Sprawozdanie

Zajęcia: Analiza procesów uczenia
Prowadzący: prof. dr hab. Vasyl Martsenyuk

Laboratorium 6
Data: 24.05.2024

Temat: Problemy NLP w uczeniu maszynowym

Wariant: 1

Agnieszka Białecka
Informatyka II stopień,
stacjonarne,
1 semestr,
Gr.1a

1. Cel ćwiczenia

Celem ćwiczenia była analiza tekstów za pomocą list częstotliwości, chmur słów i n-gramów.

2. Wstęp teoretyczny

3. Przebieg ćwiczenia

Zadanie dotyczy analizy tekstu, w tym listę częstotliwości słów, budowanie chmury słów, kojarzenie, sentiment analysis, emotion analysis, bigramów, grafów powiązań. Warianty zadania są określone tekstem w języku angielskim umieszczonym na portalu en.wikipedia.org

https://en.wikipedia.org/wiki/Machine_learning

```
install.packages("tm")
install.packages("SnowballC")
install.packages("wordcloud")
install.packages("RColorBrewer")
install.packages("syuzhet")
install.packages("ggplot2")

library("tm")
library("SnowballC")
library("wordcloud")
library("RColorBrewer")
library("syuzhet")
library("ggplot2")
```

Zdjęcie 1. Wgranie bibliotek

```
# read file
text <- readLines("Machine learning.txt", warn=FALSE)

TextDoc <- Corpus(VectorSource(text))

toSpace <- content_transformer(function(x, pattern) gsub(pattern, " ", x))
TextDoc <- tm_map(TextDoc, toSpace, "/")
TextDoc <- tm_map(TextDoc, toSpace, "@")
TextDoc <- tm_map(TextDoc, toSpace, "\\|")
TextDoc <- tm_map(TextDoc, toSpace, ":")
TextDoc <- tm_map(TextDoc, toSpace, ";")
TextDoc <- tm_map(TextDoc, toSpace, ",")
TextDoc <- tm_map(TextDoc, toSpace, "/")
TextDoc <- tm_map(TextDoc, removeNumbers)
TextDoc <- tm_map(TextDoc, removeWords, stopwords("english"))
TextDoc <- tm_map(TextDoc, removeWords, c("[", "]"))
TextDoc <- tm_map(TextDoc, removePunctuation)
TextDoc <- tm_map(TextDoc, stripWhitespace)
TextDoc <- tm_map(TextDoc, stemDocument)
TextDoc <- tm_map(TextDoc, content_transformer(tolower))

# build text matrix
TextDoc_dtm <- TermDocumentMatrix(TextDoc)
dtm_m <- as.matrix(TextDoc_dtm)
dtm_v <- sort(rowSums(dtm_m), decreasing = TRUE)
dtm_d <- data.frame(word = names(dtm_v), freq = dtm_v)
head(dtm_d, 5)
```

Zdjęcie 2. Wgranie tekstu oraz wyczyszczenie tekstu z niepotrzebnych znaków

```r
# build text matrix
TextDoc_dtm <- TermDocumentMatrix(TextDoc)
dtm_m <- as.matrix(TextDoc_dtm)
dtm_v <- sort(rowSums(dtm_m), decreasing = TRUE)
dtm_d <- data.frame(word = names(dtm_v), freq = dtm_v)
head(dtm_d, 5)

# plot of most frequent words
barplot(
  dtm_d[1:20, ]$freq,
  las = 2,
  names.arg = dtm_d[1:20, ]$word,
  col = "lightgreen",
  main = "Top 20 most frequent words",
  ylab = "Word frequency"
)
```

Zdjęcie 3. Rysowane wykresu najczęstszych słów

```r
# word associations
findAssocs(
  TextDoc_dtm,
  terms = c("learn", "machine", "algorithm", "train"),
  corlimit = 0.5
)
findAssocs(
  TextDoc_dtm,
  terms = findFreqTerms(TextDoc_dtm, lowfreq = 20),
  corlimit = 0.5
)

# sentiment analysis
syuzhet_vector <- get_sentiment(text, method = "syuzhet")
bing_vector <- get_sentiment(text, method = "bing")
nrc_vector <- get_sentiment(text, method = "nrc")
rbind(
  sign(head(syuzhet_vector)),
  sign(head(bing_vector)),
  sign(head(nrc_vector))
)
```

