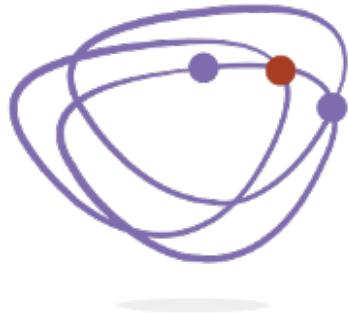


Open Stylometric System *WebSty*: Towards Multilingual and Multipurpose Workbench



CENTRUM TECHNOLOGII
JĘZYKOWYCH **CLARIN-PL**

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 ▾

Pliki lokalne dSpace URL

B120(1).html...

Usuń plik

B120(1).html...

Usuń plik

B120(2).html...

Usuń plik

B120(3).html...

Usuń plik

B120(4).html...

Usuń plik

B314(1).html...

Usuń plik

B314(10).ht...

Usuń plik

B314(100).ht...

Usuń plik

B314(101).ht...

Usuń plik

B314(101).ht...

Usuń plik

B314(103).ht...

Usuń plik

B314(104).ht...

Usuń plik

B314(104).ht...

Usuń plik

Skasuj pliki

WebSty: corpus upload



Lista korpusów w repozytorium

ID	Nazwa	Autor	Wybór
/11321/246/	Iwo Gall -teksty teatralne	Dulna-Rak, Ewa	[grid icon]
/11321/245/	ROThA_Vol. 2	Włodarczyk, Matylda	[grid icon]
/11321/239/	Test cmdi	Tomasz, Naskręt	[grid icon]
/11321/243/	Tekst reklam TVP ABC ver.2	Ka, Emilia	[grid icon]
/11321/22/	NELexicon	Marcińczuk, Michał	[grid icon]
/11321/59/	Żeromski	Żeromski, Stefan Żeromski	[grid icon]
/11321/230/	Wcrft test	Marcińczuk, Michał	[grid icon]
/11321/29/	WordNet	Piasecki, Maciej	[grid icon]
/11321/16/	KPWrr	Maziarz, Marek	[grid icon]
/11321/17/	KPWrr-lemma	Radziszewski, Adam	[grid icon]
/11321/19/	Lists of semantic relatedness	Piasecki, Maciej	[grid icon]
/11321/240/	TEST with cmdi 2	Tomasz, Naskręt	[grid icon]
/11321/221/	Mining blogs	Konieczna, Dorota	[grid icon]
/11321/95/	MWE Zapolska	Zapolska, Zapolska	[grid icon]
/11321/77/	MWE Kaczkowski	Kaczkowski, Kaczkowski	[grid icon]

Dane wejściowe ▾

Pliki lokalne dSpace URL

Adres pliku (zip)

http://ws.clarin-pl.eu/public/teksty/5_autorow_skrot

- 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)

Descriptive features (2)



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 1st and 2nd person

Descriptive features (3)



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 (*subst* in NKJP)
- depreciative forms
- main numerals
- collective numerals
- common adjectives (adj in NKJP)
- postadjective adjectives
- predicate adjectives
- postprepositional adjectives
- non-3rd person personal pronouns
- 3rd person personal pronouns
- siebie* reflexive pronouns
- winię* modal verbs
- predicates
- coordinating conjunctions
- subordinating conjunctions
- exclamation marks ()
- burkinostka (a type of multi-word lexical units)
- kublik (e.g. -ż/-że, e.g. również)

- acronyms
- non-past forms/verbs
- future forms of *być*
- agglutinant forms of *być*
- pseudo-participles
- imperatives
- non-personal verbs
- infinitives
- simultaneous converbs/transgressives
- anterior converbs
- 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, Jaccard,*
 - *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)

