개요

PDF 문서에서 자주 사용되는 단어(키워드)를 분류한 Text mining 결과를 시각적으로 표현하는 것을 word cloud라 한다.

예제 PDF(txt) 문서

(1) 영어 NIV 성경 [http://wolfpack.hnu.ac.kr/Big_Data/data/NIV_Bible.pdf]

```
[Genesis 1]
1.In the beginning God created the heavens and the earth.
2.Now the earth was formless and empty, darkness was over the
surface of the deep, and the Spirit of God was hovering over the
waters.
3.And God said, "Let there be light," and there was light.
4.God saw that the light was good, and he separated the light from
the darkness.
5.God called the light "day," and the darkness he called "night."
```

(2) 영어 성경 txt 버전 [http://wolfpack.hnu.ac.kr/Big_Data/data/NIV성경.txt]

```
[Genesis 1]
1.In the beginning God created the heavens and the earth.
2.Now the earth was formless and empty, darkness was over the surface of the deep, and the Spirit of God
3.And God said, "Let there be light," and there was light.
4.God saw that the light was good, and he separated the light from the darkness.
5.God called the light "day," and the darkness he called "night." And there was evening, and there was mo
6.And God said, "Let there be an expanse between the waters to separate water from water."
7.So God made the expanse and separated the water under the expanse from the water above it. And it was s
8.God called the expanse "sky." And there was evening, and there was morning--the second day.
9.And God said, "Let the water under the sky be gathered to one place, and let dry ground appear." And it
10.God called the dry ground "land," and the gathered waters he called "seas." And God saw that it was go
11.Then God said, "Let the land produce vegetation: seed-bearing plants and trees on the land that bear f
various kinds." And it was so.
12.The land produced vegetation: plants bearing seed according to their kinds and trees bearing fruit wit
God saw that it was good.
13.And there was evening, and there was morning--the third day.
```

PDF 데이터 텍스트 데이터로 읽기 : pdf_text() 함수

```
#install.packages("pdftools")
#Reading PDF File into Text data

library(pdftools)
#PDF 문서 txt 데이터 변환
txt <- pdf_text('http://wolfpack.hnu.ac.kr/Big_Data/data/NIV_Bible.pdf')
head(txt,1) #읽은 첫문자 콘솔 출력
# Read text data from text file.
txt <- readLines("http://wolfpack.hnu.ac.kr/Big_Data/data/NIV성경.txt")
head(txt,10)
```

도움 필요 : pdf_text() 함수는 2단 문서는 오른쪽 문단만 가져오는 문제 있음 - 해결방법

PDF 문서는 페이지 단위로 읽어 들여 한 라인에 저장된다.

```
> head(txt,1) #읽은 첫문자 콘솔 출력
[1] "[Genesis 1]\n1.In the beginning God created the heavens and the earth.\n2.Now the earth w
as formless and empty, darkness was over the\nsurface of the deep, and the Spirit of God was h
overing over the\nwaters.\n3.And God said, \"Let there be light,\" and there was light.\n4.God
saw that the light was good, and he separated the light from\nthe darkness.\n5.God called the
light \"day,\" and the darkness he called \"night.\"\nAnd there was evening, and there was mor
ning--the first day.\n6.And God said, \"Let there be an expanse between the waters to\nseparat
```

txt 문서는 라인(행 단위로 읽어 들여 저장한다.

```
> head(txt,10)
 [1] "[Genesis 1]"
 [2] "1.In the beginning God created the heavens and the earth."
 [3] "2.Now the earth was formless and empty, darkness was over the surface of the dee
p, and the Spirit of God was hovering over the waters."
 [4] "3.And God said, \"Let there be light,\" and there was light."
 [5] "4.God saw that the light was good, and he separated the light from the darkness.
"
 [6] "5.God called the light \"day,\" and the darkness he called \"night.\" And there
was evening, and there was morning--the first day."
 [7] "6.And God said, \"Let there be an expanse between the waters to separate water f
rom water.\""
 [8] "7.So God made the expanse and separated the water under the expanse from the wat
```

텍스트 문장 자연어 처리 : Corpus() 함수

본 예제 데이터는 PDF 문서를 기본으로 하여 적용한 결과임

```
#Natural Language Process
library(NLP)
library(tm)
docs <- Corpus(VectorSource(txt)) #말뭉치 만들기 - VectorSource 문자벡터 말뭉치
inspect(docs[2]) #말뭉치 작성 확인 - 페이지 지정
```

