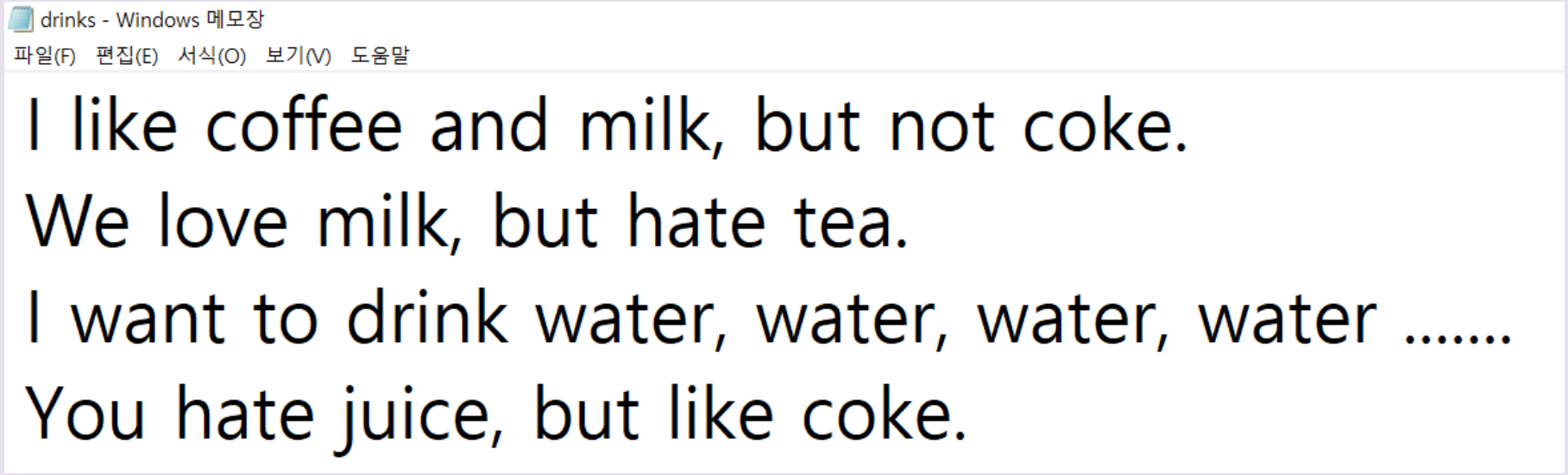
R
영어 텍스트 분석

2017010698
수학과 오서영

1 영어 텍스트 분석 tm 패키지 사용


drinks - Windows 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말

I like coffee and milk, but not coke.

We love milk, but hate tea.

I want to drink water, water, water, water .....

You hate juice, but like coke.

1

영어 텍스트 분석
tm 패키지 사용

```
# Set work directory
setwd("my_work_directory")
```

```
# Install Packages
install.packages("wordcloud")
install.packages("tm")
library("wordcloud")
library("tm")
```

1

영어 텍스트 분석
tm 패키지 사용

```
# Load and Explore datasets
mydata1 <- readLines("drinks.txt")
mydata1
```

```
> data1 <- readLines("drinks.txt")
> data1
[1] "I like coffee and milk, but not coke."
[2] "We love milk, but hate tea."
[3] "I want to drink water, water, water, water ......."
[4] "You hate juice, but like coke."
```

1 영어 텍스트 분석 tm 패키지 사용

```
# Convert 'data1' to 'Corpus' form
corp1 <- VCorpus(VectorSource(mydata1))
corp1
```

```
> corp1 <- VCorpus(VectorSource(data1))
> corp1
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 4
```

Document : tm패키지가 작업할 수 있는 특별한 형태
-> 1줄 : 1 document

1 영어 텍스트 분석 tm 패키지 사용

```
# Explore corpus
inspect(corp1)
```

```
> inspect(corp1)
<<VCorpus>>
Metadata:  corpus specific: 0,
 document level (indexed): 0
Content:  documents: 4

[[1]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 37

[[2]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 27
```

```
[[3]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 50

[[4]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 30
```