WebSty: filtering and transformation



Clustering options▼

Filtering method

with smaller number of occurrence then

occurring with number of documents then

Feature weighing method

trigramy

mi simple

Dimension reduction method

Singular value de

10

Similarity coefficient

ratio

Number of groups

2

Clustering



- 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 ▾

Filtering method

with smaller number of occurrence then

occurring with number of documents then

Feature weighing method

Dimension reduction method

Similarity coefficient

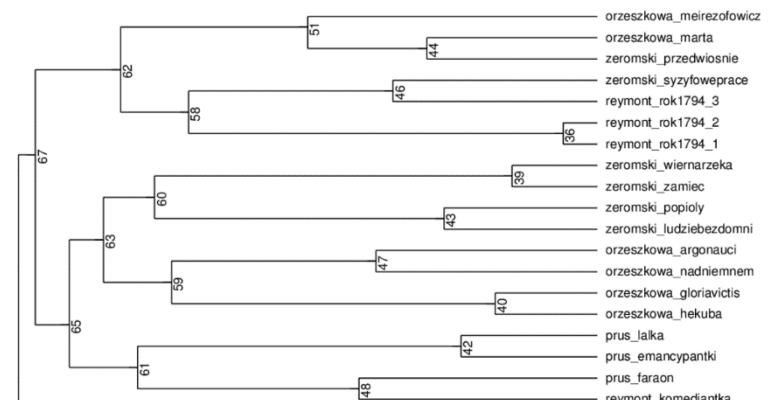
Number of groups

- ratio
- kosinusowa
- dice
- Jaccarda
- ratio**
- shd
- euklidesowa
- manhattan
- argamon
- delta
- Eder
- simple

