Text Mining using R

Veerasak Kritsanapraphan

Get Text Mining Library

- Needed <- c("tm", "SnowballCC", "RColorBrewer", "ggplot2", "wordcloud", "biclust", "cluster", "igraph", "fpc")

- install.packages(Needed, dependencies=TRUE)

Load File

- http://www.cs.cornell.edu/people/pabo/movie-review-data/review_polarity.tar.gz

- # Load Text file

- cname <- file.path("~", "Downloads", "text")

- cname

- dir(cname)

Text Mining

- library(tm)

- docs <- Corpus(DirSource(cname))

- summary(docs)

Text Mining Basic

- # Remove Punctuation

- docs <- tm_map(docs, removePunctuation)

- for(j in seq(docs))

- {

-     docs[[j]] <- gsub("/", " ", docs[[j]])

-     docs[[j]] <- gsub("@", " ", docs[[j]])

-     docs[[j]] <- gsub("\\|", " ", docs[[j]])

- }

- #remove Number

- docs <- tm_map(docs, removeNumbers)

- docs <- tm_map(docs, tolower)

- # remove stop words

- docs <- tm_map(docs, removeWords, stopwords("english"))

Clean Data

- # remove ing s, es

- library(SnowballC)

- docs <- tm_map(docs, stemDocument)

- docs <- tm_map(docs, stripWhitespace)

- # tells R to treat your preprocessed documents as text documents.

- docs <- tm_map(docs, PlainTextDocument)

Step of Text Mining

- # Create Document Term Matrix

- dtm <- DocumentTermMatrix(docs)

- dtm

- # Create Term Document Matrix

- tdm <- TermDocumentMatrix(docs)

- tdm

Explore Data

- # Explore Data

- freq <- colSums(as.matrix(dtm))

- length(freq)

- ord <- order(freq)

- # Start by removing sparse terms:

- dtms <- removeSparseTerms(dtm, 0.1) # This makes a matrix that is 10% empty space, maximum.

- inspect(dtms)

- freq[head(ord)]

- freq[tail(ord)]

- head(table(freq), 20)

Explore Data

- \# we can view a table of the terms we selected when we removed sparse terms

- freq <- colSums(as.matrix(dtms))

- freq <- sort(colSums(as.matrix(dtm)), decreasing=TRUE)

- head(freq, 14)

- findFreqTerms(dtm, lowfreq=500)

- wf <- data.frame(word=names(freq), freq=freq)

- head(wf)

Create Word Cloud

- library(ggplot2)

- p <- ggplot(subset(wf, freq>500), aes(word, freq))

- p <- p + geom_bar(stat="identity")

- p <- p + theme(axis.text.x=element_text(angle=45, hjust=1))

- p

Thank you