

RAPORT DATA MINING – EKSPLORACJA TEKSTU W

ŚRODOWISKU R

Emil Filipowicz

Raport eksploracji danych tekstowych w języku angielskim zawierających informację z obszaru wiedzy opartego o sztuczną inteligencję (ang. AI). W raporcie przedstawione zostanie etapowe przekształcanie tekstu, komendy użyte do wykonania określonych czynności oraz wnioski z wykonanej analizy.

1.ZAŁADOWANE PAKIETY R

- library(tm)
- library(SnowballC)
- library(ggplot2)
- library(wordcloud)

2.PRZYGOTOWANIE ZBIORU DANYCH

2.1. W pierwszym kroku eksploracja danych tekstowych rozpoczęta została od załadowania katalogu z korpusem plików, aby to zrobić została użyte komendy:

- setwd() – określa katalog roboczy określony ścieżką

```
setwd("/Users/emilfilipowicz/Desktop/korpus")
```

- getwd() – określa bieżący katalog roboczy – sprawdzenie katalogu roboczego.

```
> getwd()  
[1] "/Users/emilfilipowicz/Desktop/korpus"
```

2.2. Tworzenie zmiennej z katalogiem roboczym i sprawdzenie jej zawartości za pomocą komendy dir()

- tworzenie zmiennej

```
wd <- "/Users/emilfilipowicz/Desktop/korpus"
```

- sprawdzenie zawartości katalogu

```
> dir(wd)
[1] "Dangerous AI.txt"           "finance AI.txt"
[3] "Health care and criminal justice.txt" "Myths AI.txt"
[5] "Qualities AI.txt"          "safe AI.txt"
[7] "Security AI.txt"           "Smart citiesAI.txt"
[9] "Start AI.txt"              "Transportation AI.txt"
```

Funkcja sprawdziła zawartość katalogu roboczego. W nim znajdują się 10 plików tekstowych.

2.3. Utworzenie zbioru dokumentów za pomocą funkcji Corpus() pod zmienną docs.

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

Po utworzeniu zmiennej docs, została ona wywołana przekazując kolejne informacje o korpusie z danymi tekstowymi.

```
> docs <- Corpus(DirSource(wd))
> docs
<<SimpleCorpus>>
Metadata: corpus specific: 1, document level (indexed): 0
Content: documents: 10
```

Zmienna pokazuje ogólne informacje o katalogu, specyfikację, poziom dokumentu oraz ilość dokumentów jaka znajduje się pliku. Analizie poddane zostanie 10 dokumentów tekstowych.

2.4. Dodatkowo do eksploracji danych zwartych w katalogu użyta została funkcja writeLines(), która umożliwia sprawdzenie konkretnego dokumentu w korpusie.

```
writeLines(as.character(docs[[1]]))
```

2.5. SPRAWDZENIE ILOŚCI SŁÓW ORAZ ZNAKÓW W ANALIZOWANYM KORPUSIE.

- Słowa

```
> dtm
<<DocumentTermMatrix (documents: 10, terms: 1767)>>
Non-/sparse entries: 2463/15207
Sparsity             : 86%
Maximal term length: 22
Weighting             : term frequency (tf)
```

W badanym korpusie przed podjęciem wstępnej obróbki danych ilość słów wynosiła 1767.

-

ZNAKI

```
> corpus_chars <- 0
>
> for(i in seq(docs)){
+   corpus_chars <- corpus_chars + nchar(content(docs[[i]]))
+ }
>
> corpus_chars
[1] 28756
```

Do zliczenia znaków w korpusie użyta została funkcja for, która wykazała, że w całym korpusie przed wstępną obróbką danych znajdowało się 28756 znaków.

3.WSTĘPNA OBRÓBKA DANYCH

Wstępna analiza danych obejmuje:

- usuwanie znaków interpunkcyjnych
- usuwanie liczb
- przekształcanie słów w dokumencie na małe litery
- pomijanie słów powszechnych
- Odfiltrowywanie niechcianych terminów
- usuwanie białych znaków w dokumencie

Do przeprowadzenia wstępnej analizy dokumentu została użyta funkcja getTranformations(), która umożliwia przeprowadzenie czynności przedstawionych powyżej.

```
> getTranformations()
[1] "removeNumbers"      "removePunctuation" "removeWords"        "stemDocument"       "stripWhitespace"
```

1.3 . USUWANIE ZNAKÓW INTERPUNKCYJNYCH (PUNCTATION MARKS AND NUMBERS)

