

Reprezentacja dokumentów oparta na macierzy częstości

Przetwarzanie języka naturalnego
Ćwiczenia 1.

Rok akademicki: 2016/2017

MACIERZ CZĘSTOŚCI – ASPEKTY TEORETYCZNE

Reprezentacja dokumentów tekstowych

- **Reprezentacja unigramowa** (model przestrzeni wektorowej, reprezentacja bag-of-words, BOW)
- *A vector space model for automatic indexing* (1975), by G. Salton, A. Wong, C. S. Yang, Communications of the ACM

$\mathbf{X} = \begin{bmatrix} & \text{Dokumenty} \\ & \\ & \\ & \end{bmatrix}$

Wyrazy

x_{ij} – liczba wystąpień i -tego wyrazu w j -tym dokumencie

Nie uwzględnia kolejności wyrazów w tekście!

Najpopularniejszy sposób reprezentacji dokumentów.

Paweł Lula, Katedra Systemów Obliczeniowych, Uniwersytet Ekonomiczny w Krakowie

3

Tworzenie macierzy częstości (BOW – bag-of-words)

- Podział dokumentów na wyrazy,
- Usunięcie wyrazów nieistotnych (zawartych na stop-liście),
- Przekształcenie wyrazów do formy podstawowej,
- Utworzenie macierzy częstości,
- Przekształcenie macierzy częstości.

Paweł Lula, Katedra Systemów Obliczeniowych, Uniwersytet Ekonomiczny w Krakowie

4

Wyznaczanie macierzy częstości BOW (1)

- Podział dokumentów na wyrazy

Mowa jest srebrem, lecz milczenie złotem.

↓

*mowa
jest
srebrem
lecz
milczenie
złotem*

Wyznaczanie macierzy częstości BOW (2)

- Usunięcie słów nieistotnych (*stop-lista*)

Mowa jest srebrem, lecz milczenie złotem.

↓

*mowa
jest
srebrem
lecz
milczenie
złotem*

Wyznaczanie macierzy częstości BOW (3)

- Przekształcenie wyrazów do formy podstawowej

Mowa jest srebrem, lecz milczenie złotem.



mowa - mowa

jest - być

srebrem - srebro

lecz - lecz

milczenie - milczenie

złotem - złoto

Stemming i lematyzacja

- Redukcja do rdzenia (rdzeń – nieodmienna część wyrazu) = stemming
- Redukcja do formy podstawowej (bezokolicznik, mianownik) = lematyzacja

Stemming i lematyzacja

- Metody redukcji do rdzenia (stemming):
 - regułowa
 - algorytm Lovins – opisany w: Julie Beth Lovins (1968) Development of a stemming algorithm. Mechanical Translation and Computational Linguistics, 11: 22-31.
 - algorytm Portera – opisany w: M.F. Porter, 1980, An algorithm for suffix stripping, Program, 14(3) pp 130-137
 - słownikowa
 - bazująca na słowniku morfologicznym.

Wyznaczanie macierzy częstości BOW (4)

- Utworzenie wspólnej listy dla wszystkich dokumentów

Milczenie - przyjaciel który, nigdy nie zdradza

Książka to przyjaciel, który nigdy nie zdradzi

książka, który, milczenie, nie, nigdy, przyjaciel, to, zdradzać

Wyznaczanie macierzy częstości BOW (5)

- Utworzenie macierzy częstości

$$\mathbf{X} = \begin{matrix} & \text{Dokumenty} \\ \begin{matrix} \text{Wyrazy} \end{matrix} & \left[\begin{array}{c} \\ \\ \\ \end{array} \right] \end{matrix}$$

x_{ij} – liczba wystąpień i -tego wyrazu w j -tym dokumencie

Przetwarzanie macierzy częstości BOW

- zmiana wartości przechowywanych w macierzy częstości (bez zmiany rozmiarów macierzy) – w celu lepszej reprezentacji informacji zawartych w dokumencie,
- redukcja wymiarów macierzy częstości.