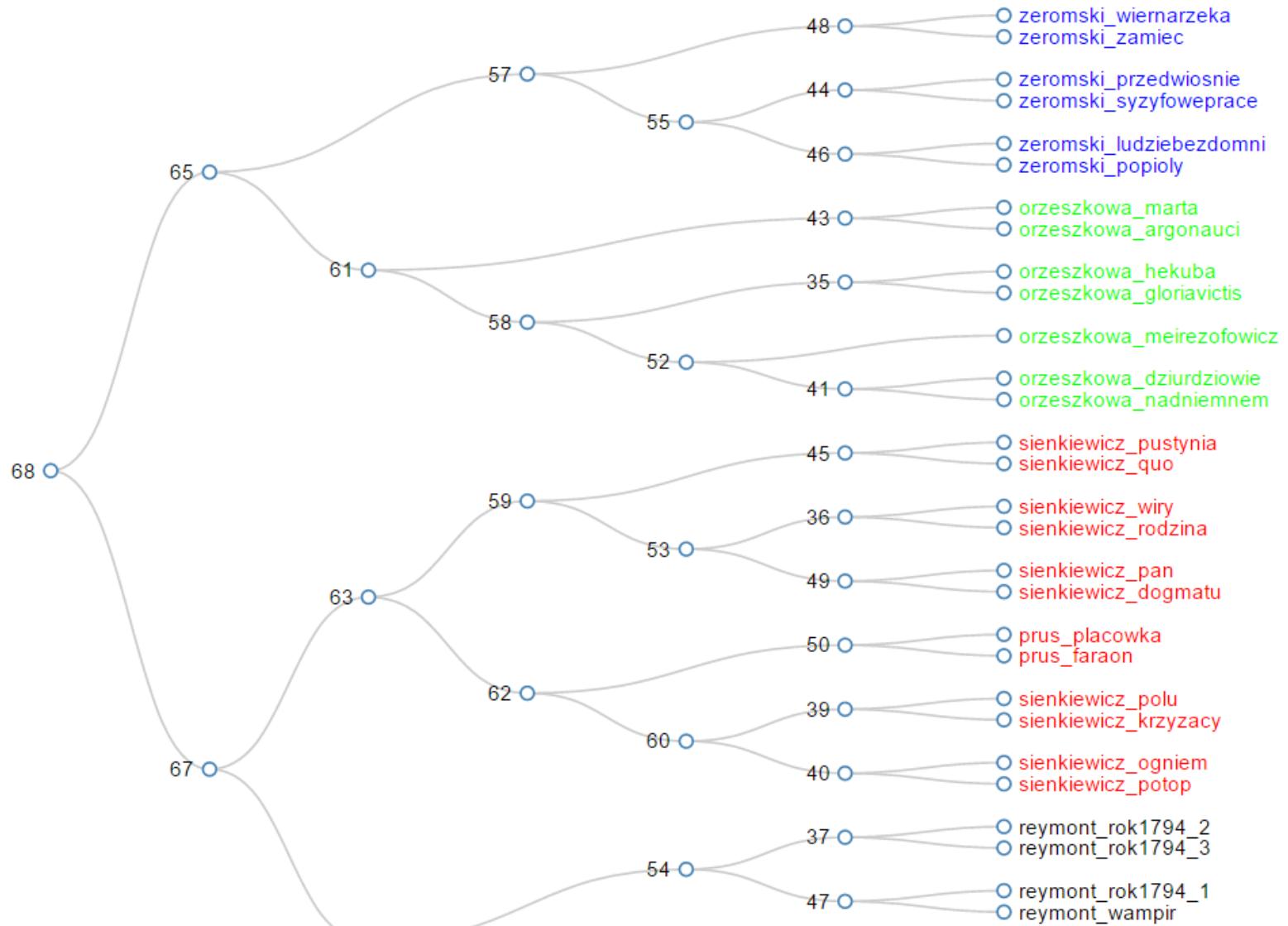
Analyze

Data visualisation

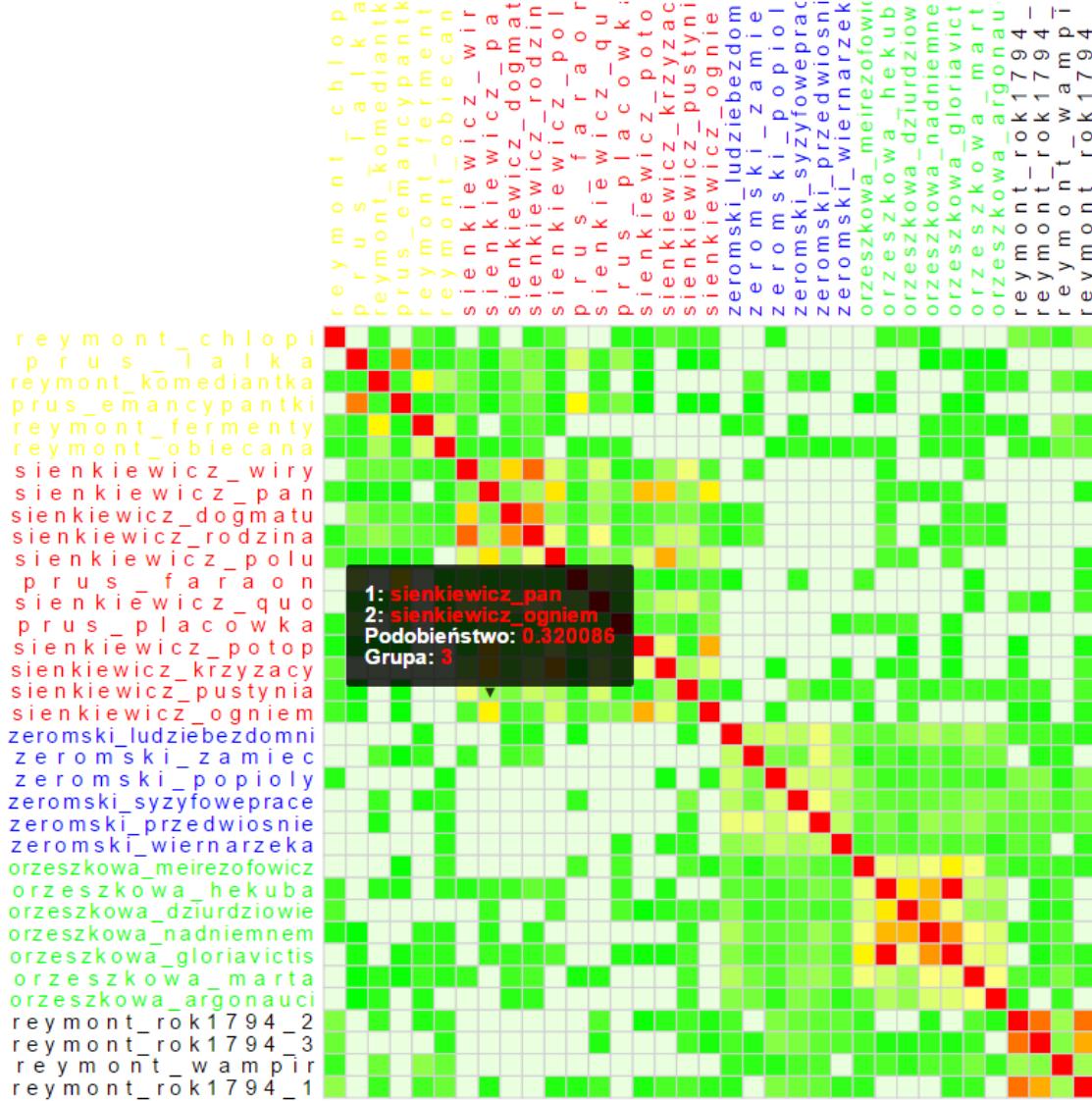
- Results:
 - For each text (file) - N
 - Texts could be automatically split into parts
 - Clustering results – group id (vector: Nx1)
 - Dendrogram (binary tree)
 - Similarities (matrix: NxN, values 1-0)
 - Distance (matrix: NxN, values 0- $+\infty$)
 - Formats: JSON, XLSX



