HD.R

aqeel

Mon Apr 25 03:26:44 2016

```
rm(list=ls())
setwd('/home/aqeel/Study/DM/CourseProject/')
# Word cloud code adopted from http://www.r-bloggers.com/word-clouds-us
ing-text-mining/
#install.packages("tm")
library(readr)
library(sqldf)
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Loading required package: RSQLite
```

```
## Loading required package: DBI
```

```
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```
library(tm)
```

```
## Loading required package: NLP
```

```
#cat("Reading data\n")
train <- read_csv('Search Relatives/df_train_stemmed.csv')
test <- read_csv('Search Relatives/df_test_stemmed.csv')
desc <- read_csv('Search Relatives/df_pro_desc_stemmed.csv')
attr <- read_csv('Search Relatives/df_attr_stemmed.csv')
n_words <- function(keyword){
  return(length(unlist(strsplit(keyword,"\\S+"))))
}

get_word_frequency <- function(vec){
  trainCorpus <- Corpus(VectorSource(vec))
  trainCorpus = tm_map(trainCorpus, content_transformer(tolower))
  trainCorpus = tm_map(trainCorpus, removePunctuation)
  trainCorpus = tm_map(trainCorpus, removeWords, stopwords("english"))
  dtm_matrix = TermDocumentMatrix(trainCorpus, control = list(minWordLe
ngth = 1))
  m = as.matrix(dtm_matrix)
  v = sort(rowSums(m), decreasing = TRUE)
  return(v)
}


## Exploring search terms
st_train <- unique(train$search_term)
st_test <- unique(test$search_term)
diff_terms_1 <- setdiff(st_train,st_test)
diff_terms_2 <- setdiff(st_test,st_train)
common_terms <- intersect(st_train,st_test)

(paste("Number of search terms in train :",(length(st_train)),sep=" "))
```

```
## [1] "Number of search terms in train : 11748"
```

```
(paste("Number of search terms in test :",(length(st_test)),sep=" "))
```

```
## [1] "Number of search terms in test : 22210"
```

```
print(paste("Number of terms in train not in test :",(length(diff_terms
_1)),sep=" "))
```

```
## [1] "Number of terms in train not in test : 2520"
```

```
print(paste("Number of terms in test not in train :",(length(diff_terms
_2)),sep=" "))
```

```
## [1] "Number of terms in test not in train : 12982"
```

```
print(paste("Number of common terms in test and train :",(length(common
_terms)),sep=" "))
```

```
## [1] "Number of common terms in test and train : 9228"
```

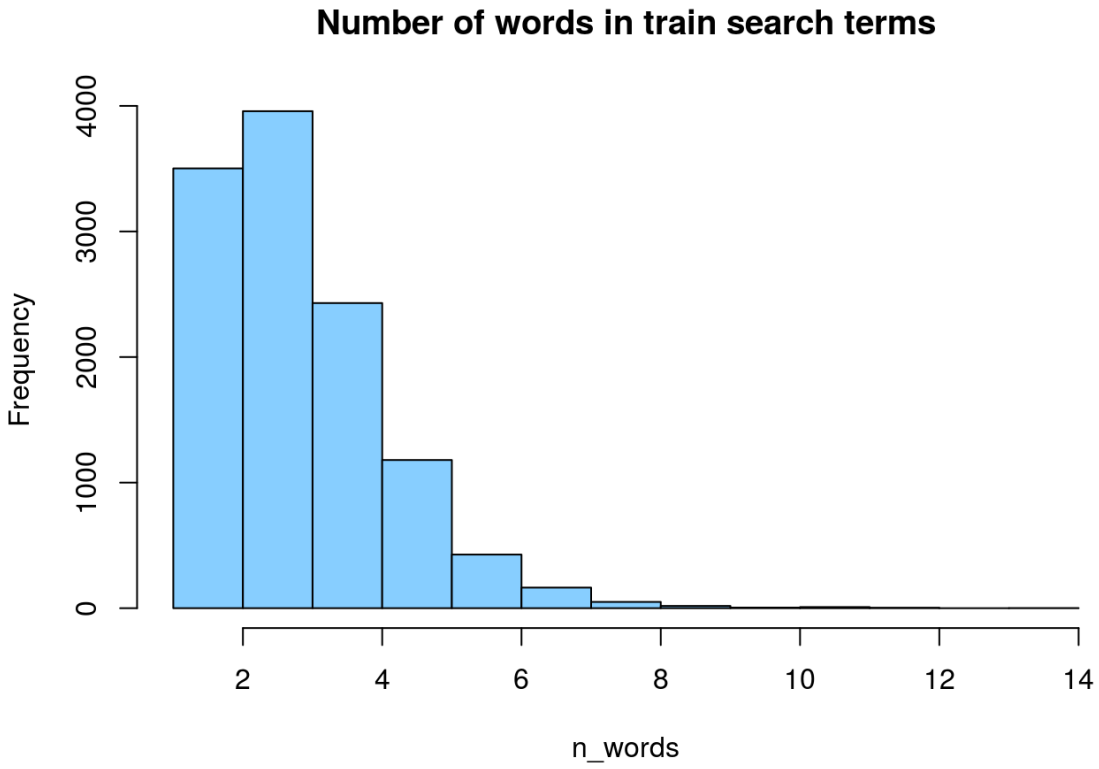
```
train_term_words <- sapply(st_train,n_words)
cat('Number of words in train terms\n')
```

```
## Number of words in train terms
```

```
print(table(train_term_words))
```

```
## train_term_words
##    1    2    3    4    5    6    7    8    9   10   11   12   14
##  641 2861 3958 2430 1180  427  164   50   18    5    9    4    1
```

```
hist(train_term_words,main="Number of words in train search terms", xla
b="n_words",ylab="Frequency",col="skyblue1")
```



