Laboratorium Analizy Procesów Uczenia.

Data wykonania ćwiczenia:	**24.05.2024**
Rok studiów:	**1**
Semestr:	1
Grupa studencka:	**1b**
Grupa laboratoryjna:	**-**

Ćwiczenie nr	6

Temat: Problemy NLP w uczeniu maszynowym.

Osoby wykonujące ćwiczenia:

1. Gracjan Wackermann

Katedra Informatyki i Automatyki

1. Cel ćwiczenia:

Celem są 1. Wstępna analiza tekstów z pomocą list częstotliwości, chmur słów, n-gramów. 2. Konstruowanie n-gramów i grafów.

2. Zadanie do wykonania:

Zadanie dotyczy analizy tekstu, w tym list częstotliwości słów, budowanie chmury słów, kojarzeń i innych.

3. https://en.wikipedia.org/wiki/Poetry

- Wariant nr. 3 -

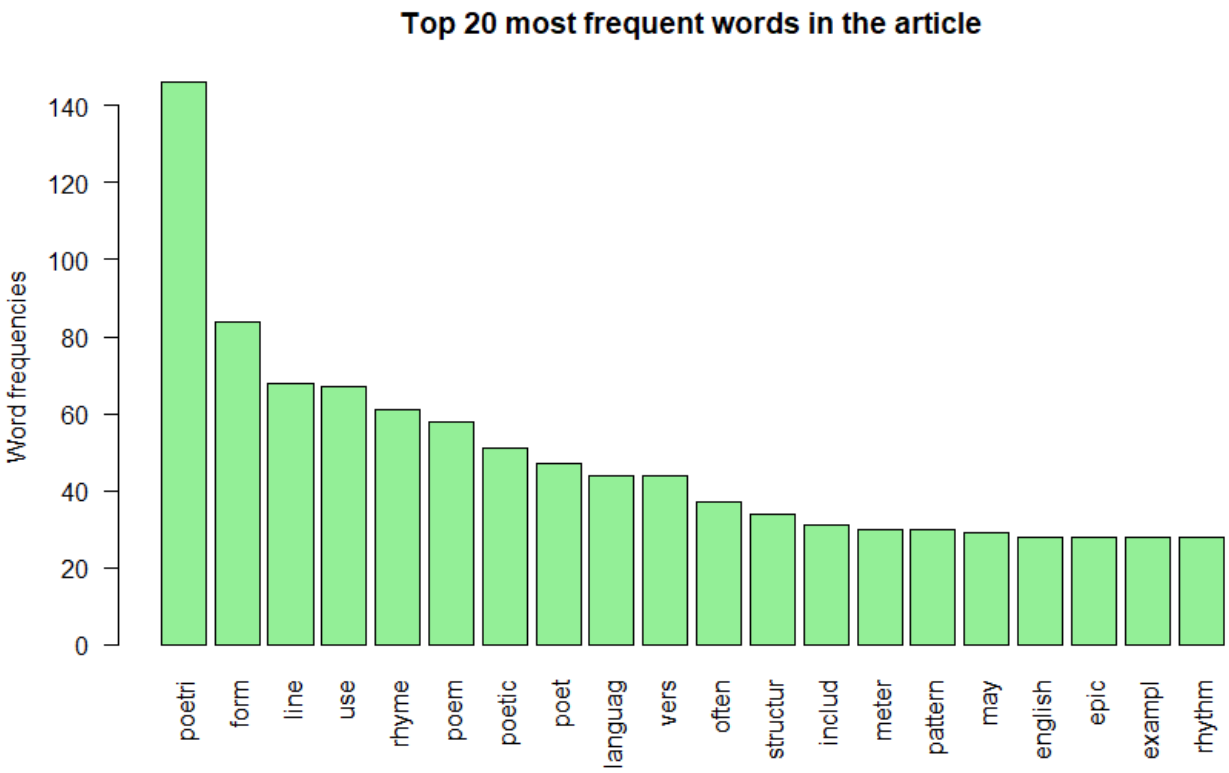
Uzyskany kod:

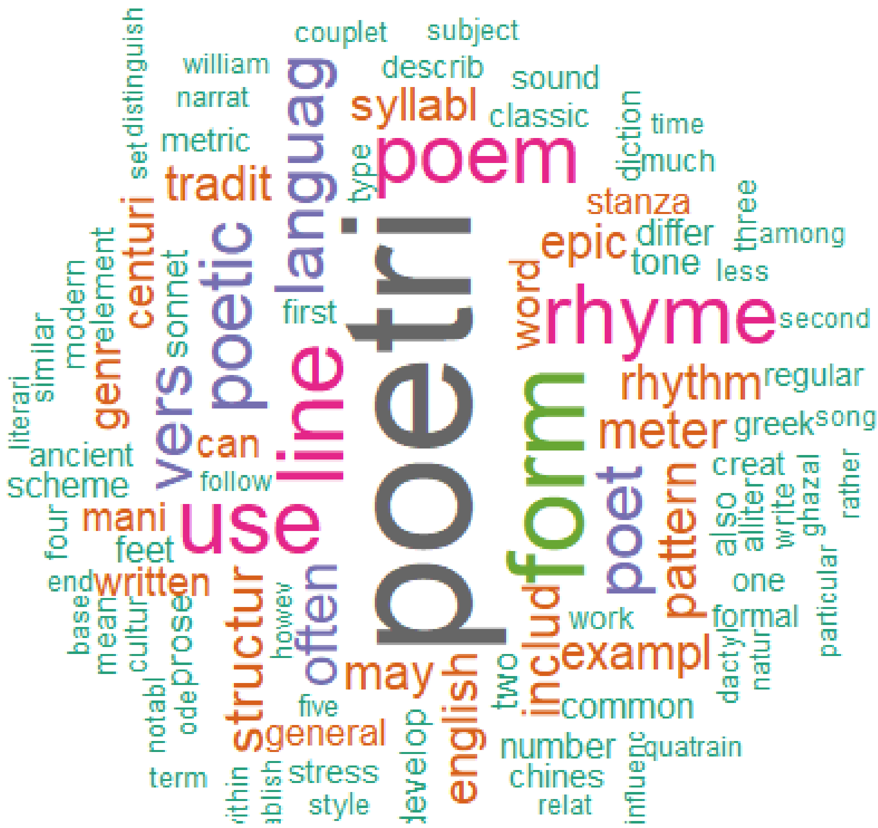
```
# Ładowanie niezbędnych paczek
library("tm")
library("SnowballC")
library("wordcloud")
library("RColorBrewer")
library("syuzhet")
library("ggplot2")
library("tidytext")
library("igraph")
library("ggraph")
library("rvest")
library("dplyr")
library("tidyr")

# Pobranie tekstu z URL
url <- 'https://en.wikipedia.org/wiki/Poetry'
webpage <- read_html(url)
text <- webpage %>%
  html_nodes("p") %>%
  html_text() %>%
  paste(collapse = " ")

# Przekształcenie tekstu do obiektu Corpus
TextDoc <- VCorpus(VectorSource(text))

# "Wyczyszczanie" tekstu
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, content_transformer(tolower))
TextDoc <- tm_map(TextDoc, removeNumbers)
TextDoc <- tm_map(TextDoc, removeWords, stopwords("english"))
TextDoc <- tm_map(TextDoc, removePunctuation)
TextDoc <- tm_map(TextDoc, stripWhitespace)
TextDoc <- tm_map(TextDoc, stemDocument)

# Budowanie macierzy dokumentu
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)

# Wyświetlanie 5 najczęstszych słów
```