Modyfikacje macierzy częstości – bez zmiany rozmiarów (1)

- Reprezentacja binarna

$$\mathbf{X} = \begin{bmatrix} 2 & 0 & 4 & \dots & 4 \\ 1 & 0 & 3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 1 & 2 & \dots & 1 \end{bmatrix} \xrightarrow{\text{red arrow}} \mathbf{X}^{\text{bin}} = \begin{bmatrix} 1 & 0 & 1 & \dots & 1 \\ 1 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 1 & 1 & \dots & 1 \end{bmatrix}$$

Modyfikacje macierzy częstości – bez zmiany rozmiarów (2)

- Reprezentacja logarytmiczna

$$\mathbf{X} = \begin{bmatrix} 2 & 0 & \dots & 4 \\ 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 1 & \dots & 2 \end{bmatrix} \xrightarrow{\text{red arrow}} \mathbf{X}^{\text{log}} = \begin{bmatrix} 1,301 & 0,000 & \dots & 1,602 \\ 1,000 & 0,000 & \dots & 0,000 \\ \dots & \dots & \dots & \dots \\ 0,000 & 1,000 & \dots & 1,301 \end{bmatrix}$$

$$x_{ij} \rightarrow 1 + \log(x_{ij})$$

Modyfikacje macierzy częstości – bez zmiany rozmiarów (3)

- Ważona reprezentacja logarytmiczna (model TFIDF)

Reprezentacja logarytmiczna

$$x_{ij} \rightarrow 1 + \log(x_{ij})$$

Ważona reprezentacja logarytmiczna

$$x_{ij} \rightarrow (1 + \log(x_{ij})) * \log(N/df_i)$$

N - liczba wszystkich dokumentów

df_i - liczba dokumentów zawierających i-ty wyraz

Redukcja wymiarów macierzy częstości

- Dwa podejścia do zagadnienia redukcji
 - wybór reprezentantów – usuwane są informacje dotyczące mniej istotnych wyrazów:
 - zastosowanie stop listy,
 - usunięcie informacji o wyrazach występujących tylko w jednym dokumencie,
 - usunięcie wyrazów występujących bardzo rzadko,
 - usunięcie wyrazów występujących bardzo często,
 - stworzenie nowego zestawu cech opisujących dokumenty/wyrazy
 - analiza głównych składowych,
 - dekompozycja według wartości osobliwych.

MACIERZ CZĘSTOŚCI – ASPEKTY PRAKTYCZNE

Paweł Lula, Katedra Systemów Obliczeniowych, Uniwersytet Ekonomiczny w Krakowie

17

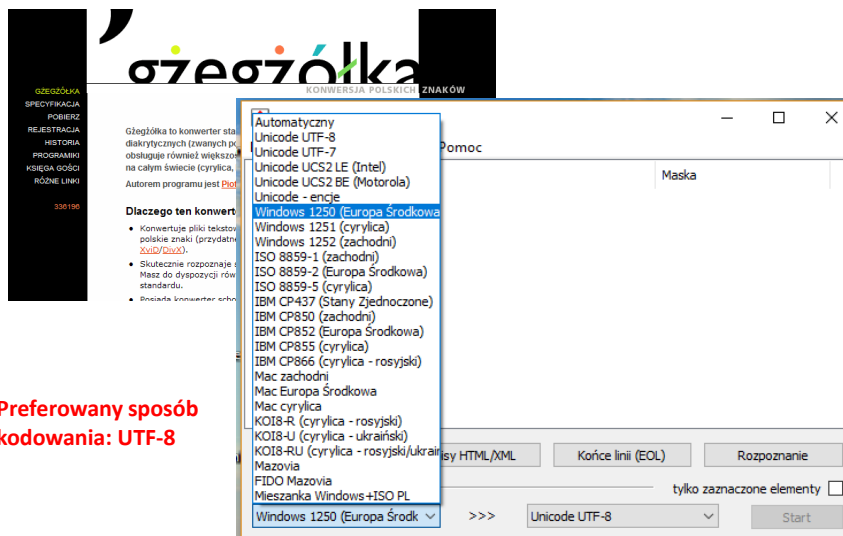
Wstępne przetworzenie dokumentów

- Transformacja dokumentów do postaci tekstowej,
- Usunięcie znaków formatujących,
- Ujednolicenie sposobu kodowania znaków.