Usuwanie znaków interpunkcyjnych zachodzi po wprowadzeniu kodu:

```
docs<-tm_map(docs,removePunctuation)
docs<-tm_map(docs, removeNumbers)
#docs<-tm_map(docs, PlainTextDocument)
```

Następnie sprawdzono czy wszystkie dokumenty w zawarte w katalogu zostały wyczyszczone z znaków interpunkcyjnych. Do sprawdzenia została użyta funkcja `writeLines()`.

```
writeLines(as.character(docs[[3]]))
```

Oto część tekstu, wyczyszczona z znaków i wywołana przez funkcję powyżej:

How can AI be dangerous?

Most researchers agree that a superintelligent AI is unlikely to exhibit human emotions like love or hate and that there is no reason to expect AI to become intentionally benevolent or malevolent. Instead, when considering how AI might become a risk, experts think two scenarios most likely.

The AI is programmed to do something devastating. Autonomous weapons are artificial intelligence systems that are programmed to kill. In the hands of the wrong person, these weapons could easily cause mass casualties. Moreover, an AI arms race could be disastrous.

Dla porównania, przedstawiono również to samą część tekstu przed wyczyszczeniem:

How can AI be dangerous?

Most researchers agree that a superintelligent AI is unlikely to exhibit human emotions like love or hate, and that there is no reason to expect AI to become intentionally benevolent or malevolent. Instead, when considering how AI might become a risk, experts think two scenarios most likely:

The AI is programmed to do something devastating: Autonomous weapons are artificial intelligence systems that are programmed to kill. In the hands of the wrong person, these weapons could easily cause mass casualties. Moreover, an AI arms race could be disastrous.

Jak widać udało się przeprowadzić pierwszą część obróbki danych, z całego korpusu plików tekstowych zniknęły znaki interpunkcyjne. Co pozwala przejść do kolejnego etapu.

2.3. USUWANIE ZNAKÓW SPECJALNYCH

Kolejny etap, który został objęty działaniem podczas wstępnej obróbki danych to usuwanie znaków specjalnych. Do tego została użyta pętla `for`, która dodatkowo usunęła znaki nie wykluczone przez poprzednią funkcję. Przedstawia się ona w następujący sposób:

```
for (j in seq(docs)) {  
  docs[[j]] <-gsub("/", " ", docs[[j]])  
  docs[[j]] <-gsub("@", " ", docs[[j]])  
  docs[[j]] <-gsub("-", " ", docs[[j]])  
  docs[[j]] <-gsub("'", " ", docs[[j]])  
  docs[[j]] <-gsub("\"", " ", docs[[j]])  
  docs[[j]] <-gsub("...", " ", docs[[j]])  
  docs[[j]] <-gsub("`", " ", docs[[j]])  
  docs[[j]] <-gsub(">", " ", docs[[j]])  
  docs[[j]] <-gsub("<", " ", docs[[j]])  
}  
#docs<-tm_map(docs, PlainTextDocument)  
writeLines(as.character(docs[[1]]))
```

Wywołany plik 1 nie posiada przy tym etapie żadnych znaków interpunkcyjnych oraz znaków specjalnych, to oznacza, że na tym etapie cały korpus danych jest wyczyszczony z tej kategorii obiektów tekstowych.

3.3. PRZEKSZTAŁCENIE PLIKÓW TEKSTOWYCH NA MAŁE LITERY.

Ze względu na rozróżnianie przez środowisko R małych i dużych liter warto, aby globalnie przetransformować pliki tekstowe na tą samą wielkość liter. Do wykonania tej czynności został użyta komenda:

```
docs<-tm_map(docs, tolower)
#docs<-tm_map(docs, PlainTextDocument)
writeLines(as.character(docs[[1]]))
```

Po użyciu przedstawionej funkcji w całym korpusie doszło do transformacji dużych liter w małe litery. Co sprawiło, że na tym etapie tekst wygląda w następujący sposób:

how can ai be dangerous

most researchers agree that a superintelligent ai is unlikely to exhibit human emotions like love or hate and that there is no reason to expect ai to become intentionally benevolent or malevolent instead when considering how ai might become a risk experts think two scenarios most likely

the ai is programmed to do something devastating autonomous weapons are artificial intelligence systems that are programmed to kill in the hands of the wrong person these weapons could easily cause mass casualties moreover an ai arms race could

