Open Stylometric System *WebSty*: Towards Multilingual and Multipurpose Workbench

Maciej Piasecki, Tomasz Walkowiak
Wrocław University of Science and Technology
& CLARIN-PL Language Technology Centre
maciej.piasecki@pwr.edu.pl
tomasz.walkowiak@pwr.edu.pl

Maciej Eder
Institute of Polish Language, PAS
& Pedagogical University of Kraków
maciejeder@gmail.com
Stylometry

- **Stylometry**
  - identification of textual similarities and dissimilarities between texts
  - grouping (clustering) texts according to their linguistic characteristic
  - aimed at detecting signals in texts, e.g.
    - authorship, genre, gender, origin, style, etc.

- **Typical features for texts**
  - word form (words)
  - word form features: morphological and grammatical
  - collocations
  - syntactic properties: phrases and/or sentences
Stylometry - applications

- Authorship
  - attribution
  - recognition (from a closed set)
  - discovering (from texts or unlimited set)
- Period of writing
- Style recognition and analysis
- Genre recognition
- Origin
- Author features, e.g. gender, mother tongue
- Analysis of translations: source language, native language of the translator
- ...

...
Stylometry - barriers

- Technological
  - computer enough efficient for processing larger amounts of text
  - programming environment

- Knowledge
  - in programming
  - statistics, clustering methods, Machine Learning
  - Natural Language Engineering
    - interpretation of their results

- Language technology
  - limitations on the depth of analysis
  - definition of more sophisticated features, e.g. grammatical classes of words

- Lack of robust language tools
WebSty – open, web-based stylometric system

- **Idea:**
  - Web-based application that does not require installation
  - Equipped with Language Tools enabling definition of a rich set of features
    - only open LTs
    - robust in terms of coverage and accuracy
  - Integrated with access to many open tools for data analysis
    - feature transformation, similarity calculation, clustering, machine learning
    - visualisation and supporting analysis of the results
  - Lowering barriers in application of the stylometric tools by SS&H users
WebSty – scheme of processing

1. Corpus uploading
   - any format, text advised
   - descriptive file names or meta-data (CDMI)

2. Choice of the features

3. Setting up processing parameters
   - clustering vs classification
   - feature processing, e.g. transformation

4. Automated, feature-driven text pre-processing
   - automated pipeline of language tools

5. Feature extraction
   - mostly frequencies

6. Filtering and/or feature transformation

7. Main processing:
   - clustering
   - or/and classification

8. Presentation of the results
   - visualization
   - and/or export numerical data (CSV, Excel)
WebSty: corpus upload

Dane wejściowe

<table>
<thead>
<tr>
<th>Pliki lokalne</th>
<th>dSpace</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>B120(1).html</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B120(1).html</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B120(2).html</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B120(3).html</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B120(4).html</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B314(1).html</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B314(10).ht...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B314(100).ht...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B314(101).ht...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B314(103).ht...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B314(104).ht...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Usuń plik

Skasuj pliki
WebSty: corpus upload

- A corpus from the D-Space based repository of CLARIN-PL
- A corpus packed (Zip) from URL
Descriptive features (1)

- **Assumptions:**
  - possible to be identified on the appropriate level of accuracy
  - as little sensitive to the text semantics as possible

1. **Document level-features**
   - length of: a document, paragraph or sentence

2. **Morphological features (frequencies)**
   - word forms and tokens
     - all or from a predefined list, e.g. most frequent in NCP
   - punctuation marks
   - lemmas
     - all or from a predefined list (Polish, derived from the most frequent)
     - Recognised by a morpho-syntactic tagger (e.g. WCRFT2 for Polish)
3. Grammatical classes
   - 35 grammatical classes from the tagset of the National Corpus of Polish (WCRFT2 tagger)
     - e.g. pseudo-past participle, non-past form, ad-adjectival adjective, etc.