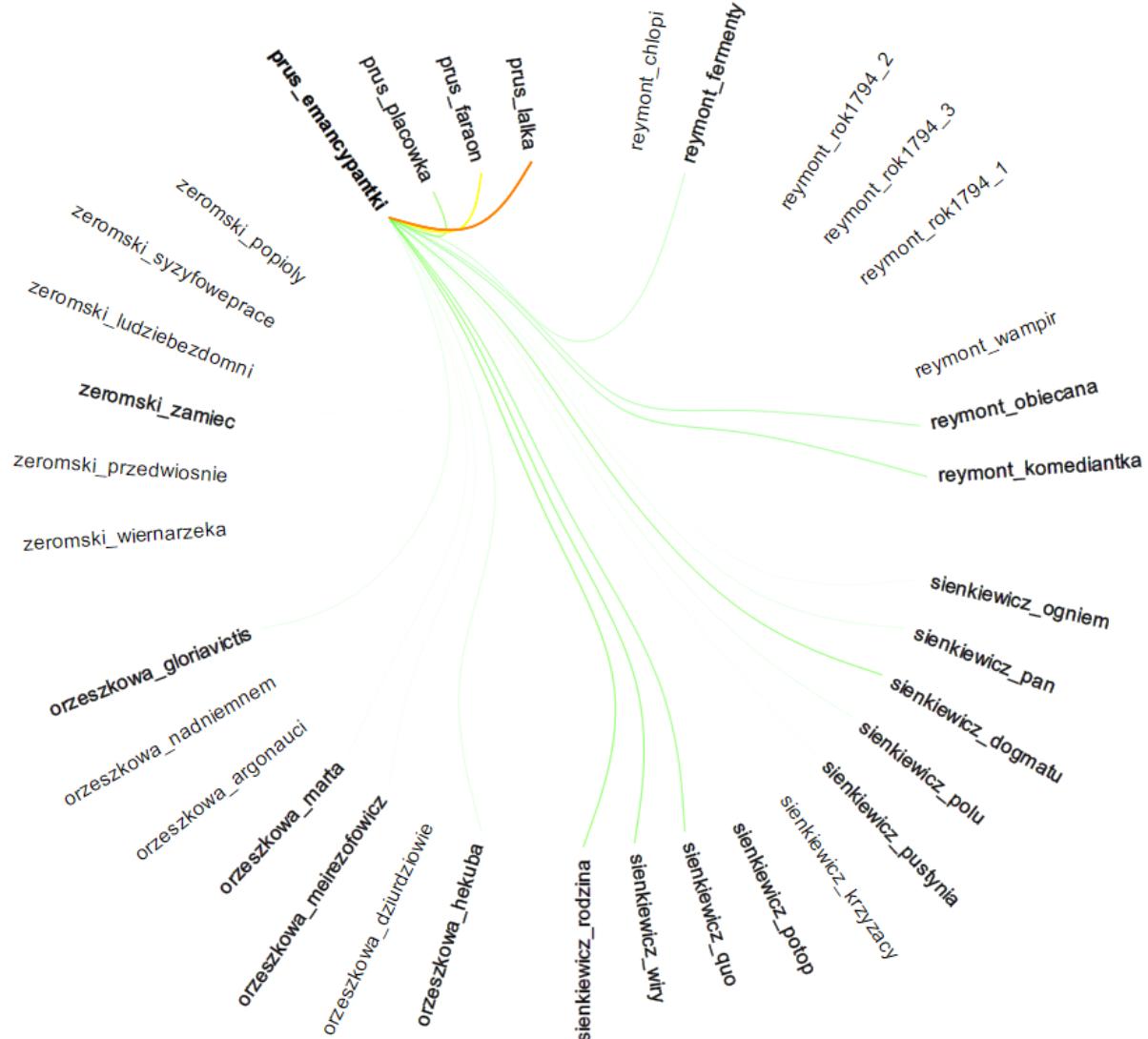
Data visualisation - interactive dendrogram



