Swiftkey_Capstone

Jeet Tanna

8/02/2020

Read in the appropriate data

```
news <- readLines("C:/Users/JeetsPC-1/Desktop/Study Material/R DataSets/Coursera-SwiftKey/final/
en_US/en_US.news.txt", encoding="UTF-8", skipNul = TRUE, warn = TRUE)
```

```
## Warning in readLines("C:/Users/JeetsPC-1/Desktop/Study Material/R DataSets/
## Coursera-SwiftKey/final/en_US/en_US.news.txt", : incomplete final line
## found on 'C:/Users/JeetsPC-1/Desktop/Study Material/R DataSets/Coursera-
## SwiftKey/final/en_US/en_US.news.txt'
```

```
twitter <- readLines("C:/Users/JeetsPC-1/Desktop/Study Material/R DataSets/Coursera-SwiftKey/fin
al/en_US/en_US.twitter.txt",encoding="UTF-8", skipNul = TRUE, warn = TRUE)
```

load the libraries

```
library(ggplot2)
library(NLP)
```

```
##
## Attaching package: 'NLP'
```

```
## The following object is masked from 'package:ggplot2':
##
##     annotate
```

```
library(tm)
```

```
## Warning: package 'tm' was built under R version 3.5.3
```

```
library(RWeka)
```

```
## Warning: package 'RWeka' was built under R version 3.5.3
```

```
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 3.5.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Sample the data

```
set.seed(1130)
samp_size = 1500

news_samp <- news[sample(1:length(news),samp_size)]
twitter_samp <- twitter[sample(1:length(twitter),samp_size)]

df <-rbind(news_samp,twitter_samp)
rm(news,twitter)
```

Create the corpus

```
corp <- VCorpus(VectorSource(df))

corp <- tm_map(corp, tolower)
corp <- tm_map(corp, removePunctuation)
corp <- tm_map(corp, removeNumbers)
corp <- tm_map(corp, stripWhitespace)
corp <- tm_map(corp, PlainTextDocument)
changetospace <- content_transformer(function(x, pattern) gsub(pattern, " ", x))
corp <- tm_map(corp, changetospace, "/|@|\\|")
```

use a tokenizer to break speeck into components that can be read my machine

```
uniGramTokenizer <- function(x) NGramTokenizer(x, Weka_control(min = 1, max = 1))
biGramTokenizer <- function(x) NGramTokenizer(x, Weka_control(min = 2, max = 2))
triGramTokenizer <- function(x) NGramTokenizer(x, Weka_control(min = 3, max = 3))
quadGramTokenizer <- function(x) NGramTokenizer(x, Weka_control(min = 4, max = 4))
OneT <- NGramTokenizer(corp, Weka_control(min = 1, max = 1))
oneGM <- TermDocumentMatrix(corp, control = list(tokenize = uniGramTokenizer))
twoGM <- TermDocumentMatrix(corp, control = list(tokenize = biGramTokenizer))
threeGM <- TermDocumentMatrix(corp, control = list(tokenize = triGramTokenizer))
fourGM <- TermDocumentMatrix(corp, control = list(tokenize = quadGramTokenizer))
```

Find the most frequently used terms

```
freqTerms1 <- findFreqTerms(oneGM, lowfreq = 5)
termFreq1 <- rowSums(as.matrix(oneGM[freqTerms1,]))
termFreq1 <- data.frame(unigram=names(termFreq1), frequency=termFreq1)
termFreq1 <- termFreq1[order(-termFreq1$frequency),]
unigramlist <- setDT(termFreq1)
save(unigramlist,file="unigram.Rds")
freqTerms2 <- findFreqTerms(twoGM, lowfreq = 3)
termFreq2 <- rowSums(as.matrix(twoGM[freqTerms2,]))
termFreq2 <- data.frame(bigram=names(termFreq2), frequency=termFreq2)
termFreq2 <- termFreq2[order(-termFreq2$frequency),]
bigramlist <- setDT(termFreq2)
save(bigramlist,file="bigram.Rds")
freqTerms3 <- findFreqTerms(threeGM, lowfreq = 2)
termFreq3 <- rowSums(as.matrix(threeGM[freqTerms3,]))
termFreq3 <- data.frame(trigram=names(termFreq3), frequency=termFreq3)
trigramlist <- setDT(termFreq3)
save(trigramlist,file="trigram.Rds")
freqTerms4 <- findFreqTerms(fourGM, lowfreq = 1)
termFreq4 <- rowSums(as.matrix(fourGM[freqTerms4,]))
termFreq4 <- data.frame(quadgram=names(termFreq4), frequency=termFreq4)
quadgramlist <- setDT(termFreq4)
save(quadgramlist,file="quadgram.Rds")
```