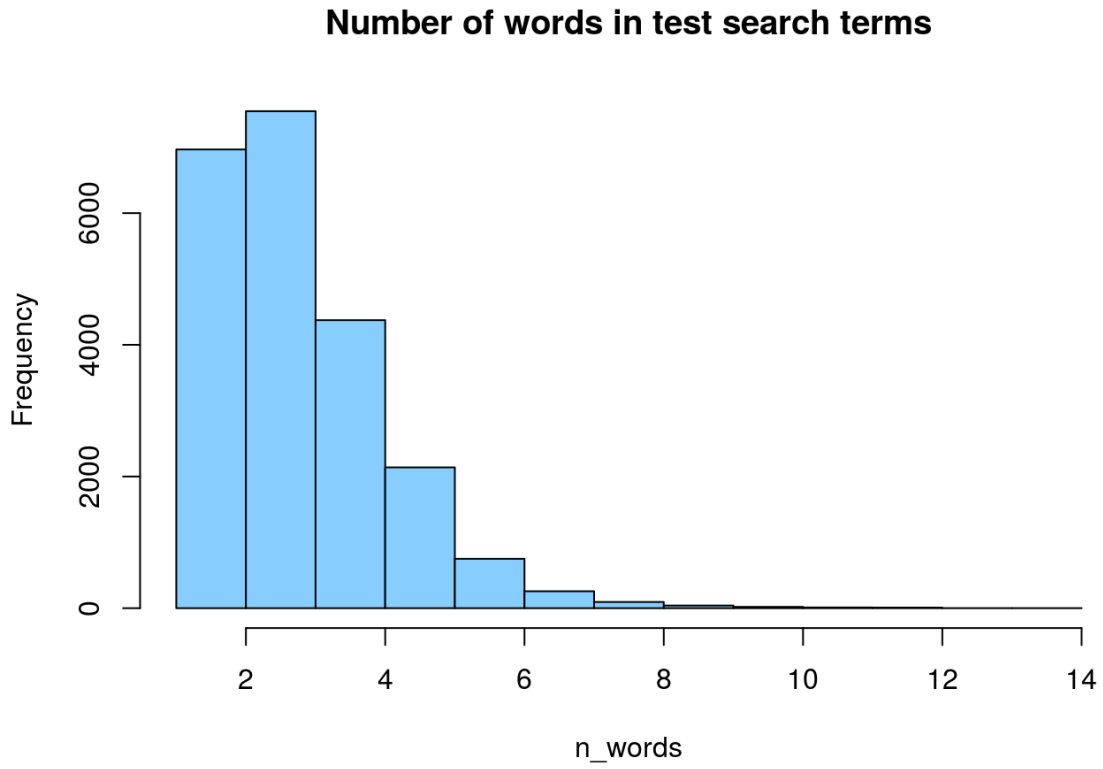

```
test_term_words <- sapply(st_test,n_words)
cat('Number of words in test terms\n')
```

```
## Number of words in test terms
```

```
print(table(test_term_words))
```

```
## test_term_words
##    1    2    3    4    5    6    7    8    9   10   11   12   13   1
4
## 1273 5693 7548 4374 2138  750  257   94   41   20   11    9    1
1
```

```
hist(test_term_words,main="Number of words in test search terms", xlab=
"n_words",ylab="Frequency",col="skyblue1")
```




```
cat('Words in train terms\n')
```

```
## Words in train terms
```

```
train_term_frequency <- get_word_frequency(st_train)
set.seed(1)
wordcloud(names(train_term_frequency), train_term_frequency, min.freq =
 60)
```


```
#cat('Words in test terms\n')
test_term_frequency <- get_word_frequency(st_test)
set.seed(1)
wordcloud(names(test_term_frequency), test_term_frequency, min.freq = 6
0)
```


```
## Exploring product data
train_pid <- unique(train$product_uid)
test_pid <- unique(test$product_uid)
common_pid <- intersect(train_pid,test_pid)
unique_train_pid <- setdiff(train_pid,test_pid)
unique_test_pid <- setdiff(test_pid,train_pid)

print(paste("Number of unique product ids in train :",(length(train_pid
)),sep=" "))
```

```
## [1] "Number of unique product ids in train : 54667"
```

```
print(paste("Number of unique product ids in test :",(length(test_pid))
,sep=" "))
```

```
## [1] "Number of unique product ids in test : 97460"
```

```
print(paste("Number of common product ids :",(length(common_pid)),sep="
 "))
```

```
## [1] "Number of common product ids : 27699"
```

```
print(paste("Number of pid in train not in test :",(length(unique_train
_pid)),sep=" "))
```

```
## [1] "Number of pid in train not in test : 26968"
```

```
print(paste("Number of pid in test not in train :",(length(unique_test_
pid)),sep=" "))
```

```
## [1] "Number of pid in test not in train : 69761"
```

