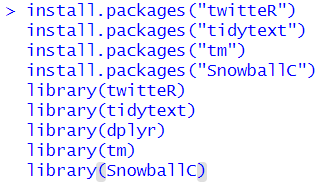
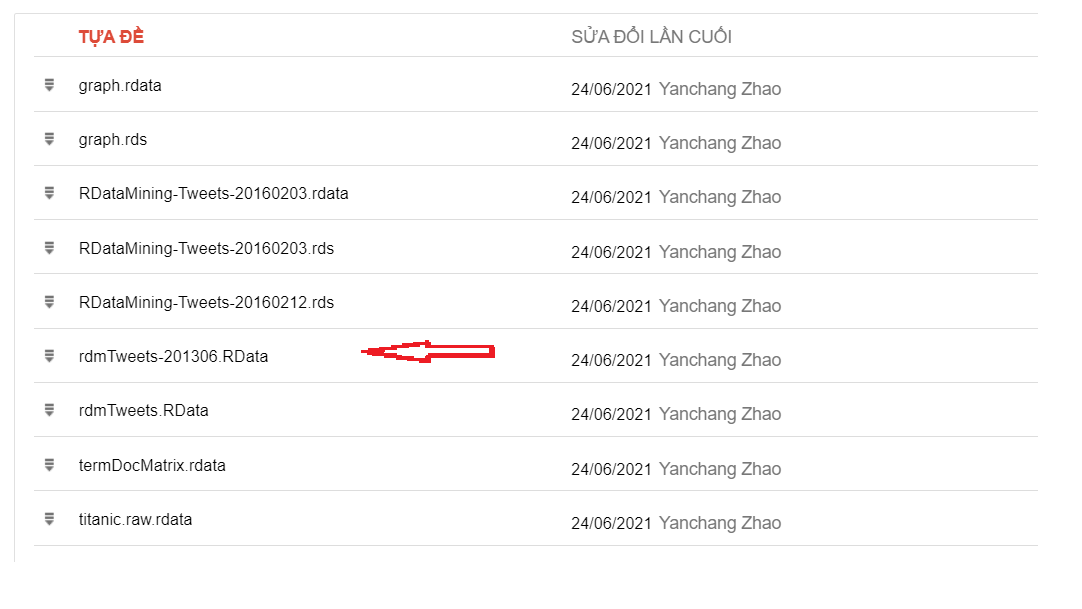
Cài đặt + import thư viện



đặt giá trị của biến workingDir thành đường dẫn tới thư mục C:/a/demo



Tải file dataset



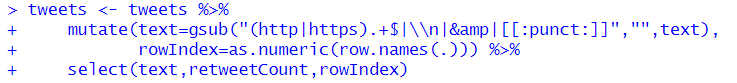
Load file vừa tải vào Rstudio Chuyển file trên thành df và gàn vào biến tweets



Chuyển đổi dữ liệu trên dataframe tweets trong R. Loại bỏ các URL, dấu xuống dòng, "&", và dấu câu từ cột text.

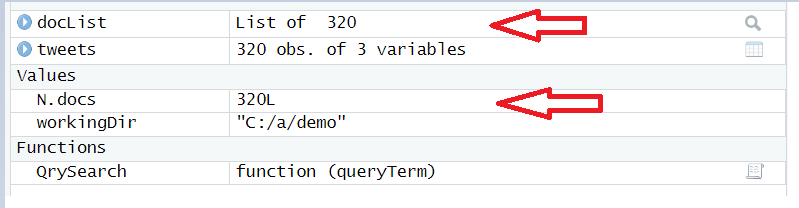
Tạo một cột mới có tên rowIndex để lưu trữ chỉ số dòng của mỗi dòng trong dataframe.

Chọn chỉ ba cột từ dataframe, bao gồm text, retweetCount, và rowIndex, và loại bỏ các cột khác



Tạo danh sách mới và tính số lượng phần tử của nó





Tạo hàm qry để tìm kiếm và đánh giá kquả tìm kiếm

> QrySearch <- function(queryTerm) {

+ start.time <- Sys.time()

+ my.docs <- VectorSource(c(docList, queryTerm))

+ my.corpus <- VCorpus(my.docs) %>%

+ tm\_map(stemDocument) %>%

+ tm\_map(removeNumbers) %>%

+ tm\_map(content\_transformer(tolower)) %>%

+ tm\_map(removeWords,stopwords("en")) %>%

+ tm\_map(stripWhitespace)

+ term.doc.matrix.stm <- TermDocumentMatrix(my.corpus,

+ control=list(

+ weighting=function(x)

+ weightSMART(x,spec="ltc"),

+ wordLengths=c(1,Inf)))

+ term.doc.matrix <- tidy(term.doc.matrix.stm) %>%

+ group\_by(document) %>%

+ mutate(vtrLen=sqrt(sum(count^2))) %>%

+ mutate(count=count/vtrLen) %>%

+ ungroup() %>%

+ select(term:count)

+ docMatrix <- term.doc.matrix %>%

+ mutate(document=as.numeric(document)) %>%

+ filter(document<N.docs+1)

+ qryMatrix <- term.doc.matrix %>%

+ mutate(document=as.numeric(document)) %>%

+ filter(document>=N.docs+1)

+ searchRes <- docMatrix %>%

+ inner\_join(qryMatrix,by=c("term"="term"),

+ suffix=c(".doc",".query")) %>%

+ mutate(termScore=round(count.doc\*count.query,4)) %>%

+ group\_by(document.query,document.doc) %>%

+ summarise(Score=sum(termScore)) %>%

+ filter(row\_number(desc(Score))<=10) %>%

+ arrange(desc(Score)) %>%

+ left\_join(tweets,by=c("document.doc"="rowIndex")) %>%

+ ungroup() %>%

+ rename(Result=text) %>%

+ select(Result,Score,retweetCount) %>%

+ data.frame()

+ end.time <- Sys.time()

+ time.taken <- round(end.time - start.time,4)

+ print(paste("Used",time.taken,"seconds"))

+ return(searchRes)

+ }

Truyền vào tham số > QrySearch("data science") ta được kết quả