Data visualisation - heat map



Data visualisation - schemaball



Data visualisation – multidimensional scaling

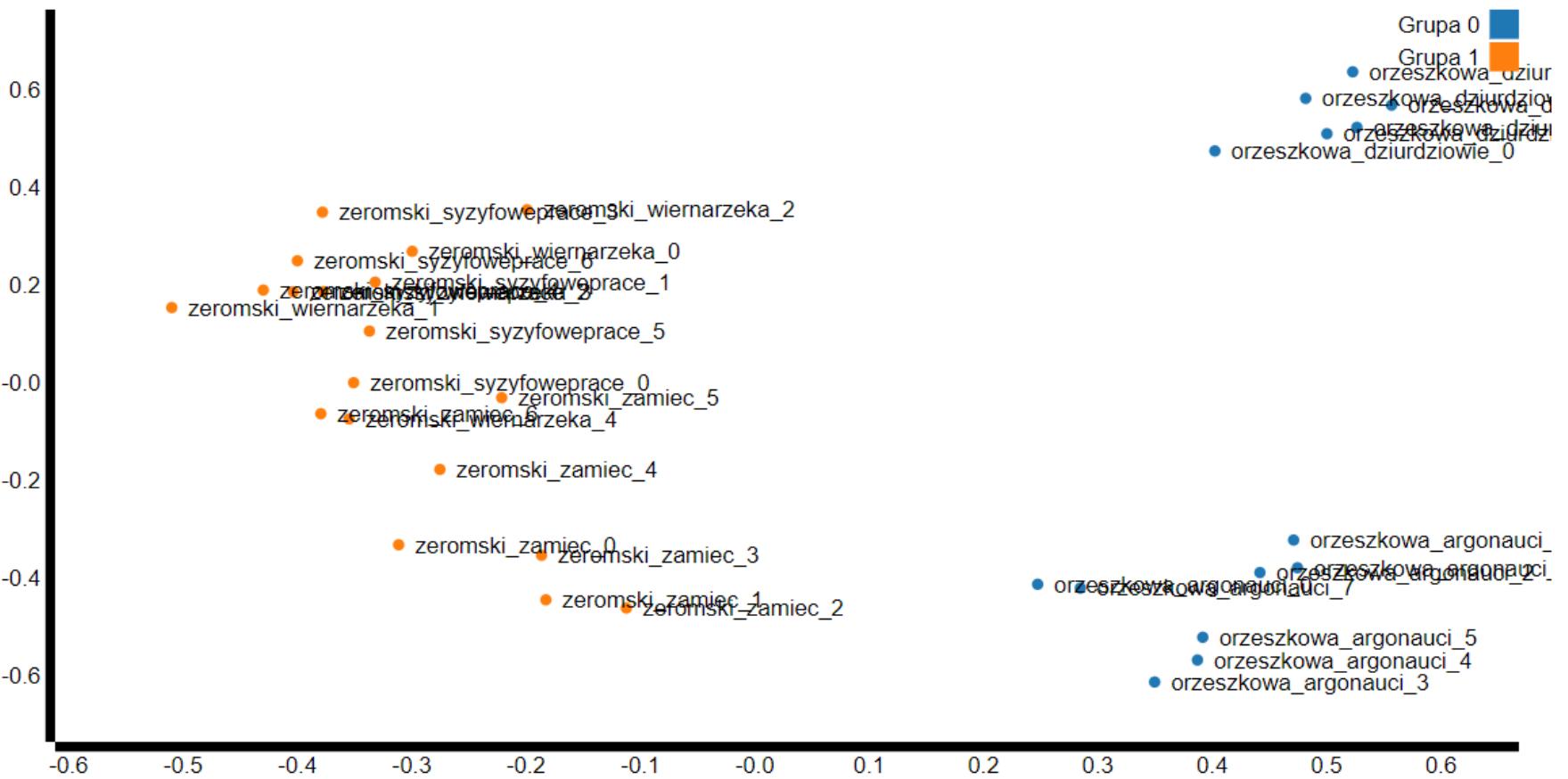


- Texts as points in 2D or 3D,
 - distance between points reflects texts similarities
- Multidimensional scaling:
 - *metric*, preserving distances,
 - *non-metric*, preserving orders in distances,
 - ***t-distributed Stochastic Neighbor Embedding***
 - preserving similarities,
 - ***spectral embedding***,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