4. Parts of Speech
   - by grouping grammatical classes
   - Universal Part of Speech tags

5. Combinations: grammatical classes and selected categories (WCRFT2)
   - Verbs in 1\textsuperscript{st} and 2\textsuperscript{nd} person
6. Sequences of simple feature
   - bigrams of grammatical classes
   - trigrams of grammatical classes
     - some hints about the grammatical structures

7. Classes of Proper Names
   - e.g. person names, geographical names etc.
   - Recognised by a Named Entity Recogniser (Liner2 for Polish)
   - too much semantic features
# WebSty: feature selection

## Choice of features

### Number of occurrences in a document:

#### Elements:
- lemmas
- word forms

#### Punctuation:
- selected marks (list)
- all marks

#### Word classes:
- verbs
- nouns
- adjectives
- adverbs
- prepositions

### Grammatical classes:
- common nouns (sub in NKJP)
- depreciative forms
- main numerals
- collective numerals
- common adjectives (adj in NKJP)
- postadjective adjectives
- predicate adjectives
- postpositional adjectives
- non-3rd person personal pronouns
- 3rd person personal pronouns
- siebie reflexive pronouns
- winien modal verbs
- predicates
- coordinating conjunctions
- subordinating conjunctions
- exclamation marks
- burkinostka (a type of multi-word lexical units)
- kublik (e.g. -2/-2e, e.g. również)
- acronyms
- non-past forms/verbs
- future forms of być
- agglutinative forms of być
- pseudo-participles
- imperatives
- non-personal verbs
- infinitives
- simultaneous converses/transgressives
- anterior converses
- gerund
- present active participles
- present passive participle
- verbs in 1st or 2nd person

### Sequences of grammatical classes:
- with 2 elements (i.e. bigrams)
- with 3 elements (i.e. trigrams)
Filtering

- **Infrequent features**
  - minimal occurrences in the corpus
    - typically 20
  - minimal number of documents (fragments) including a feature
    - typically 5 (depends on the corpus size)

- **Planned**
  - pattern-based filtering, e.g. selected grammatical classes or bigrams matching a pattern
  - minimal value after feature transformation
Transformations

- Dimensionality reduction
  - Singular Value Decomposition (SVD)
  - Latent Semantic Analysis (SVD plus preprocessing)
  - Random Projection

- Feature weighting
  - heuristic transformations,
    - tf, tf.idf, normalisation
  - statistical association measures,
    - Chi2, tscore
  - based on Information Theory
    - Pointwise Mutual Information, Lin’s PMI
Similarity measures

- Applied to feature vectors representing documents (or text fragments)

- Distance measures
  - Manhattan, Canberra, euclidean, Simple (L1 on vectors normalised by a square root function) (Eder, 2016)

- Geometrical
  - cosine

- Heuristic
  - Dice, Jacquard,
  - ratio (average ratio of commonality), shd (precision of mutual rendering)
  - Burrows’s Delta, Argamon (Euclidean distance combined with Z-score normalisation), and Eder’s delta (Eder, 2016)
### Clustering options

<table>
<thead>
<tr>
<th>Filtering method</th>
<th>1</th>
<th>with smaller number of occurrence then 1 occurring with number of documents then</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature weighing method</td>
<td>trigramy</td>
<td>mi simple</td>
</tr>
<tr>
<td>Dimension reduction method</td>
<td>Singular value decompose</td>
<td>10</td>
</tr>
<tr>
<td>Similarity coefficient</td>
<td>ratio</td>
<td></td>
</tr>
<tr>
<td>Number of groups</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>
Clustering - *agglomerative-flat* clustering method from Cluto (Zhao & Karypis, 2005)
- pairwise hierarchy of similarity
- and flat division into a predefined, expected number of clusters

**Parameters**
- number of clusters

**Automated division of documents into fragments**
- for longer documents or size differences

**Pre-defined settings**
- authorship attribution, style analysis etc.
- tested on 1000 Books Corpus of literary works in Polish
### WebSty: similarity and clustering

**Clustering options**

<table>
<thead>
<tr>
<th>Filtering method</th>
<th>with smaller number of occurrence then</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>occurring with number of documents then</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature weighing method</th>
</tr>
</thead>
<tbody>
<tr>
<td>trigramy</td>
</tr>
<tr>
<td>mi simple</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dimension reduction method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singular value de</td>
</tr>
<tr>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Similarity coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>ratio</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>kosinusowa</td>
</tr>
<tr>
<td>dice</td>
</tr>
<tr>
<td>Jaccarda</td>
</tr>
<tr>
<td>ratio</td>
</tr>
<tr>
<td>shd</td>
</tr>
<tr>
<td>euklidesowa</td>
</tr>
<tr>
<td>manhattan</td>
</tr>
<tr>
<td>argamon</td>
</tr>
<tr>
<td>delta</td>
</tr>
<tr>
<td>Eder</td>
</tr>
<tr>
<td>simple</td>
</tr>
</tbody>
</table>