Paweł Lula, Katedra Systemów Obliczeniowych, Uniwersytet Ekonomiczny w Krakowie

18

Gzeglółka – zmiana sposobu kodowania



Preferowany sposób
kodowania: UTF-8

Paweł Lula, Katedra Systemów Obliczeniowych, Uniwersytet Ekonomiczny w Krakowie

19

<http://sgjp.pl/morfeusz/morfeusz.html>

in English



Analizator morfologiczny Morfeusz

Podstawowe pojęcia

Słowem nazywamy ciąg znaków w tekście w języku naturalnym zwykle wydzielony odstępami lub znakami interpunkcyjnymi. **Leksem** to abstrakcyjna jednostka języka, wyraz słownikowy. **Formą wyrazową** nazywamy słowo zinterpretowane — przypisane do konkretnego leksemu i opisane co do jego funkcji gramatycznej.

Analiza morfologiczna polega na określeniu dla danego słowa wszystkich form wszystkich leksatów, których może ono być wykładnikiem. W procesie tym nie uwzględnia się kontekstu, w którym wystąpiło dane słowo. W językoznawstwie termin analiza morfologiczna odnosi się raczej do rozkładania wyrazów na elementarne składniki morfologiczne (morfemy), być może sensowniej byłoby więc mówić o **analizie fleksyjnej**, niestety wydaje się, że to ten pierwszy termin utarł się w środowisku językoznawstwa komputerowego.

Ujednoznacznianiem morfologicznym nazywamy określanie na podstawie kontekstu, jaką formę realizuje dane wystąpienie słowa.

Następujące po sobie analizę i ujednoznacznianie morfologiczne nazywa się żargonowo **tagowaniem**.

Celem **hasłowania (lematyzacji)** jest wskazanie dla każdego słowa tekstowego opisującej je jednostki słownika morfologicznego (leksemu). Jest to więc analiza morfologiczna (lub tagowanie) ograniczona tylko do części informacji o formach — do lematów.

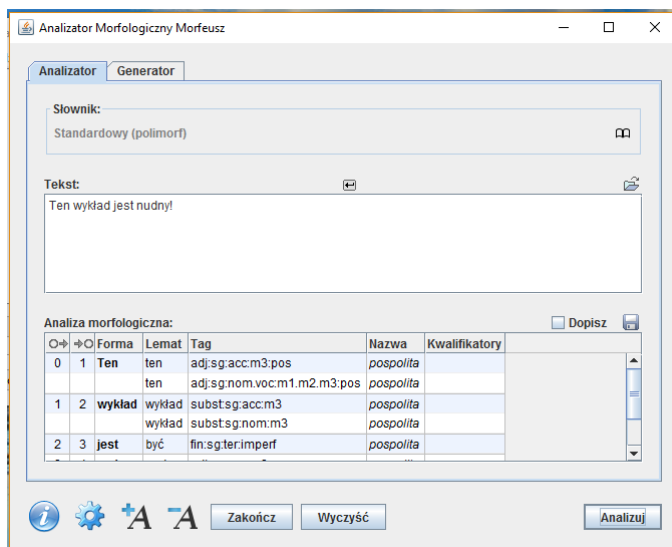
Przybliżone hasłowanie polegające na odcięciu ze słów części zmieniającej się przy odmianie bywa nazywane **stemowaniem**. Metoda ta ma sens dla języków o ograniczonej fleksji, ale dla polskiego daje wyniki wysoce niezadowalające. W kontekście Morfeusza mówimy więc o prawdziwym hasłowaniu.

Operacją odwrotną do analizy morfologicznej jest **synteza morfologiczna** — utworzenie wykładnika formy odmiany danej przez wskazanie lematu (identyfikatora leksemu) i żądanej charakterystyki fleksyjnej.