3.4. USUWANIE NAJCZĘSTSZYCH SŁÓW WYSTĘPUJĄCYCH W KORPUSIE.

Na tym etapie tekstów usunięte zostały przedimki, spójniki, czasowniki pospolite oraz kwalifikatory, przy użyciu komendy:

```
docs<-tm_map(docs, removeWords, stopwords("English"))
#docs<-tm_map(docs, PlainTextDocument)
writeLines(as.character(docs[[1]]))
```

Po użyciu danej funkcji tekst został oczyszczony z określonych elementów i przedstawia się w następujący sposób:

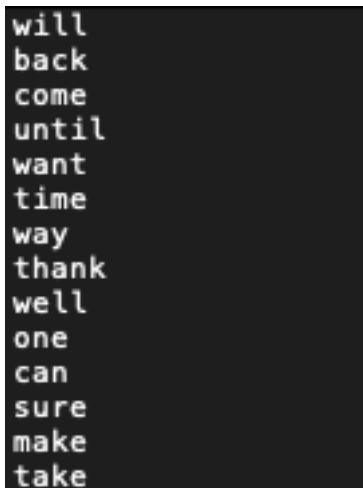
can ai dangerous

researchers agree superintelligent ai unlikely exhibit human emotions like love hate reason expect ai become intentionally benevolent malevolent instead considering ai might become risk experts think two scenarios likely

ai programmed something devastating autonomous weapons artificial intelligence systems programmed kill hands wrong person weapons easily cause mass casualties moreover ai arms race inadvertently lead ai war also results may

3.5. USUWANIE DODATKOWYCH SŁÓW

Ponad to po zlikwidowaniu popularnych słów, ta czynność została rozszerzona o usunięcie dodatkowych słów. Którymi były słowa:



```
will  
back  
come  
until  
want  
time  
way  
thank  
well  
one  
can  
sure  
make  
take
```

Zostały one usunięte za pomocą określonej metody:

```
StW<-read.table("/Users/emilfilipowicz/Desktop/StopWords.txt")  
StW
```

```
StWW<-as.character(StW$V1)  
StWW
```

```
docs <-tm_map(docs, removeWords, StWW)  
#docs<-tm_map(docs, PlainTextDocument)  
writeLines(as.character(docs[[1]]))
```

W ten sposób z korpusu zniknęły dodatkowe słowa zawarte na liście powyżej.

3.6. USUWANIE BIAŁYCH ZNAKÓW

Do usunięcia białych znaków z tekstu użyta została komenda:

```
docs <-tm_map(docs, stripWhitespace)  
#docs<-tm_map(docs, PlainTextDocument)  
writeLines(as.character(docs[[1]]))
```

Wywołany tekst po usunięciu białych znaków wygląda w następujący sposób:

```
> writeLines(as.character(docs[[1]]))
artificial intelligence transforming world people familiar concept artificial intelligence ai illustration senior busi
ness leaders united states asked ai percent said familiar number affect particular companies understood considerable po
tential altering business processes clear ai deployed within organizations despite widespread lack familiarity ai techn
ology transforming every walk life wideranging tool enables people rethink integrate information analyze data use resul
ting insights improve decisionmaking hope comprehensive overview explain ai audience policymakers opinion leaders inter
ested observers demonstrate ai already altering world raising important questions society economy governance paper disc
uss novel applications finance national security health care criminal justice transportation smart cities address issue
s data access problems algorithmic bias ai ethics transparency legal liability ai decisions contrast regulatory approac
hes us european union close making number recommendations getting ai still protecting important human values order maxi
mize ai benefits recommend nine steps going forward encourage greater data access researchers without compromising user
sd personal privacy invest government funding unclassified ai research promote new models digital education ai workforc
e development employees skills needed stcentury economy create federal ai advisory committee make policy recommendation
s engage state local officials enact effective policies regulate broad ai principles rather specific algorithms take bi
as complaints seriously ai replicate historic injustice unfairness discrimination data algorithms maintain mechanisms h
uman oversight control penalize malicious ai behavior promote cybersecurity
```

3.7. STEAMING

Usuwanie końcówki fleksyjnej, przekształcenie słów do ich form pierwotnych:

Etap 1:

```
library(SnowballC)
stemDocument("modelling", language = "english")
stemDocument("modeller", language = "english")
stemDocument("models", language = "english")
```

