Weibo Scraping

Paine (HE Peiwei, 54471952)

Research Topic: Scraping Weibo contents

First, require for all the packages.

```
library(plyr)
library(httr)
library(jiebaR)
```

```
## Loading required package: jiebaRD
```

```
library(rJava)
library(Rwordseg)
```

```
## # Version: 0.2-1
```

Second, set up for the Weibo API.

```
endpoint = oauth_endpoint(
  authorize = 'https://api.weibo.com/oauth2/authorize',
  access = 'https://api.weibo.com/oauth2/access_token')
APP_KEY = "3756418011"
APP_SECRET = "a754ecf89b984b6196d8983a45291ecb"
app = oauth_app(endpoint, key = APP_KEY, secret = APP_SECRET)
```

```
system("defaults write org.R-project.R force.LANG en_US.UTF-8")
```

```
## Warning: 运行命令'defaults write org.R-project.R force.LANG en_US.UTF-8'的
## 状态是127
```

Third, create funcion to get Weibo contents by using the API.

```
weibo_get <- function(token, path, ...) {
  url = paste0("https://api.weibo.com/2/", path, ".json")
  query = c(list(access_token="2.00ixXfJCRvXNGEde90cd86260UZF73"), list(...))
  res = httr::GET(url, query=query)
  stop_for_status(res, task = paste("Query weibo:", url))
  content(res)
}
```

Fourth, get the statuses of my bilateral friends, the statuses on my home timeline, and the statuses of my followings. As the API can only return 20 once, so I scrap the weibo for 3 times, and save the weibo contents as RData, there are 60 weibo contents in total.

```
w=weibo_get(token, 'statuses/bilateral_timeline')
weibotext = list()
for(i in 1:20){
  wt=w$statuses[[i]]$text
  weibotext <- c(wt,weibotext)
}

ft=weibo_get(token, 'statuses/friends_timeline')
fwtext = list()
for(i in 1:20){
  fwt=ft$statuses[[i]]$text
  fwtext <- c(fwt,fwtext)
}
alltext=cbind(weibotext ,fwtext)
save(alltext, file="scrapweibo.RData")
```

```
ht=weibo_get(token, 'statuses/home_timeline')
```

```
httext = list()
for(i in 1:20){
  htt=ht$statuses[[i]]$text
  httext <- c(htt,httext)
}
load("scrapweibo.RData")
alltext=cbind(alltext ,httext)
alltext=ldply(alltext, rbind)
colnames(alltext)[1]="text"
```

Fitht, I save the dataframe into txt file, so as to do the Chinese words segment.

```
write.table(alltext, file = "alldf.txt", sep = ",", col.names = colnames(alltext))

all2 <- read.table(file = "alldf.txt",sep = ",", header = TRUE, stringsAsFactors = FALSE)
```

Sixth, by using the Rwordseg, there is a txt file output of the segment words. And I count the frequency of these words, and get the top 100 words to a dataframe.

```
text=segmentCN("alldf.segment.2017-03-06_19_24_27.txt",returnType = "tm")
```

```
## Output file:  alldf.segment.2017-03-06_19_24_27.segment.txt
```

```
text1=readLines("alldf.segment.2017-03-06_19_24_27.segment.txt ", encoding = "UTF-8")
word = lapply(X = text1, FUN = strsplit, "\\s")
word1=unlist(word)
df=table(word1)
df=sort(df, decreasing = T)
```

```
d=data.frame(df)
newd = head(d, n=100)
newd<-newd[-1,]
head(newd)
```

```
##   word1 Freq
## 2     B  273
## 3     U  273
## 4   200  259
## 5     了  259
## 6     马  217
## 7     在  189
```

Finally, create a wordcloud base on the segment.

```
library(wordcloud2)
wordcloud2(newd, color = "random-light", backgroundColor = "white")
```