Paweł Lula, Katedra Systemów Obliczeniowych, Uniwersytet Ekonomiczny w Krakowie

20

Morfeusz



Paweł Lula, Katedra Systemów Obliczeniowych, Uniwersytet Ekonomiczny w Krakowie

21

<http://www.ipipan.waw.pl/~wolinski/publ/znakowanie.pdf>

MARCIN WOLIŃSKI

System znaczników morfosyntaktycznych w korpusie IPI PAN

Niniejszy artykuł opisuje zasady znakowania¹ morfosyntaktycznego tekstów języka polskiego przyjęte dla korpusu tekstów tworzonego w ramach projektu 7 T11C 043 20 finansowanego przez Komitet Badań Naukowych w latach 2001–2004 i realizowanego w Instytucie Podstaw Informatyki PAN pod kierunkiem Adama Przepiórkowskiego.


Omawiany system znaczników opracowali Adam Przepiórkowski i Marcin Woliński. Pracy tej sekundowali Łukasz Dębowski i Elżbieta Hajnicz. W końcowej fazie do dyskusji włączył się Zygmunt Saloni.

Znaczniki programu Morfeusz


Paweł Lula, Katedra Systemów Obliczeniowych, Uniwersytet Ekonomiczny w Krakowie

22

Spejd - <http://ws.clarin-pl.eu/spejd.shtml>



CLARIN-PL
Common Language Resources and Technology Infrastructure



CENTRUM TECHNOLOGII
JEZYKOWYCH **CLARIN-PI**

Morpho Tager Chunker Ner Serel Spatial **Spejd** Parser WSD Indeks usług API

Spejd

Interfejs webowy dla narzędzia Spejd. Wykorzystane narzędzia:

- Konwerter plików do tekstu Apache Tika, analizator morfologiczny Morfeusz 2 ze słownikiem SGJP, tager WCRFT2
- [Spejd](#)

Instrukcja ▾

Ten wykład jest nudny!

Analizuj
Wyczyść

Paweł Lula, Katedra Systemów Obliczeniowych, Uniwersytet Ekonomiczny w Krakowie

23

Spejd - <http://ws.clarin-pl.eu/spejd.shtml>

```


XML
1 <?xml version="1.0" encoding="UTF-8"?>
2 <teiCorpus xmlns="http://www.tei-c.org/ns/1.0" xmlns:nkjp="http://www.nkjp.pl/ns/1.0" xmlns:xi="http://www.w3.org/2001/XInclude">
3 <TEI>
4 <text>
5 <body>
6 <p xml:id="p-1">
7 <s corresp="ann_segmentation.xml#segm_p-1.1-s" xml:id="p-1.1-s">
8 <seg corresp="ann_segmentation.xml#segm_p-1.1-seg" xml:id="morph_p-1.1-seg">
9 <fs type="morph">
10 <f name="orth">
11 <string>Ten</string>
12 </f>
13 <f name="interps">
14 <fs type="lex" xml:id="morph_1.1.1.1-lex">
15 <f name="base">
16 <string>ten</string>
17 </f>
18 <f name="ctag">
19 <symbol value="adj"/>
20 </f>

```

Paweł Lula, Katedra Systemów Obliczeniowych, Uniwersytet Ekonomiczny w Krakowie

24

<http://zil.ipipan.waw.pl/Spejd/>



Search: [Titles](#) [Text](#) [Login](#)

Spejd [Locked](#) [History](#) [Actions](#)

Menu

- Linguistic Engineering Group
- IPi PAN
- CLIP

Spejd / ♠ / SPADE

This page offers the official [GNU General Public License](#) release of Spejd, a tool for partial parsing and rule-based morphosyntactic disambiguation. By downloading the Spejd package you accept the conditions of that licence.

