

Podstawy maszynowego uczenia na przykładzie klasyfikacji tekstu

Sztuczna Inteligencja i Inżyniera Wiedzy. Ćwiczenie 4

Michał Hetmańczuk

27 05 2020

Sprawozdanie powstało jak dokument typu R Markdown.

Etap 1

Data processing

W pierwszej kolejności rozpakowano piliki zip.

Następnie umieszczono zawartość plików w strukturze danych typu Data Frame. Z nazwy pliku wyekstrahowano kategorię dokumentu. Referencja na tak utworzoną strukturę. została nazwana 'texts_df'.

Opis struktury:

```
str(texts_df, vec.len = 0)
```

```
## 'data.frame': 9837 obs. of 2 variables:
## $ category: Factor w/ 34 levels "Albania","Amerykanscy-prozaicy",...: NULL ...
## $ text : chr ...
```

```
levels(texts_df$category)
```

```
## [1] "Albania"
## [2] "Amerykanscy-prozaicy"
## [3] "Arabowie"
## [4] "Astronautyka"
## [5] "Choroby"
## [6] "Egipt"
## [7] "Ekologia-roslin"
## [8] "Filmy-animowane"
## [9] "Galezie-prawa"
## [10] "Gry-komputerowe"
## [11] "Karkonosze"
## [12] "Katolicyzm"
## [13] "Komiksy"
## [14] "Komputery"
## [15] "Kotowate"
## [16] "Kultura-Chin"
## [17] "Monety"
## [18] "Muzyka-powazna"
## [19] "Narcciarstwo"
```

```
## [20] "Narkomania"
## [21] "Niemieccy-wojskowi"
## [22] "Optyka"
## [23] "Pierwiastki-chemiczne"
## [24] "Pilka-nozna"
## [25] "Propaganda-polityczna"
## [26] "Rachunkowosc"
## [27] "Samochody"
## [28] "Samoloty"
## [29] "Sporty-silowe"
## [30] "System-opieki-zdrowotnej-w-Polsce"
## [31] "Szachy"
## [32] "Wojska-pancerne"
## [33] "Zegluga"
## [34] "Zydzi"
```

Aby ujednolicić tekst oraz wstępnie usunąć potencjalnie zbędne cechy, dane zostały wyczyszczone poprzez usunięcie: wielkich liter, liczb, znaków interpunkcyjnych, zbędnych białych znaków oraz wyrazów ze stop-listy (plik stopwords.txt z polską stop-listą został zaczerpnięty z <https://github.com/bieli/stopwords/blob/master/polish.stopwords.txt>). Aby wykorzystać w tym celu funkcję `tm_map`, zawartość dokumentów umieszczono w przeznaczony do tego strukturze - `VCorpus` (taki typ danych jest obsługiwany przez `tm_map`).

```
text_labels <- texts_df$category
text_corpus <- VCorpus(VectorSource(texts_df$text))
text_corpus_clean <- text_corpus %>%
  tm_map(content_transformer(tolower)) %>%
  tm_map(removeNumbers) %>%
  tm_map(removePunctuation) %>%
  tm_map(stripWhitespace)
stopwords.pl <- readLines(".\\stopwords.txt", encoding = 'UTF-8')
text_corpus_clean <- tm_map(text_corpus_clean, removeWords, stopwords.pl)
```

Wykorzystując tak wyczyszczone korpusy dokumentów utworzono macierz pojęć, zawierającą częstotliwości występowania danego pojęcia w danym dokumencie.

```
text_dtm <- DocumentTermMatrix(text_corpus_clean)
text_dtm
```

```
## <<DocumentTermMatrix (documents: 9837, terms: 264253)>>
## Non-/sparse entries: 1524005/2597932756
## Sparsity           : 100%
## Maximal term length: 173
## Weighting           : term frequency (tf)
```

Przykładowe wizualizacje przygotowanych danych. Kategorie: "Albania", "Choroby", "Astronautyka"



Propozycja metody selekcji cech i implementacji modeli

Ze względu na fakt, iż mamy do czynienia z szukaniem zależności między dwoma kategorycznymi zmiennymi, selekcja cech zostanie przeprowadzona z wykorzystaniem testów χ^2 . Implementacja klasyfikatora Naiwnego Bayesa zostanie zaczerpnięta z biblioteki “fastNaiveBayes”. Biblioteka tak obsługuje wszystkie rodzaje klasyfikatora (Bernoulli, Gaussian, Multinomial), co może być bardzo przydatne z perspektywy przeprowadzania badań. Wykorzystana zostanie również implementacja drzewa decyzyjnego z pakietu “tree”.