Etap 2:

```
for (j in seq(docs)) {
  docs[[j]]<-stemDocument(docs[[j]], language = "english")
} #docs <-tm_map(docs, PlainTextDocument)
writeLines(as.character(docs[[1]]))
```

Tekst po przeprowadzeniu steamingu:

```
ai danger research agre superintellig ai unlik exhibit human emot like love hate reason expect ai becom intent benevol male
vol instead consid ai might becom risk expert think two scenario like ai program someth devast autonom weapon artifici inte
llig system program kill hand wrong person weapon easili caus mass casualti moreov ai arm race inadvert lead ai war also re
sult mass casualti avoid thwart eneml weapon design extrem difficult simpli dturn offd human plausibl lose control situat r
```

Steaming był ostatnim etapem wstępnej obróbki danych, w kolejnym etapie ponownie przeprowadzono sprawdzenie ilości słów w korpusie.

4. SPRAWDZENIE ILOŚCI SŁÓW ORAZ ZNAKÓW W WYCZYSZCZONYM TEKŚCIE.

- Słowa

```
> dtm
<<DocumentTermMatrix (documents: 10, terms: 1170)>>
Non-/sparse entries: 1781/9919
Sparsity           : 85%
Maximal term length: 19
Weighting          : term frequency (tf)
```

Po wykonaniu wstępnej obróbki danych w badanym korpusie, ilość słów w całym katalogu spadła do 1170 słów.

- Znaki

```
> corpus_chars <- 0
>
> for(i in seq(docs)){
+   corpus_chars <- corpus_chars + nchar(content(docs[[i]]))
+ }
>
> corpus_chars
[1] 17096
```

Po wykonaniu wstępnej obróbki danych w badanym korpusie, ilość słów w całym katalogu spadła do 17096 znaków.

5. ANALIZA KORPUSU

5.1 Częstotliwość najczęściej występujących słów w korpusie.

- 14 najczęściej występujących słów w korpusie

```
> head(freq, 14)
  human intellig system data technolog goal research machin car citi learn
    27      25     22   22      16    15      14     14   12    11     11
  mani  advanc autonom
    11      10      10
```

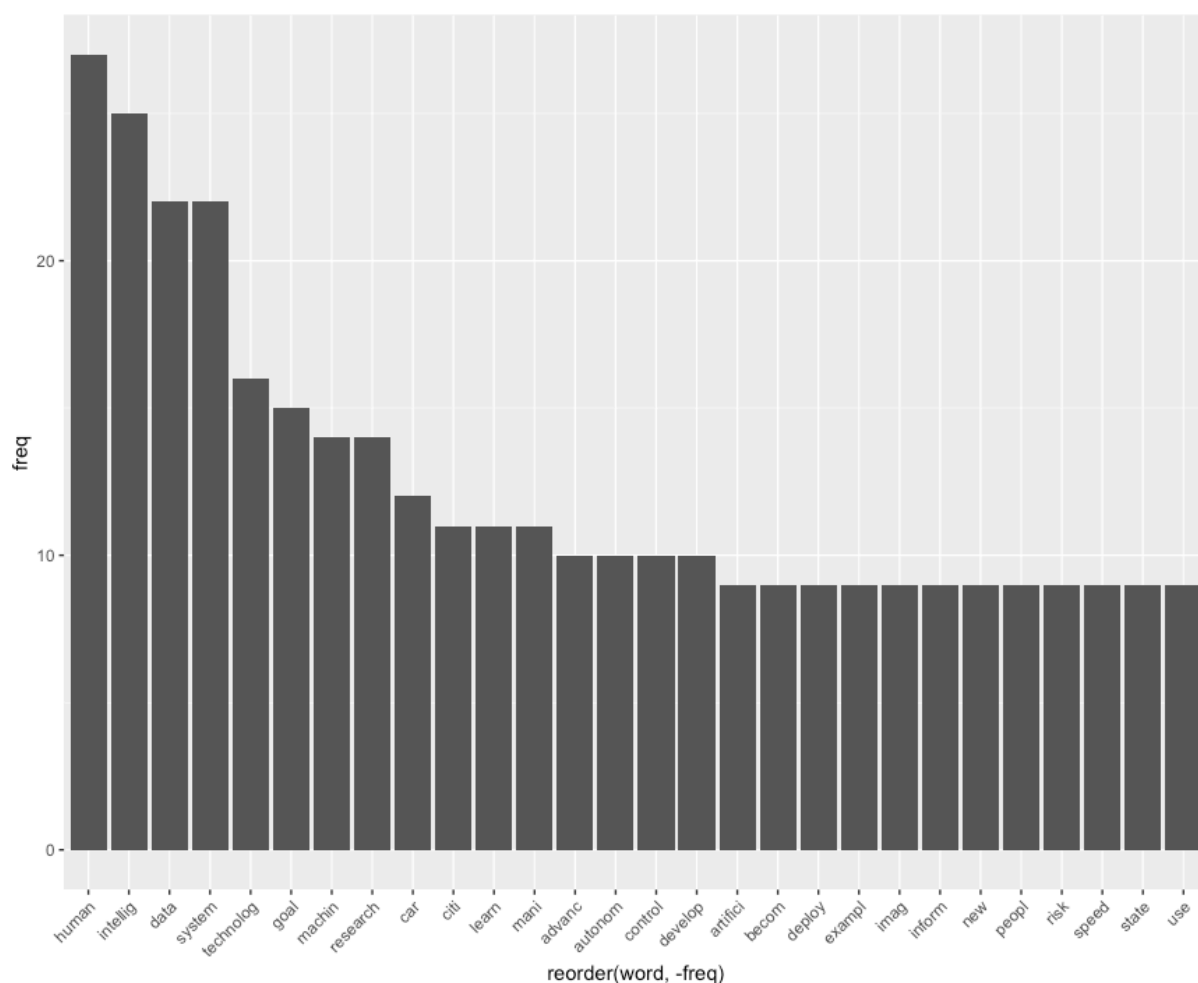
- słowa występujące w korpusie więcej niż 20 razy

```
> wd2 <- findFreqTerms(dtmr, lowfreq=20)
> wd2
[1] "human"    "intellig" "system"   "data"
```