Principal developer: [Bartosz Zaborowski](#)
License: GPL v.3

Documentation

Readme file of the 1.3.6 version, in English:

- pdf version [README.pdf](#)
- text version [README.txt](#)

Detailed user manual for the 1.3.6 version, in English:

- pdf version [spejd-manual.pdf](#)

Paweł Lula, Katedra Systemów Obliczeniowych, Uniwersytet Ekonomiczny w Krakowie

25

<http://morfologik.blogspot.com/>

Morfologik

Strona projektu morfologik - analizator morfologiczny + słownik morfologiczny + korektor gramatyczny + biblioteki

27.3.16

Walery Pisarek — Słownik języka niby-polskiego

Małopolska Biblioteka Cyfrowa udostępnia [książkę profesora Walerego Pisarka Słownik języka niby-polskiego w formacie PDF](#). Co prawda, opisane w niej błędy w większości pochodzą z prasy z lat siedemdziesiątych, ale wiele błędów typowych dla napuszonego i pretensjonalnego języka niestety trzyma się nadal w polszczyźnie.

Autor [Marcin Miłkowski](#) o 27.3.16 [Brak komentarzy](#) [Linki do tego posta](#) [Poleć to w Google](#)

15.2.16

polimorfologik 2.1

Od ostatniego wydania słowników morfosyntaktycznych z serii Morfologik minęło trochę czasu, a warto było wprowadzić trochę kosmetycznych poprawek, m.in. usunąć niepotrzebne formy (takie jak czasowniki „dzielić”, „bożyć” czy „cienić”) oraz dodać trochę geograficznych nazw własnych. Nowe wydanie dostępne jest na [githubie](#): [Polimorfologik 2.1](#). W pliku opis zmian plus wersje tekstowe i binarne słowników.

Autor [Marcin Miłkowski](#) o 15.2.16 [Brak komentarzy](#) [Linki do tego posta](#) [Poleć to w Google](#)

uruchom

[LanguageTool - Java Web Start](#)

pliki

[languageTool - najnowsza wersja](#)

[morfologik-stemming - najnowsza wersja biblioteki](#)

[morfologik - słownik](#)

materiały

[Java API](#)

[Powered by morfologik-stemming](#)

[zgłaszanie błędów w korektorze](#)

[gromadzimy błędy językowe](#)

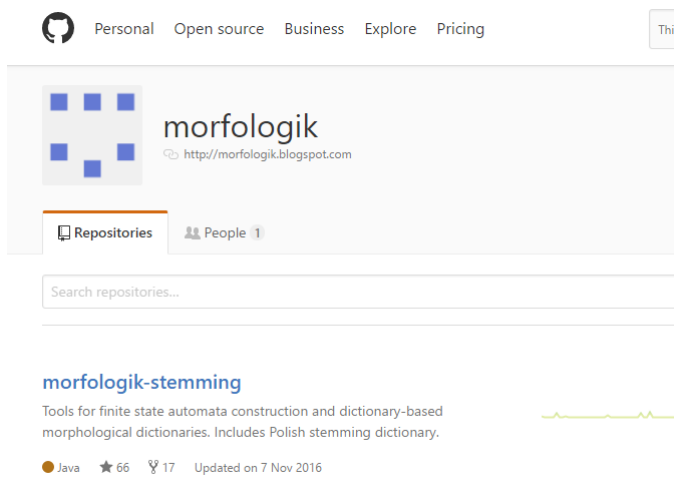
poradnie

[Dziękuję Panu Miłkowski](#)

Paweł Lula, Katedra Systemów Obliczeniowych, Uniwersytet Ekonomiczny w Krakowie

26

<https://github.com/morfologik/>



The screenshot shows the GitHub profile page for the user 'morfologik'. At the top, there are navigation links: Personal, Open source, Business, Explore, and Pricing. The user's profile picture is a 3x3 grid of blue squares. The username 'morfologik' is displayed, along with a link to their website: <http://morfologik.blogspot.com>. Below the profile information, there are tabs for 'Repositories' and 'People'. The 'Repositories' tab is selected, showing a search bar and a list of repositories. The first repository listed is 'morfologik-stemming', which is described as 'Tools for finite state automata construction and dictionary-based morphological dictionaries. Includes Polish stemming dictionary.' It has 66 stars and 17 forks, and was updated on 7 Nov 2016.