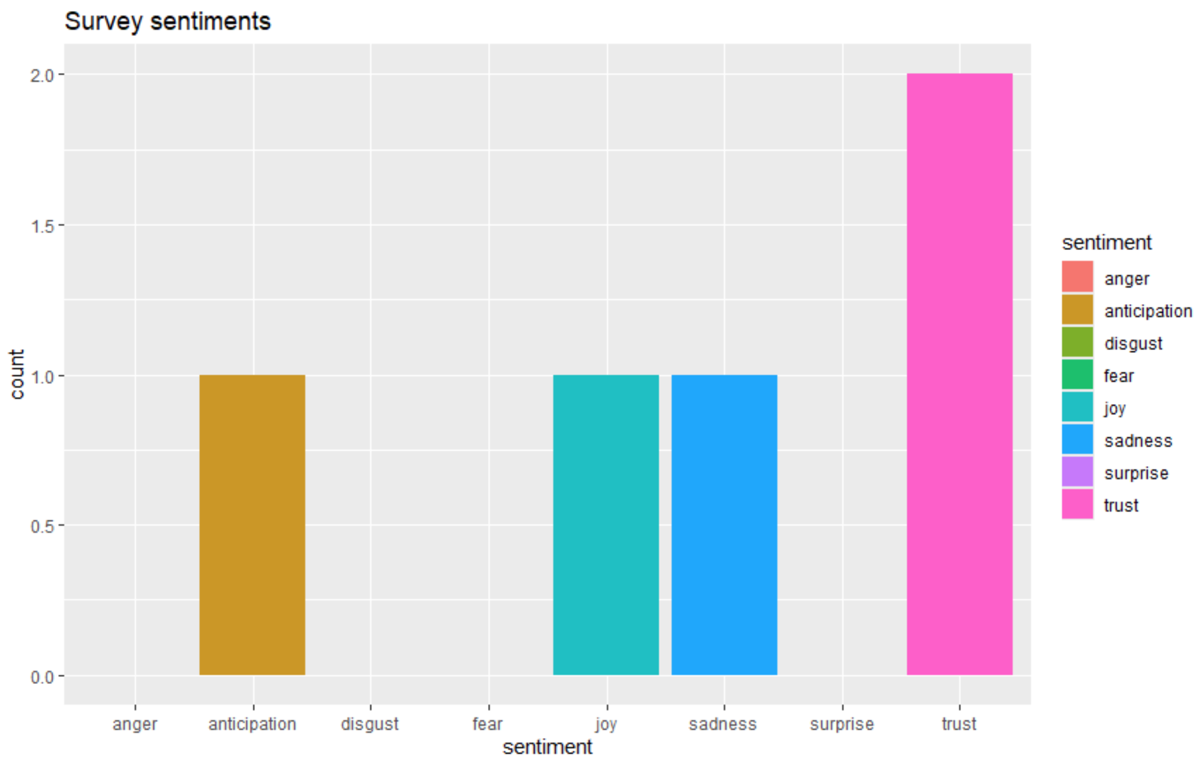
```r
head(dtm_d, 5)

# Wykres słupkowy najczęstszych słów
barplot(dtm_d[1:20,]$freq, las = 2, names.arg = dtm_d[1:20,]$word,
        col ="lightgreen",
        main ="Top 20 most frequent words in the article",
        ylab = "Word frequencies")

# Generowanie chmury słów
set.seed(1234)
wordcloud(words = dtm_d$word, freq = dtm_d$freq, scale=c(5,0.5),
          min.freq = 1, max.words=100, random.order=FALSE,
          rot.per=0.40, colors=brewer.pal(8, "Dark2"))

# Kojarzenia słów
findAssocs(TextDoc_dtm, terms = c("poetry","form","literary","work"), corlimit =
0.5)

# Analiza sentymentu
syuzhet_vector <- get_sentiment(text, method="syuzhet")
summary(syuzhet_vector)

bing_vector <- get_sentiment(text, method="bing")
summary(bing_vector)

afinn_vector <- get_sentiment(text, method="afinn")
summary(afinn_vector)

# Analiza emocji
d <- get_nrc_sentiment(as.vector(dtm_d$word))
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,]

# Wykres liczby słów związanych z każdym uczuciem
ggplot(td_new2, aes(x = sentiment, y = count, fill = sentiment)) +
  geom_bar(stat = "identity") +
  ggtitle("Survey sentiments") +
  ylab("count")

# Bigramy
text_df <- tibble(line = 1, text = text)
tidy_text <- text_df %>%
  unnest_tokens(word, text)

data(stop_words)
tidy_text <- tidy_text %>%
  anti_join(stop_words, by = "word")

# Lista częstotliwości słów
tidy_text %>%
  count(word, sort = TRUE)

# Bigramy
text_bigrams <- text_df %>%
  unnest_tokens(bigram, text, token = "ngrams", n = 2)

text_bigrams %>%
  count(bigram, sort = TRUE)

bigrams_separated <- text_bigrams %>%
  separate(bigram, c("word1", "word2"), sep = " ")

bigrams_filtered <- bigrams_separated %>%
  filter(!word1 %in% stop_words$word) %>%
```

