경주시 민원자료 분석

김진석

2016년 9월 27일

경주시 민원자료

- 메타데이터 읽기

```
  library(XLConnect)
# Home <- "H:/lectures/2016-2/bigdata/민원데이터/경주시/"
# metaFile <- paste0(Home, "meta/meta-data-경주시.xlsx")
  meta <- readWorksheetFromFile(metaFile, sheet=1)
  head(meta)
```

경주시 민원자료

- 민원 데이터 읽기

```
fileNamesDir <- dir(paste0(Home, "/data"))
fileNames <- paste0(Home, "data/",meta$file_id, ".txt")
id <- match(gsub(".txt", "", fileNamesDir), as.character(meta$file_id))
fileNamesDir[which(is.na(id))]

texts <- sapply(fileNames, function(x){
        doc <- readLines(x, n=-1, warn=FALSE)
        doc <- paste(doc, collapse=" ")
      }
    )
texts <- gsub("[[:punct:]]", "", texts)
```

경주시 민원자료

- 단어 추출

```
##
library(KoNLP)
useSejongDic()
x1 <- extractNoun(texts[1]) #x2 <- lapply(texts[1:2], extractNoun)
MorphAnalyzer(texts[1])
```

경주시 민원자료

- 문서-단어 행렬(document-term matrix)

```
m <- simple_triplet_matrix(i = i, j = j, v = rep(1, length(j)),
                           nrow = length(x),
                           ncol = length(u.keywords),
                           dimnames =
                             list(doc = as.character(seq_along(x)),
                              terms=u.keywords))
class(m) <- c("DocumentTermMatrix", "simple_triplet_matrix")
weighting <- control$weighting
dtm <- weighting(m)
```

경주시 민원자료

- n-gram

```
library("RWeka")
library("tm")
BigramTokenizer <- function(x) NGramTokenizer(x, Weka_control(min = 2, max =
tdm <- TermDocumentMatrix(crude, control = list(tokenize = BigramTokenizer))
#
NGramTokenizer(paste(x1[[1]], collapse=" "), Weka_control(min = 1, max = 2))
```