Występująca najczęściej słowa wskazują na tematykę korpusu. Na tym etapie można stwierdzić, że dane tekstowe zawierają treści odnoszące się do obszarów wiedzy opartych na inteligentnej technologii związanych z ludźmi.

6. PRZEDSTAWIENIE GRAFICZNE ANALIZOWANEGO KORPUSU

6.1 Histogram częstotliwości występujących słów (Zipf's law).



Histogram pokazujący 28 słów występujących w korpusie, ich ilość oraz częstotliwość, ułożonych w kolejności od najczęściej występujących do tych mniej. Wskazują identyczną tendencję jak w pod punkcie 5.1. z wyjątkiem kolejności słów „machin” i „research” jednak należy zwrócić uwagę, że słowa te posiadają taką samą ilość.

6. Wordcloud

- wordcloud w kolorze z występowaniem słów kluczowych co najmniej do 70 razy.

[illegible]

Wnioski i interpretacja

- po przeprowadzeniu analizy korpusu, zwracając uwagę na ilość występujących słów i znaków oraz po wstępnej obróbce tekstu udało się wykazać, które słowa nadają sens oraz kontekst w badanym korpusie oraz w jakiej częstotliwości słowa te występują.

- Teksty znajdujące się w korpusie związane były z obszarem technologii zajmującym się sztuczną inteligencją, poruszały tematykę Sztucznej inteligencji w następujących kategoriach: ryzyko, zagrożenia, mity, możliwości, rozwój, zdrowie, transport, inteligentne miasta oraz finanse. Co zauważyć można w histogramie oraz w przedstawionych wordcloudach.

- Najczęściej występującymi słowami były human, intellig, data i system. Co ciekawe słowo artifi-ci, czyli pierwszy człon rozwinięcia skrótu AI (artificial intelligence), występuje rzadziej. Może to świadczyć o tym, że teksty znajdujące się w korpusie częściej poruszają tematykę AI z punktu społecznego niż technologicznego.

- Jednak to nie jest tak, że wszystkie teksty poruszają tylko kwestie społeczne, w przedstawionych analizach często występują także słowa typowo technologiczne jak data, system, technolog itd. To wskazują także na obecność w korpusie treści skupiające się na samej technologii AI

- Analizując słowa kluczowe, oraz podkategorie słów, można stwierdzić, że teksty zawarte w korpusie poruszają tematykę użytkowania technologii sztucznej inteligencji w społeczeństwie, i mogą odpowiedzieć na pytania:

- jakie korzyści może przynieść AI?
- Jakie dziedziny wiedzy oraz gałęzi biznesu AI może rozwinać?
- Jakie zagrożenia niesie za sobą sztuczna inteligencja?

