

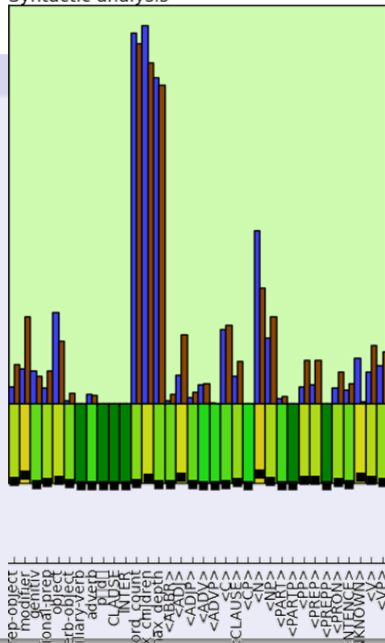
Syntactic Analysis

- Extract features using **syntactic analyzer**



- Karel Čapek: *syntactic trees have similar depth*

Syntactic analysis



Author's characteristic features

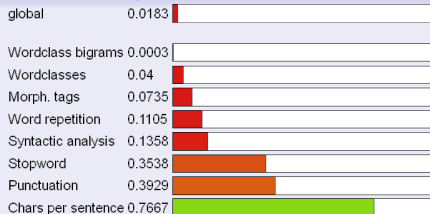
Other stylometric features

language independent:

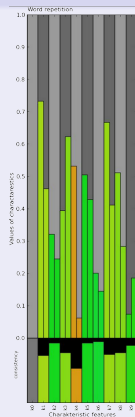
- **typography** (number of dots, spaces, emoticons, ...)
- **errors**
- **vocabulary** richness

Author's characteristic features

Document comparison



Example: comparison between two different authors

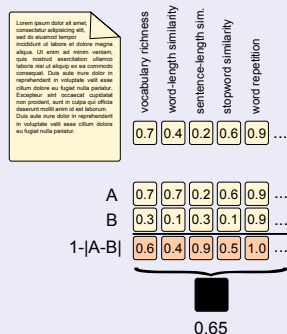


Similarity weights learning

Double-layer ML technique

binary: decide **same** vs **different** authorship

- 1 Extract document features for each author characteristic
- 2 apply learnt weights
- 3 Compare documents to obtain a **similarity vector**
- 4 ML classifier predicts probability of the same authorship



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

Authorship recognition (Czech texts)

Balanced accuracy: Current (CS) → Desired (EN)



Verification:

- books, essays: **95 %** → 99 %
- blogs, articles: **98 %** (20 % uncertain)
- twitter (>50/author): **99 %**

Attribution (for blogs):

- up to 4 candidates: **80 %** → 95 %
- up to 100 candidates: **40 %** → 60 %

Clustering:

- the evaluation metric depends on the scenario (**50–60 %**)

Minister Prize for Security Research

Propaganda detection

Propaganda detection

- 8,000 articles from 4 propaganda news servers:
`cz.sputniknews.com`, `parlamentnilisty.cz`, `ac24.cz` and `svetkolemnas.info`
- annotation for 8 manipulative techniques: **blaming**, **labelling**, **argumentation**, **emotions**, **demonizing**, **relativizing**, **fear mongering**, and **confabulation**
- **detection tool**

Usedne na Pražském hradě Havel 2.0?

cz.sputniknews.com

Je Horáček tím antiZemanem, kterého tzv. pražská kavárna usilovně hledá? Nemůže zopakovat osud Jana „Želě“ Fischera?

Konec konců, vyrazil přesně po jeho stopách. Vyvolává výrazné emoce, mobilizuje stoupence, vráží voliče protistrany, neumožní nikomu zastávat umírněný postoj. Jenže situace od minulých voleb se změnila.

Z islámské imigrace je významné téma, popularita EU dále klesla a ubylo voličů, kteří budou ochotni tolerovat proislámské postoje výměnou za schopnost nosit drahý oblek. Kdyby se duel Zeman – Schwarzenberg opakoval dnes, nedostal by kníže více než 30%. Stejně dopadne Horáček.

Zatím se prezentuje spíš jako Matěj Hollan 2.0.

Myslíte, že v pronárodním a antiuprchlickém táboře se najde kandidát, který by důstojně reprezentoval ve volbách náladu nemalé části české společnosti?

Takovým kandidátem je zcela jistě Miloš Zeman. Připomínám, že během několika let, které uplynuly od té nešťastné záležitosti s korunovačnými klenoty, se choval státnicky a neudělal nic, co by bylo možné označit za nedůstojné. Jistě, George Clooney nebo

Atributy s rozsahem

Nastav vše na NE*

| | | |
|----------------|-----------------|---|
| Místo | Česká republika | * |
| Vina | ne ano ? | * |
| Nálepkování | ne ano ? | * |
| Argumentace | ne ano ? | * |
| Obsažené emoce | rozhořčení | * |
| Démonizace | ne ano ? | * |
| Relativizace | ne ano ? | * |
| Strach | ne ano ? | * |
| Fabulace | ne ano ? | * |
| Názor | ne ano ? | * |
| Zdroj | ne ano ? | * |
| Rusko | missing | * |
| Odborník | ne ano ? | * |
| Politik 1 | Zeman | |
| Vyznění 1 | neutrální | * |
| Politik 2 | Schwarzenberg | |
| Vyznění 2 | neutrální | * |
| Politik 3 | Horacek | |
| Vyznění 3 | neutrální | * |

Current results

Propaganda detection

| label | MAX of test_f1_weighted |
|----------------------|-------------------------|
| demonizing Total | 95 % |
| relativizing Total | 92 % |
| fear mongering Total | 91 % |
| labelling Total | 83 % |
| emotions Total | 85 % |
| confabulation Total | 80 % |
| blaming Total | 74 % |
| argumentation Total | 71 % |

Remarks on the stylometric analysis tasks

- If using linear models, **discretize** or divide features (e.g. feature avg. word 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 of the data domain
- Think about your target **audience**:
 - ① just the **result** is important (automatic data classification)? Experiment with feature combinations and all possible features.
 - ② Do people want to examine results and **evidence** (court expertises)? Features must be comprehensible (add explanations of tags, don't use too complicated features). Be prepared to explain why a feature was selected (linguistic background).

References I



Bevendorff, J., Rosso, P., et al. (2020).

Overview of PAN 2020: Authorship verification, celebrity profiling, profiling fake news spreaders on twitter, and style change detection. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 372–383. Springer.



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.