КНІТ-33, Возняк Микола, Лабораторна робота 13, Варіант 2.

**Підключення бібліотеки**

> library(tm)

**1. Збереження текстових документів у змінних**

> t1 <- "Our university gives a good education."

> t2 <- "Good education provides a good job in the future."

> t3 <- "Appropriate programmes of the university give students good education."

> t4 <- "Certain programmes of the university gives bad level of mathematical education."

> t5 <- "Bad education leads to bad employment. Do all your homework."

> t6 <- "Bad education prevents success in the future. Be a good student at a university."

**2. Вилучення чисел**

> t1 <- removeNumbers(t1)

> t2 <- removeNumbers(t2)

> t3 <- removeNumbers(t3)

> t4 <- removeNumbers(t4)

> t5 <- removeNumbers(t5)

> t6 <- removeNumbers(t6)

**3. Вилучення розділових знаків**

> t1 <- removePunctuation(t1)

> t2 <- removePunctuation(t2)

> t3 <- removePunctuation(t3)

> t4 <- removePunctuation(t4)

> t5 <- removePunctuation(t5)

> t6 <- removePunctuation(t6)

**4. Вилучення зайвих пробілів**

> t1 <- stripWhitespace(t1)

> t2 <- stripWhitespace(t2)

> t3 <- stripWhitespace(t3)

> t4 <- stripWhitespace(t4)

> t5 <- stripWhitespace(t5)

> t6 <- stripWhitespace(t6)

**5. Об’єднання у вектор**

> t <- c(t1, t2, t3, t4, t5, t6)

> t

[1] "Our university gives a good education"

[2] "Good education provides a good job in the future"

[3] "Appropriate programmes of the university give students good education"

[4] "Certain programmes of the university gives bad level of mathematical education"

[5] "Bad education leads to bad employment Do all your homework"

[6] "Bad education prevents success in the future Be a good student at a university"

**6. Створення корпусу на основі вектора t**

> corpus <- Corpus(VectorSource(t))

> corpus

<<SimpleCorpus>>

Metadata: corpus specific: 1, document level (indexed): 0

Content: documents: 6

**7. Обчислення матриці частот для всіх термінів та всіх документів.**

> dtm <- DocumentTermMatrix(corpus)

> dtm

<<DocumentTermMatrix (documents: 6, terms: 25)>>

Non-/sparse entries: 44/106

Sparsity : 71%

Maximal term length: 12

Weighting : term frequency (tf)

**8. Вивід термінів, які зустрічаються не менше, ніж 2, 3 та 4 рази.**

> findFreqTerms(dtm, lowfreq = 2)

[1] "education" "gives" "good" "university" "future"

[6] "the" "programmes" "bad"

> findFreqTerms(dtm, lowfreq = 3)

[1] "education" "good" "university" "the" "bad"

> findFreqTerms(dtm, lowfreq = 4)

[1] "education" "good" "university" "the" "bad"

**9. Обчислення частоти для термінів із списку: university, programmes, education, students.**

> selected\_terms <- c("university", "programmes", "education", "students")

> freq\_selected\_terms <- colSums(as.matrix(dtm[, selected\_terms]))

> freq\_selected\_terms

university programmes education students

4 2 6 1

**10. Терміни з найбільшою частотою**

> top\_terms

$`1`

education gives good our university

1 1 1 1 1

$`2`

good education future job provides

2 1 1 1 1

$`3`

appropriate education give good programmes

1 1 1 1 1

$`4`

bad certain education gives level

1 1 1 1 1

$`5`

bad all education employment homework

2 1 1 1 1

$`6`

bad education future good prevents

1 1 1 1 1