- inspect() 함수 : 만들어진 말뭉치(docs 오브젝트) 2 페이지 결과 보기

```
ground according to their kinds. And God saw that it was good.
26.Then God said, "Let us make man in our image, in our likeness,
and let them rule over the fish of the sea and the birds of the air,
over the livestock, over all the earth, and over all the creatures
that move along the ground."
27.So God created man in his own image, in the image of God he
created him; male and female he created them.
28.God blessed them and said to them, "Be fruitful and increase in
number; fill the earth and subdue it. Rule over the fish of the sea
and the birds of the air and over every living creature that moves
on the ground."
29.Then God said, "I give you every seed-bearing plant on the face
```

```
> inspect(docs[2]) #말뭉치 작성 확인 - 페이지 지정
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 1

[1] ground according to their kinds. And God saw that it was good.\n26.Then God said, "
Let us make man in our image, in our likeness,\nand let them rule over the fish of the
sea and the birds of the air,\nover the livestock, over all the earth, and over all the
creatures\nthat move along the ground."\n27.So God created man in his own image, in the
image of God he\ncreated him; male and female he created them.\n28.God blessed them and
said to them, "Be fruitful and increase in\nnumber; fill the earth and subdue it. Rule
```

특수문자 빈칸(space) 변환 : content_transformer() 함수

```
#tm_map() 함수 활용하여 텍스트의 특수문자를 제외하기 - content_transformer() 함수
#특수문자 “/”, “@”, “|”,"\n" 을 space(공백)으로 변환

toSpace <- content_transformer(function (x , pattern ) gsub(pattern,' ', x))
docs <- tm_map(docs, toSpace, '/')
docs <- tm_map(docs, toSpace, '@')
docs <- tm_map(docs, toSpace, '\\|')
docs <- tm_map(docs, toSpace, '\n')
inspect(docs[2]) #특수문자 제외 확인
```

- toSapce 오브젝트는 content_transformer() 함수를 이용하여 적용되는 특수문자를 ‘ ’ (빈칸, 반드시 따옴표 안은 빈칸이 있도록, 그렇지 않으면 특수문자가 빈칸으로 바뀌는 것이 아니라 없어지는 효과만)
- 불러온 문서에 따라 특수문자가 상이할 수 있으므로 자연어 처리된 문장을 살펴 제외한다.

```
> inspect(docs[2]) #말뭉치 작성 확인 - 페이지 지정
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 1

[1] ground according to their kinds. And God saw that it was good. 26.Then God said, "L
et us make man in our image, in our likeness, and let them rule over the fish of the se
a and the birds of the air, over the livestock, over all the earth, and over all the cr
eatures that move along the ground." 27.So God created man in his own image, in the ima
```

워드 클라우드 불필요 단어 제외 : tm_map() 함수

```
# 텍스트 대문자를 소문자로 변환
docs <- tm_map(docs, content_transformer(tolower))
# 텍스트 숫자를 제거
docs <- tm_map(docs, removeNumbers)
# 텍스트 english 제거
docs <- tm_map(docs, removeWords, stopwords("english"))
# 텍스트 물음표 제거
docs <- tm_map(docs, removePunctuation)
# 텍스트 빈칸(white space) 제거
docs <- tm_map(docs, stripWhitespace)
# Text stemming : 동사 기본형으로 표준화 : do, done, doing -> do로 표준화
docs <- tm_map(docs, stemDocument)

# 사용자 지정 제거 단어 설정 및 문서(텍스트)에서 제거
my_custom_stopwords <- c('will', 'come','one', 'said','say','went',
    'may', 'let', 'give', 'made', 'make','came', 'hous','hand','now',

     'put','also','call','saw','done','eat','gave','can','left','know',
     'mani','ask','set','even','everi','sent','back','look','took','take')
docs <- tm_map(docs, removeWords, my_custom_stopwords)
inspect(docs[2]) #불필요 단어 제외 확인
```

```
> inspect(docs[2]) #말뭉치 작성 확인 - 페이지 지정
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 1

[1] ground accord kind god  good god   us  man imag like  rule fish sea bird air livest
ock earth creatur move along ground god creat man imag imag god creat male femal creat
god bless  fruit increas number fill earth subdu rule fish sea bird air  live creatur m
ove ground god    seedbear plant face whole earth  tree fruit seed  food beast earth bi
rd air creatur move groundeveryth breath life   green plant food god   good  morn sixth
```