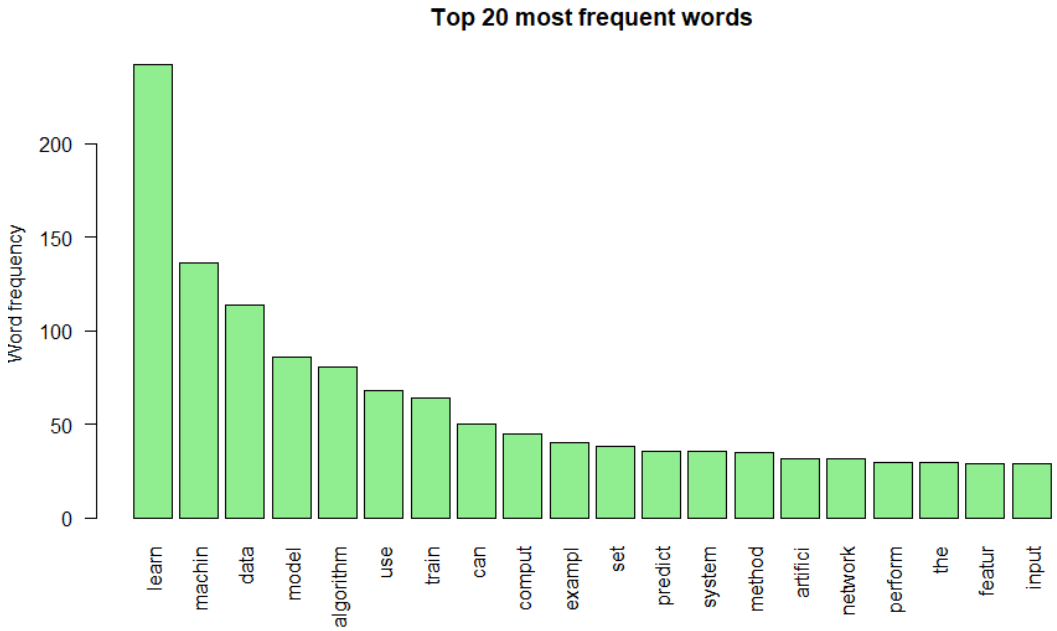
Zdjęcie 4. Kojarzenie słów

```r
# emotion classification
d <- get_nrc_sentiment(as.vector(dtm_d$word))
head(d,10)
td <- data.frame(t(d))
td_new <- data.frame(rowSums(td[1:56]))
names(td_new)[1] <- "count"
td_new <- cbind("sentiment" = rownames(td_new), td_new)
rownames(td_new) <- NULL
td_new2 <- td_new[1:8,]
quickplot(
  sentiment,
  data = td_new2,
  weight = count,
  geom = "bar",
  fill = sentiment,
  ylab = "count"
) + ggtitle("Survey sentiments")
barplot(
  sort(colSums(prop.table(d[, 1:8]))),
  horiz = TRUE,
  cex.names = 0.7,
  las = 1,
  main = "Emotions in Text",
  xlab = "Percentage"
)
```