Multidimensional scaling: 2D



Multidimensional scaling: 2D

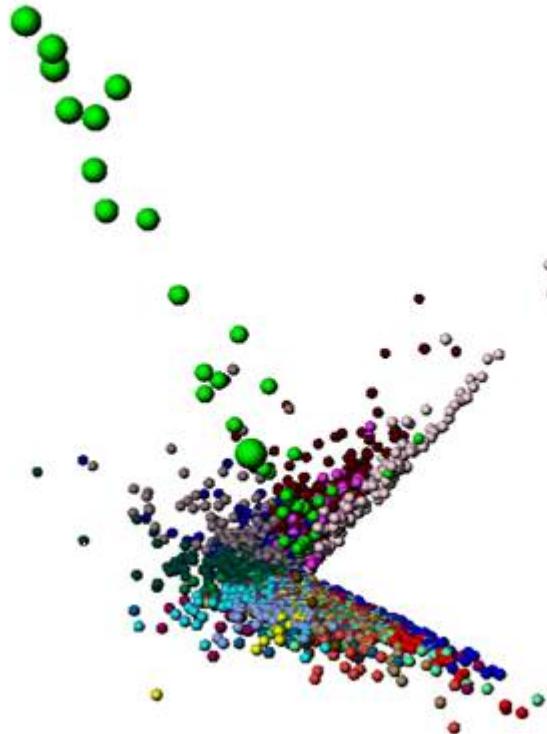


Multidimensional scaling: 3D

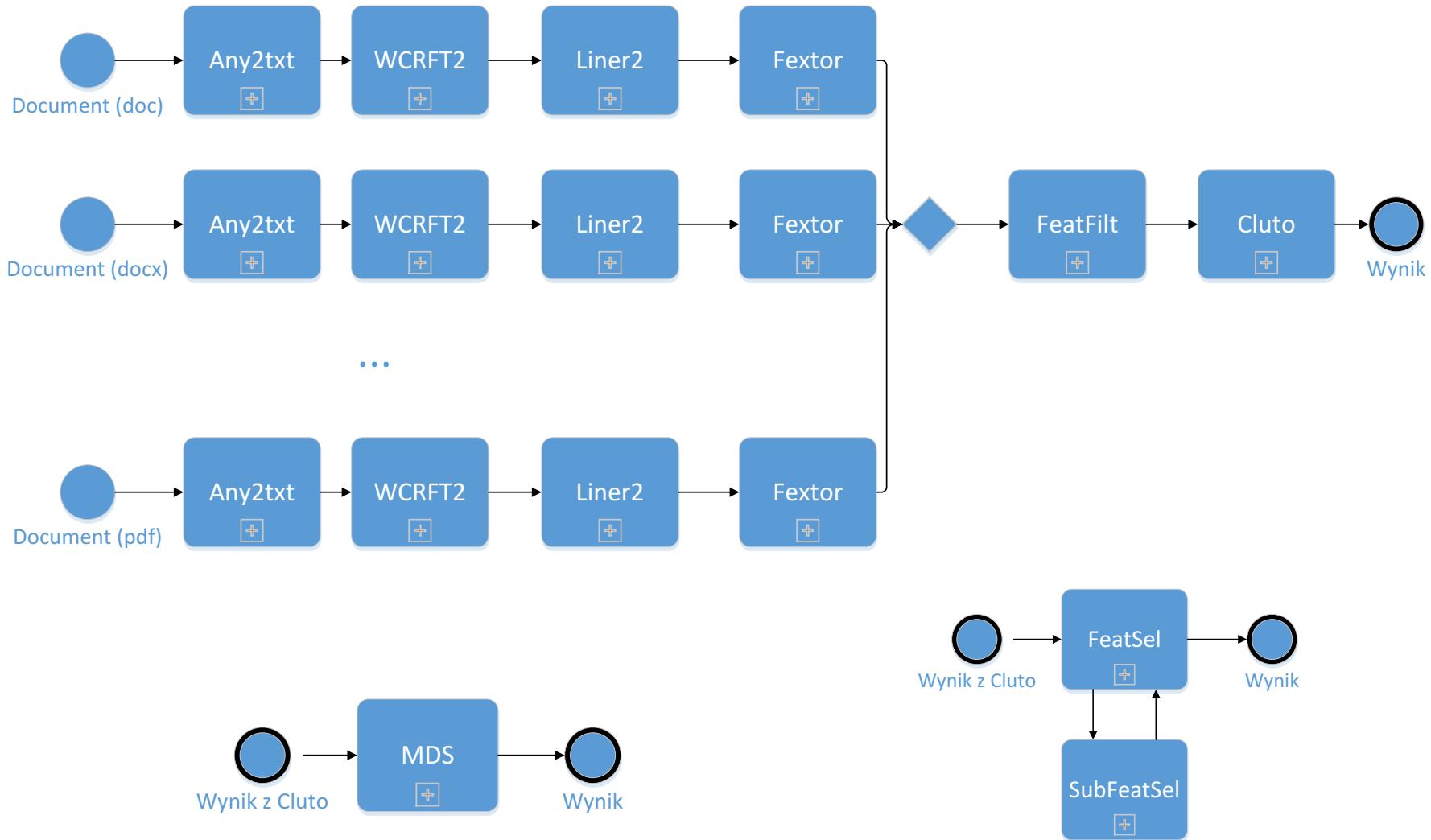


Visualisation of clustering articles from *Teksty Drugie*

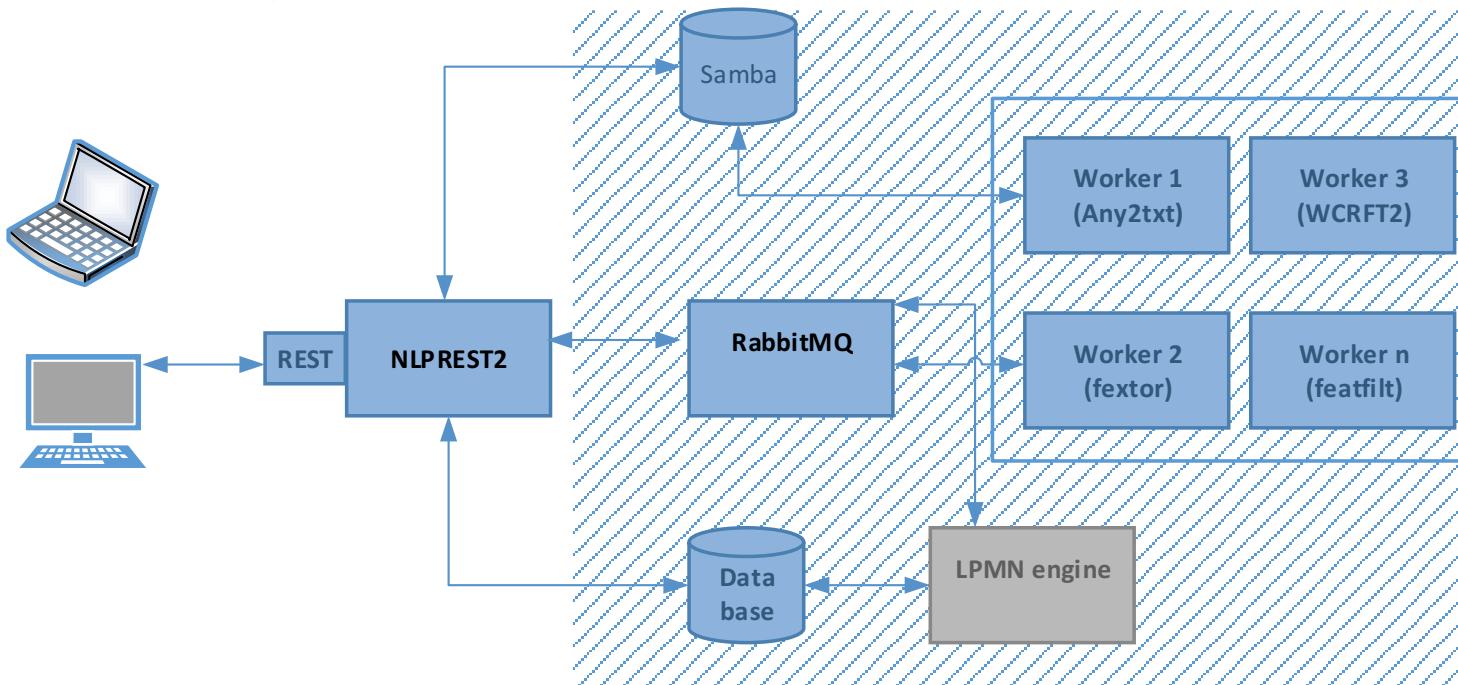
- weighting: MI-simple,
- similarity metric: ratio (from Cluto),
- number of clusters: 20,
- clustering method: agglomerative,
- visualization: the similarity matrix converted to distances and mapped to 3D by a spectral decomposition of the graph Laplacian - spectral embedding method)



WebSty Engine: LT chains

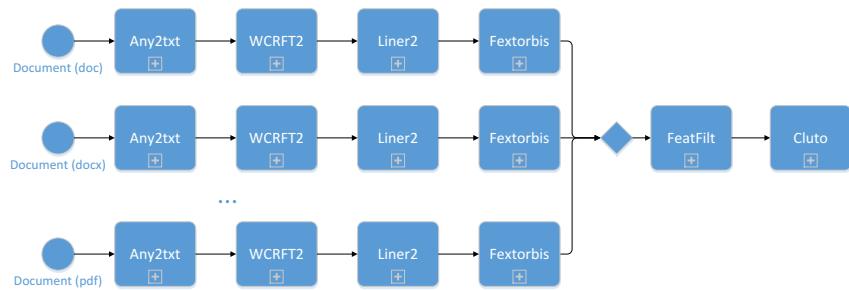


WebSty architecture: microservices



LPMN:

```
urlzip(http://ws.clarin-pl.eu/ksiazki.zip)
|any2txt|wcrft2|fextor({"features":"base"})
|dir|featfilt({"similarity":"jaccard"})
|cluto({"no_clusters":3})
```



Performance

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

WebStyEn – not only polish



Choice of features▼

Number of occurrences in a document:★

Elements:

- lemmas
- word forms

Other grammatical classes:★

- pronoun
- auxiliary
- coordinating conjunction
- determiner
- interjection
- numeral
- particle
- symbol
- other

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

Punctuation:

- selected marks (list▼)
- all marks

Word classes:★

- verbs
- nouns
- adjectives
- adverbs
- prepositions

Sequences of grammatical classes:★

- with 2 elements (i.e. bigrams)
- with 3 elements (i.e. trigrams)

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



Automatic selection of features

Options ▾

Selection method	Weka
Selection method	InfoGai
Number of features	100

Analyze

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!

<http://ws.clarin-pl.eu/websty.html?en>