문장 단어 카운트 및 결과 보기

```
#문서번호와 단어간의 사용여부 또는 빈도 카운트
dtm <- TermDocumentMatrix(docs)
#생성된 document term matrix에서 1-10번 빈도 높은단어의
#100번째에서 150번째 문서의 분포를 확인
#Sparsity           : 공백 수
inspect(dtm[1:10,100:150])

findFreqTerms(dtm, lowfreq = 500) #500회 이상 사용된 단어 출력
findFreqTerms(dtm, 500,600) #10에서 15회 사이로 사용된 단어 출력
```

```
> inspect(dtm[1:10,100:150])
<<TermDocumentMatrix (terms: 10, documents: 51)>>
Non-/sparse entries: 57/453
Sparsity           : 89%
Maximal term length: 6
Weighting          : term frequency (tf)
Sample             :
        Docs
Terms    100 112 129 135 136 137 139 141 149 150
  accord   1   1   0   0   0   0   0   0   0   1
  across   0   0   0   0   0   0   0   0   0   0
  along    2   1   1   1   0   0   1   0   0   1
```

```
> findFreqTerms(dtm, lowfreq = 500)
 [1] "day"       "earth"     "god"       "good"      "great"     "heaven"
 [7] "land"      "live"      "place"     "spirit"    "thing"     "two"
[13] "water"     "year"      "eye"       "holi"      "life"      "like"
[19] "lord"      "man"       "name"      "work"      "among"     "answer"
```

```
> findFreqTerms(dtm, 500,600)
 [1] "spirit"    "eye"       "life"      "work"      "answer"    "heard"
 [7] "head"      "destroy"   "righteous" "law"       "christ"
```

관심 단어와 동일한 문장에서 함께 사용 빈도 높은 단어 : 연관 관계 findAssocs() 함수

- 문장이 마침표(period .)가 있어야 가능한 분석임 – 요한 – 침례, 헤로디아, platter(접시), 예수

```
#findAssocs - 설정 단어(work)와 연관성이 0.3(상관계수) 이상 높은 단어
findAssocs(dtm, terms = "sin", corlimit = 0.3)
findAssocs(dtm, terms = "john", corlimit = 0.3)
```

```
> findAssocs(dtm, terms = "sin", corlimit = 0.3)
$sin
   offer    natur   commit forgiven  sonship
    0.32     0.32     0.31     0.31     0.30

> findAssocs(dtm, terms = "john", corlimit = 0.3)
$john
 baptist  herodia   baptiz  platter    jesus    herod   behead birthday     amaz
    0.53     0.48     0.44     0.42     0.40     0.40     0.38     0.36     0.32
 discipl
    0.32
```

단어 빈도 계산

```
#Build a term-document matrix
#Document Term matrix 만들기 : 단어 빈도표(word.freq) 포함.
#The function TermDocumentMatrix() 결과 기반
dtm.mat <- as.matrix(dtm)
v <- sort(rowSums(dtm.mat),decreasing=TRUE)
word.freq <- data.frame(word = names(v),freq=v)

#상위 빈도 20개 빈도표 출력
head(word.freq, 20)
```

```
> head(word.freq, 20)
          word freq
lord      lord 7809
god        god 4451
son        son 3227
king      king 2881
peopl    peopl 2375
```

Lord 빈도가 가장 높음 > God > Son 순임

워드 cloud 표현

```
#word cloud 라이브러리
library(SnowballC) ; library(RColorBrewer)
library(wordcloud)
#Generate the Word cloud
wordcloud(words = word.freq$word, freq = word.freq$freq, min.freq = 300,
    random.order=F, rot.per=0.35, colors=brewer.pal(8, 'Dark2')) #빈도 300 이상 단어

#Frequency Table
barplot(word.freq[1:20,]$freq, las = 2, names.arg = word.freq[1:20,]$word,
        col ="lightblue", main ="Most frequent words",
        ylab = "Word frequencies") #상위 빈도 20 개 단어
```