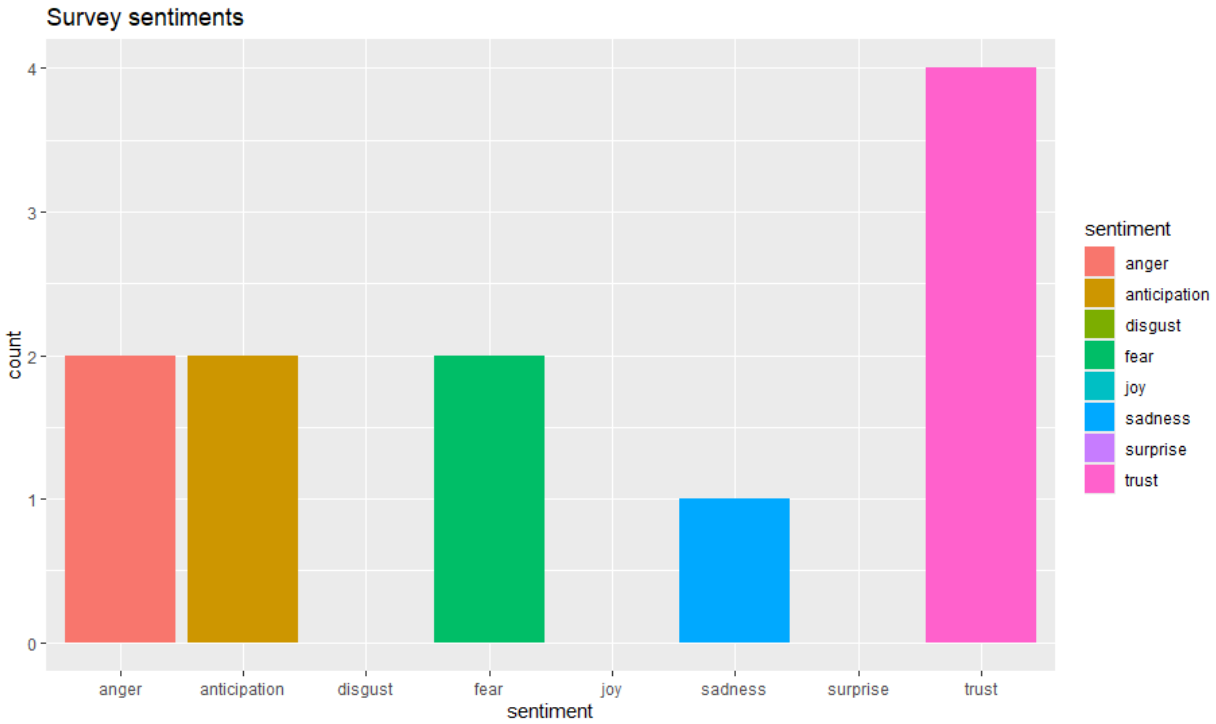
Zdjęcie 5. Klasyfikacja emocji słów

	word	freq
learn	learn	242
machin	machin	136
data	data	114
model	model	86
algorithm	algorithm	81

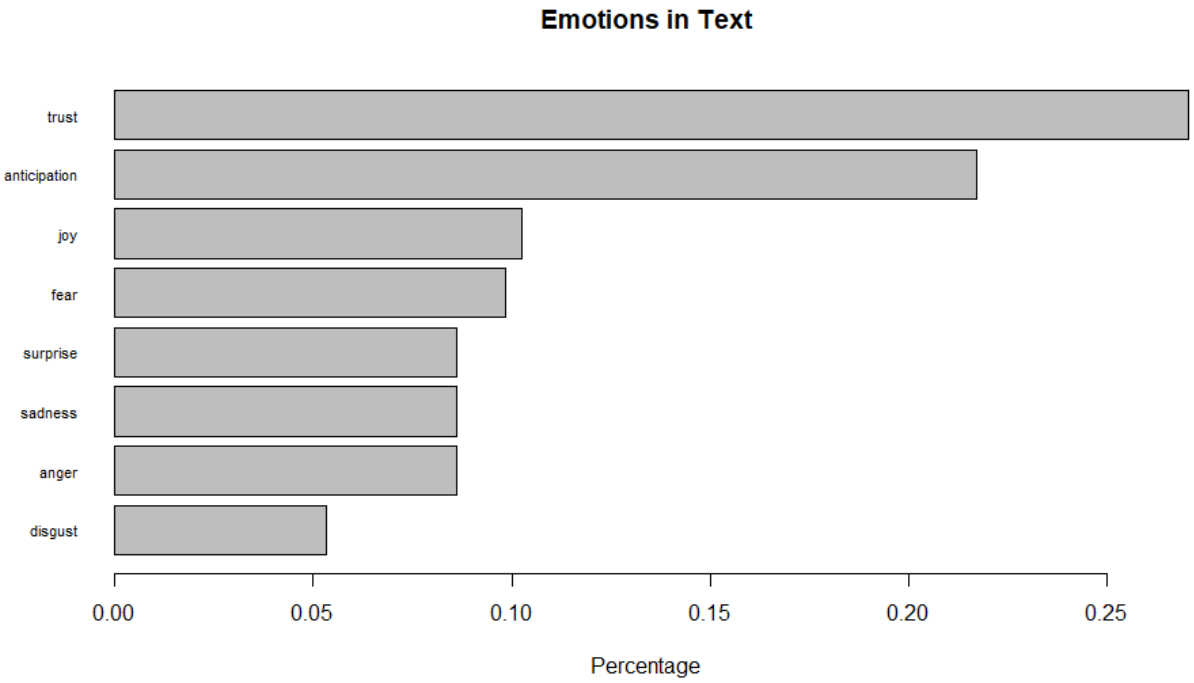
Zdjęcie 6. Najczęstsze słowa



Zdjęcie 7. Wykres 20 najczęstszych słów



Zdjęcie 8. Wykres badania nastrojów



Zdjęcie 9. Wykres emocji w tekście

4. Podsumowanie

Przeprowadzone ćwiczenie pozwoliło na dogłębne zapoznanie się z koncepcją analizy tekstu przy pomocy list częstotliwości. W ramach tego ćwiczenia zostały zaimplementowane modele list częstotliwości, chmur słów i n-gramów.