1번째 줄에
37글자

1

영어 텍스트 분석
tm 패키지 사용

```
# Convert 'Corpus' form to 'Term-Document' form
tdm <- TermDocumentMatrix(corp1)
tdm
```

```
> tdm <- TermDocumentMatrix(corp1)
> tdm
<<TermDocumentMatrix (terms: 17, documents: 4)>>
Non-/sparse entries: 23/45
Sparsity           : 66%
Maximal term length: 7
Weighting          : term frequency (tf)
```

Terms : 17 '17개의 단어'
Documents : 4 '4개의 문장'
Sparsity : 66% '66%의 공백'

1

영어 텍스
tm 패키

```
# Convert 'tdm'
m <- as.matrix(t
m
```

```
> m <- as.matrix(tdm)
> m
          Docs
Terms      1 2 3 4
  ....... 0 0 1 0
  and     1 0 0 0
  but     1 1 0 1
  coffee  1 0 0 0
  coke.   1 0 0 1
  drink   0 0 1 0
  hate    0 1 0 1
  juice,  0 0 0 1
  like    1 0 0 1
  love    0 1 0 0
  milk,   1 1 0 0
  not     1 0 0 0
  tea.    0 1 0 0
  want    0 0 1 0
  water   0 0 1 0
  water,  0 0 3 0
  you     0 0 0 1
```

1 영어 텍스트 분석 tm 패키지 사용

```
# Eliminate Stopword, Punctuation and Blank
corp2 <- tm_map(corp1, stripWhitespace)
corp2 <- tm_map(corp2, tolower)
corp2 <- tm_map(corp2, removePunctuation)
corp2 <- tm_map(corp2, PlainTextDocument)
# add stopword(and, but, not)
stopword1 <- c(stopwords('en'), "and", "but", "not")
corp2 <- tm_map(corp2, removeWords, stopword1)

# re-generate Term-Document_matrix
tdm2 <- TermDocumentMatrix(corp2)
```

불용어 (Stopword)
: 자주 등장하지만 분석에 도움이 안 되는 단어

1

영어 텍스트 분석
tm 패키지 사용

```
# re-generate Term                                    onvert to matrix
tdm2 <- TermDocume
m2 <- as.matrix(td
colnames(m2) <- c(
m2
```

```
       Docs
Terms   1 2 3 4
  coffee 1 0 0 0
  coke   1 0 0 1
  drink  0 0 1 0
  hate   0 1 0 1
  juice  0 0 0 1
  like   1 0 0 1
  love   0 1 0 0
  milk   1 1 0 0
  tea    0 1 0 0
  want   0 0 1 0
  water  0 0 4 0
```

1

영어 텍스트 분석
tm 패키지 사용

```
# Eliminate more stopwords
stopword2 <- c(stopwords('en'),"drink","hate","like","love","want")
corp3 <- tm_map(corp2, removeWords, stopword2)
tdm3 <- TermDocumentMatrix(corp3)
m3 <- as.matrix(tdm3)
colnames(m3) <- c(1:4)
m3
```

```
        Docs
Terms    1 2 3 4
  coffee 1 0 0 0
  coke   1 0 0 1
  juice  0 0 0 1
  milk   1 1 0 0
  tea    0 1 0 0
  water  0 0 4 0
```

1

영어 텍스트 분석
tm 패키지 사용

```
# Count words
count1 <- sort(rowSums(m3), decreasing = T)
head(count1, 10)
count2 <- sort(colSums(m3), decreasing = T)
head(count2, 10)
```

```
water  coke  milk coffee juice   tea
    4     2     2      1     1     1
```

```
3 1 2 4
4 3 2 2
```

1

영어 텍스트 분석
tm 패키지 사용

```
# Find Correlation
findAssocs(tdm3, "coffee", 0.5)
findAssocs(tdm3, "coffee", 0.6)
```

```
$coffee
coke milk
0.58 0.58
```

```
$coffee
numeric(0)
```

없음

1

영어 텍스트 분석
tm 패키지 사용

```
# Expression with WordCloud
library(RColorBrewer)
pic <- brewer.pal(7, "Set3")
wordcloud(names(count1), freq = count1,
          scale=c(5,1), min.freq=1, rot.per=0.5,
          color=pic, random.order=F, random.color=T)
```

끝