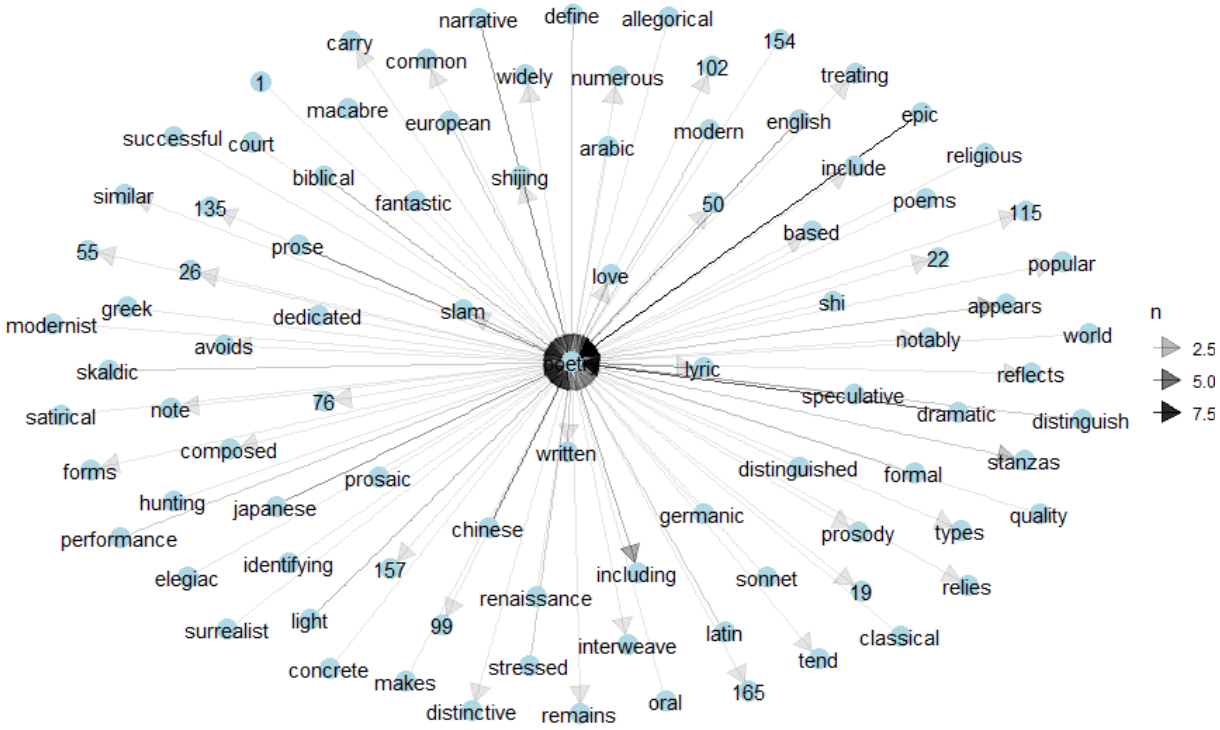
```
  filter(!word2 %in% stop_words$word)

bigram_counts <- bigrams_filtered %>%
  count(word1, word2, sort = TRUE)

bigrams_united <- bigrams_filtered %>%
  unite(bigram, word1, word2, sep = " ")

# Konstruowanie grafów
bigram_graph <- bigram_counts %>%
  filter(word1 == "poetry" | word2 == "poetry") %>%
  graph_from_data_frame()

# Wyświetlanie grafów
ggraph(bigram_graph, layout = "fr") +
  geom_edge_link(aes(edge_alpha = n), show.legend = TRUE, arrow = arrow(type =
"closed", length = unit(.15, "inches")), end_cap = circle(.07, 'inches')) +
  geom_node_point(color = "lightblue", size = 5) +
  geom_node_text(aes(label = name), position = "identity") +
  theme_void()
```

Wyniki z konsol:




poetri
poem
rhyme
form
line
use
poetic
languag
vers
poet
tradit
syllabl
epic
rhythm
meter
pattern
exampl
includ
structur
english
may
often
centuri
genr
sonnet
written
mani
can
word
stanza
common
general
differ
tone
first
couplet
subject
describ
sound
classic
william
narrat
metric
set
distinguish
type
diction
time
much
three
among
less
second
regular
greek
song
creat
also
alliter
write
ghazal
rather
one
formal
particular
work
dactyl
natur
number
quatrain
chines
relat
two
howev
five
stress
style
term
notabl
ode
prose
cultur
mean
base
end
feet
four
scheme
ancient
follow
literari
similar
modern
element


Survey sentiments
count
2.0
1.5
1.0
0.5
0.0
anger
anticipation
disgust
fear
joy
sadness
surprise
trust
sentiment
sentiment
anger
anticipation
disgust
fear
joy
sadness
surprise
trust



3. **Wnioski:**

- Na podstawie wyników uzyskanych z analizy tekstu artykułu na Wikipedii dotyczącego poezji, można wyciągnąć następujące wnioski:

Najczęściej występujące słowa to: **poetry, form, line, use, rhyme**

Wskazuje to na to, że artykuł skupia się na definicji i strukturze poezji, omawiając różne formy, linie, użycia oraz rymy w poezji. Chmura słów wizualizuje te wyniki, pokazując, że słowo "poetry" jest zdecydowanie najczęściej używanym słowem w tekście, co nie jest zaskakujące, biorąc pod uwagę temat artykułu.

Próba znalezienia kojarzeń słów dla poetry, form, literary, work przy korelacji 0.5 nie zwróciła żadnych wyników (numeric(0)). Może to oznaczać, że żadne z tych słów nie są wystarczająco silnie powiązane z innymi słowami w tekście przy ustawionym progu korelacji.

Ogólnie rzecz biorąc, analiza tekstu artykułu na Wikipedii dotyczącego poezji pokazuje, że jest on skoncentrowany na omawianiu definicji, form i struktur poezji. Artykuł jest napisany w neutralnym tonie, co jest zgodne z oczekiwaniami wobec treści encyklopedycznych. Najczęściej używane słowa i bigramy są zgodne z tematem, a analiza emocji i sentymentu potwierdza obiektywność i informacyjny charakter tekstu.

Link do repozytorium: https://github.com/fireinx/apu