Paweł Lula, Katedra Systemów Obliczeniowych, Uniwersytet Ekonomiczny w Krakowie

27

<https://www.r-project.org/>



[\[Home\]](#)

Download

[CRAN](#)

R Project

[About R](#)

[Logo](#)

[Contributors](#)

[What's New?](#)

[Reporting Bugs](#)

[Development Site](#)

[Conferences](#)

[Search](#)

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred [CRAN mirror](#).

If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

News

- **R version 3.3.3 (Another Canoe) prerelease versions** will appear starting Friday 2017-02-24. Final release is scheduled for Monday 2017-03-06.

Paweł Lula, Katedra Systemów Obliczeniowych, Uniwersytet Ekonomiczny w Krakowie

28

Text Mining Package



CRAN
[Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

About R
[R Homepage](#)
[The R Journal](#)

tm: Text Mining Package

A framework for text mining applications within R.

Version: 0.6-2
 Depends: R (≥ 3.1.0), [NLP](#) (≥ 0.1-6.2)
 Imports: parallel, [slam](#) (≥ 0.1-31), stats, tools, utils, graphics
 Suggests: [filehash](#), methods, Rcampdf, [Rgraphviz](#), [Rpoppler](#), [Sj](#)
 Published: 2015-07-03
 Author: Ingo Feinerer [aut, cre], Kurt Hornik [aut], Artifex Software [aut], Ghostscript [aut]
 Maintainer: Ingo Feinerer <feinerer at logic.at>
 License: [GPL-3](#)
 URL: <http://tm.r-forge.r-project.org/>
 NeedsCompilation: no

Tworzenie korpusu

```
library(tm)
```

```
katalog <- "C:/Projects/P102/"
```

```
korpus <- VCorpus(DirSource(katalog,encoding="UTF-8"), readerControl =  
list(reader=readPlain))
```

```
korpus<-tm_map(korpus,removeNumbers)  
stoplista <-readLines("C:/Projects/P102/stoplista_PL.txt",encoding="UTF-8")
```

```
korpus<-tm_map(korpus,removeWords,stoplista)
```

Wstępne przetworzenie korpusu

```
docs <- Corpus(DirSource(cname))
```

```
docs <- tm_map(docs, removePunctuation)
# inspect(docs[3]) # Check to see if it worked.
```

```
docs <- tm_map(docs, removeNumbers)
# inspect(docs[3]) # Check to see if it worked.
```

Wstępne przetworzenie korpusu

```
docs <- tm_map(docs, tolower)
# inspect(docs[3]) # Check to see if it worked.
```

```
docs <- tm_map(docs, removeWords, stopwords("english"))
# inspect(docs[3]) # Check to see if it worked.
```

```
docs <- tm_map(docs, removeWords, c("department", "email"))
# Just replace "department" and "email" with words that you would like to remove.
```


Tworzenie macierzy częstości

```
dtm <- DocumentTermMatrix(docs)
dtm
```

```
## A document-term matrix (6 documents, 2197 terms)
##
## Non-/sparse entries: 3867/9315
## Sparsity           : 71%
## Maximal term length: 40
## Weighting          : term frequency (tf)
```

Tworzenie macierzy częstości

```
tdm <- TermDocumentMatrix(docs)
tdm
```

```
## A term-document matrix (2197 terms, 6 documents)
##
## Non-/sparse entries: 3867/9315
## Sparsity           : 71%
## Maximal term length: 40
## Weighting          : term frequency (tf)
```

Eksport do pliku CSV

```
If you prefer to export the matrix to Excel:  
m <- as.matrix(dtm)  
dim(m)  
write.csv(m, file="dtm.csv")
```

Ważenie macierzy częstości

```
dtm <- DocumentTermMatrix(corpus, control = list(weighting = weightTfIdf))
```

Podstawowe metody ważenia:

- weightBin
- weightTf
- weightTfIdf