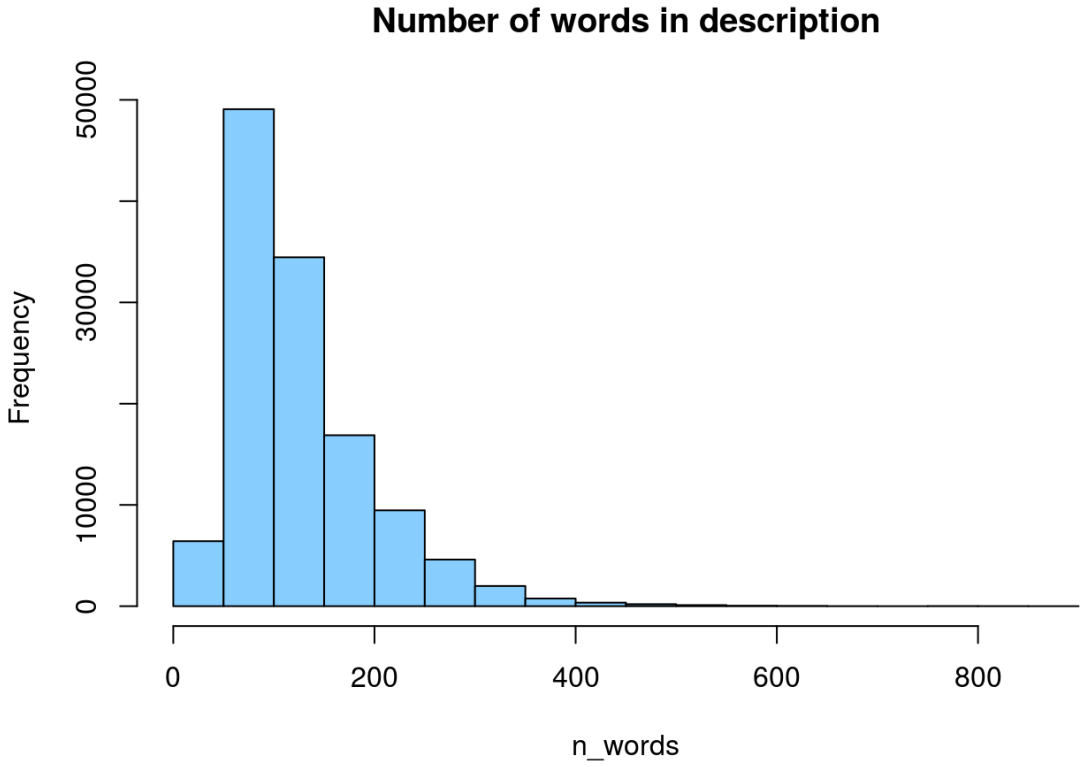
```
## Exploring product description
desc_length <- sapply(desc$product_description,n_words)
cat('Distribution of number of words in description\n')
```

```
## Distribution of number of words in description
```

```
print(quantile(desc_length,probs = seq(0,1,.1)))
```

```
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1   58   71   82   94  108  124  145  174  222  889
```

```
hist(desc_length,main="Number of words in description", xlab="n_words",
ylab="Frequency",col="skyblue1")
```




```
## Exploring attribute data
nids <- unique(attr$product_uid)
print(paste("Number of ids with attributes:",(length(nids)),sep=" "))
```

```
## [1] "Number of ids with attributes: 86264"
```

```
nattr <- sqldf('select product_uid,count(*) as n_attr from attr group b
y product_uid')
```

```
## Loading required package: tcltk
```

```
nattr <- nattr[!is.na(nattr$product_uid),]
cat('Distribution of number of attributes for products\n')
```

```
## Distribution of number of attributes for products
```

```
print(quantile(nattr$n_attr,probs = seq(0,1,.1)))
```

```
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    5   13   15   18   20   22   25   28   31   36   88
```

```
hist(nattr$n_attr,main="Number of attributes per product", xlab="N Attr
ibutes",ylab="Frequency",col="skyblue1")
```