한글 텍스트 추출 및 코퍼스 생성

김진석

2018-03-04

필요한 R 패키지

```
library(tm)
library(KoNLP)
```

- KoNLP: 형태소 분석기인 한나눔(http://semanticweb.kaist.ac.kr/home/index.php/HanNanum) Java프로그램을 전희원이 R패키지로 재개발

예제 텍스트 문서: 반야바라밀다심경 한글판

```
docs <- readLines("반야심경.txt")
head(docs)
```

```
## [1] "관자재보살이 깊은 반야바라밀다를 행할 때, 오온이 공한 것을 비추어 보고 온갖 고통에서 건너
## [2] "사리자여! 색이 공과 다르지 않고, 공이 색과 다르지 않으며, 색이 곧 공이고 공이 곧 색이니,
## [3] "사리자여! 모든 존재는 텅 빈 것이므로, 생겨나지도 없어지지도 않으며, 더럽지도 깨끗하지도
## [4] "그러므로 공의 관점에서는 실체가 없고 감각, 생각, 행동, 의식도 없으며,"
## [5] "눈도, 귀도, 코도, 혀도, 몸도, 의식도 없고, 색깔도, 소리도, 향기도, 맛도, 감촉도, 법도 없
## [6] "무명도 무명이 다함까지도 없으며, 늙고 죽음도 늙고 죽음이 다함까지도 없고,"
```

전처리(Pre-processing)

```
# 행 라벨 붙이기
names(docs) <- paste("doc", 1:length(docs), sep="")
```

```
# 문장부호 제거
docs <- gsub("[[:punct:]]", " ", docs)
# 숫자 제거
docs <- gsub("[[:digit:]]", "", docs)
```

사전 추가

```
useSejongDic()
```

```
## Backup was just finished!
## 370957 words dictionary was built.
```

```
new_term <- c("관자재보살", "사리자", "반야바라밀다")
new_dic <- data.frame(new_term , "ncn")
mergeUserDic(new_dic)
```

```
## 3 words were added to dic_user.txt.
```

형태소 분석(POS: Part of Speech; Morph analysis)

- 한글의 형태소(KAIST 품사 태그)
  - MorphAnalyzer: 세분류(영문소문자 3자리 표현)
    - ncn: 비서술형 보통명사, ncp: 서술형 보통명사, ···
  - SimplePos09: 9개 대분류
    - N:체언, P:용언, M:수식언, I:독립언(감탄사등), J: 관계언, ···
  - SimplePos22: 중분류 22개
    - NC:보통명사,NQ: 고유명사,..
    - PV:동사, PA: 형용사
    - MM:관형사, MA:부사,···
    - II:감탄사

형태소 분석(POS: Part of Speech; Morph analysis)

```
x <- unlist(SimplePos09(docs[1]))
i <- grep("/P¦/M¦/I", x)
x
```

```
##          관자재보살이                  깊은          반야바라밀다를
##   "관자재보살/N+이/J"    "깊/P+은/E" "반야바라밀다/N+를/J"
##                  행할                   때                  오온이
##        "행하/P+ㄹ/E"         "때/N"        "오온/N+이/J"
##                  공한                  것을                  비추어
##             "공한/N"    "것/N+을/J"        "비추/P+어/E"
##                  보고                  온갖                  고통에서
##           "보/P+고/E"       "온갖/M"      "고통/N+에서/J"
##                건너느니                     라
##       "건너/P+느니/E"          "라/N"
```

```
x[i]
```

```
##          깊은          행할          비추어          보고          온갖
##  "깊/P+은/E" "행하/P+ㄹ/E" "비추/P+어/E" "보/P+고/E"     "온갖/M"
##        건너느니
## "건너/P+느니/E"
```

형태소 분석(POS: Part of Speech; Morph analysis)

- 명사 추출

```
# Extract Nouns
extractNoun(docs[1])
```

```
## [1] "관자재보살"   "반야바라밀다" "때"           "오온"         "공한"
## [6] "것"           "고통"         "라"
```

```
docs1 <- sapply(docs, function(x){
  paste(extractNoun(x),  collapse=" ")
})
```

코퍼스 생성 및 분석

```
docs.corp <- Corpus(VectorSource(docs1))
# Document-Term Matrix
tdm <- TermDocumentMatrix(docs.corp, control=list(tokenize="scan",
                #weighting = weightTfIdf,
                weighting = weightTf,
                wordLengths=c(1,Inf)))
wordFreq <- slam::row_sums(tdm)
sort(wordFreq, decreasing=TRUE)[1:5]
```

```
##           라 반야바라밀다          색         의식       깨달음
##            7           5          4          4          4
```