**Analyze**
Data visualisation

- Results:
  - For each text (file) - N
    - Texts could be automatically split into parts
  - Clustering results – group id (vector: Nx1)
  - Dendrogram (binary tree)
  - Similarities (matrix: NxN, values 1-0)
  - Distance (matrix: NxN, values 0-+∞)
  - Formats: JSON, XLSX
Data visualisation - interactive dendrogram
Data visualisation - heat map
Data visualisation - schemaball
Data visualisation – multidimensional scaling

- Texts as points in 2D or 3D,
  - distance between points reflects texts similarities
- Multidimensional scaling:
  - *metric*, preserving distances,
  - *non-metric*, preserving orders in distances,
  - *t-distributed Stochastic Neighbor Embedding*
    - preserving similarities,
  - *spectral embedding*,
    - preserving local neighborhood
Multidimensional scaling: 2D
Multidimensional scaling: 2D
Multidimensional scaling: 3D

Visualisation of clustering articles from *Teksty Drugie*

- weighting: MI-simple,
- similarity metric: ratio (from Cluto),
- number of clusters: 20,
- clustering method: agglomerative,
- visualization: the similarity matrix converted to distances and mapped to 3D by a spectral decomposition of the graph Laplacian - spectral embedding method)
WebSty Engine: LT chains
WebSty architecture: microservices

- parallel processing
- private cloud, scalability
- asynchronous processing
- Zipped files

Performance

LPMN:
urlzip(http://ws.clarin-pl.eu/ksiazki.zip) | any2txt|wcrft2|fextor({"features":"base"}) |dir|featfilt({"similarity":"jaccard"}) |cluto({"no_clusters":3})
WebStyEn – not only polish

<table>
<thead>
<tr>
<th>Choice of features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of occurrences in a document: ★</td>
</tr>
</tbody>
</table>

### Elements:
- lemmas: wyszystkie
- word forms: wyszystkie

### Punctuation:
- selected marks (list)
- all marks

### Word classes:
- verbs
- nouns
- adjectives
- adverbs
- prepositions

### Other grammatical classes:
- pronoun
- auxiliary
- coordinating conjunction
- determiner
- interjection
- numeral
- particle
- symbol
- other

### Sequences of grammatical classes:
- with 2 elements (i.e. bigrams)
- with 3 elements (i.e. trigrams)

### Named entities:
- persons
- Nationalities or religious or political groups
- Buildings, airports, highways, bridges, etc.
- Companies, agencies, institutions, etc.
- Countries, cities, states
- Locations: mountain ranges, bodies of water
- Objects, vehicles, foods, etc.
- Events
- Titles of books, songs, etc
- language

### Values:
- Percentage, including "%"
- Monetary values, including unit
- Times smaller than a day.
- Absolute or relative dates or periods.
- "first", "second", etc.
- Measurements, as of weight or distance
- Numerals that do not fall under another type

---

**English - spaCy**
Feature selection

Tools: Weka, scipy, scikit-learn

Group of methods:
- statistical tests
  - for example Mann-Whitney
- information metrics
  - for example InfoGain,
- recursive feature elimination using supervised classifiers
  - like Naive Bayes
- feature importances available in tree based classifiers,
  - e.g. Random Forest
Example of the features

Corpus: 1000 polish books; features: bases, punctuation, bigrams; weighting PMI; feature selection Mann-Whitney

Kraszewski_syn_jazdona_1880
Kraszewski_krakow-za-loktka_1880
Kraszewski_pogrobek_1880
Kraszewski_kunigas_1882
Kraszewski_boleszczyce_1877
Kraszewski_stara-basn-tom-III_1876
Kraszewski_bracia-zmartwychwstancy_1876
Kraszewski_banita_1885
Kraszewski_strzemienczyk_1883
Kraszewski_stara-basn-tom-I_1876
Kraszewski_bialy-ksiaze_1882
Kraszewski_jelita_1881
Kraszewski_caprea-i-roma_1860
Kraszewski_stara-basn-tom-II_1876
Stryjkowski_stryjkowski_kronika-polska-litewska-zmudzka-i-wszystkiej-rusi_1846

bigrams:inf_imps
bigrams:inf_praet
bigrams:ppron3_pcon
bigrams:ppas_pcon
bigrams:imps_interp
bigrams:ppron3_pant
bigrams:pant_interp
lex_classes:imps_count
bigrams:subst_pant
bigrams:interj_inf
base:wszyscy
bigrams:siebie_pcon
base:on
base:choć
base:gdy
bigrams:praet_pant
bigrams:ppron3_imps
bigrams:adj_pant
bigrams:pant_pact

...
Thank you very much for your attention!