Assignment 5 Bonus Question

Mirrien Liang

04/03/2022

Import necessary libraries and data:

Latest tweet has been imported and saved as a local file

In the following code, along with an *access.r* file with user key and secret, one can load data from Twitter

```
library(tidyverse)
# library(rtweet)
library(twitteR)

# library(tidytext)
library(stringr)

# source('access.r') # contains key and secret

# Load 3200 tweets without retweets and replies

# data = userTimeline(
#   'Translink',
#   n = 3200,
#   includeRts = F,
#   excludeReplies = T
# )
#
# data[[1]]$text
# data[[1]]$created
#
# save(file = 'translink.data', data) # save to local file

load('translink.data')
```

We want to *tidy* data so that the function can run faster.

First, we extract post content and date/time:

```
time=c()
tweet=c()

for (i in seq_along(data)) {
```

```
  time[i]=as.character(data[[i]]$created)
  tweet[i]=data[[i]]$text
}

f=data.frame(Time = time, Post = tweet)
# head(f)
```

Second, we remove post rows that are irrelevant to bus routes.

```
# It's too long that I have to cut it and use str_c to combine strings
pattern_non_bus = str_c("Skytrain|SkyTrain|StationAlert|Elevator|elevator|",
                        "Expo|WCE|West Coast Express|board|time|Time|SeaBus|",
                        "Compass|Handy|morning|Morning|evening|desk|Desk|",
                        "Transit|transit|Good|Rehab| night |tonight| call|",
                        "will be|changes|Congrats|Pattullo Bridge|Valentines|",
                        "year|Year|multi")
df <- f[!grepl(pattern_non_bus,f$Post),]
```

Third, remove url substrings that can cause confusion:

```
pattern_url = "(.*) (.*)" #Greedy match to the last space
for (i in seq_along(df$Post)) {
  if (str_detect(df$Post[i],"http")) {
    df$Post[i] <- str_replace(df$Post[i],pattern_url,"\\1")
  }
}
```

Now, extract all types of numbers (route#, road/street#, time, date, duration):

```
pattern_number = str_c("([A-Z]|Bay\\s|Hwy\\s|stop\\s|Regular.*|Jan\\s|Feb\\s|",
                       "Feb.\\s|\\February\\s|Mar\\s)?",
                       "(\\d){1,3}(\\d+|\\s|:\\d+)?",
                       "(Ave|St |St,|St/|St...|Rd|th|st|nd|rd|pm|",
                       "PM|am|AM|min|minutes|due|Station|...)?")

df1 <- mutate(df,Numbers = str_extract_all(Post,pattern_number))
```

Next, extract bus routes.

```
pattern_bus_route = "^[A-Z]?\\d{1,3}\\s?$"

for (i in seq_along(df1$Numbers)){
  df1$Routes[[i]] <- na.omit(str_extract(df1$Numbers[[i]],pattern_bus_route))
}

df2 <- df1 %>% filter(!(Routes=="character(0)")) %>%
  select(-Numbers)
```

Lastly, include Status of posts.

```
pattern_status = str_c("^(?!.*( clear|Clear|CLEAR| ended| back| over|cancel|",
                       "return|ease )).*(regular route|Regular route|onward|",
                       "detour|experienc|suspend|off )")
df2 <- mutate(df2, Status = str_detect(Post,pattern_status))
```

Overview of the tidy data frame

Let's take a look into the tidy data frame:

```
head(df2,n=5)
```

```
##                  Time
## 1 2022-03-04 18:10:47
## 2 2022-03-04 15:16:55
## 3 2022-03-04 01:34:54
## 4 2022-03-04 00:53:10
## 5 2022-03-04 00:51:09
##
## 1 #RiderAlert Update. 301 Brighouse Station/Newton Exchange service is back on regular schedules aft
## 2      #RiderAlert 301 Brighouse Station/Newton Exchange service is experiencing delays due to motor
## 3                 #RiderAlert Update - 23 Main Street Station detour has been cleared. Buses are resu
## 4      #RiderAlert 123 New Westminster Station detour: Regular route to Willingdon and Dawson, then
## 5      #RiderAlert 25 UBC detour: Regular route to Dawson and Willingdon, then cont. Dawson, Rosser,
##   Routes Status
## 1   301   FALSE
## 2   301    TRUE
## 3    23   FALSE
## 4   123    TRUE
## 5    25    TRUE
```

Construct the `translink()` function as in the previous questions

```
translink <- function(y,m,d,h){
  # take four arguments of time, e.g., 2020,1,26,3

  # Create an empty list to store results
  ret = list(start=c(),stop=c())

  # format input ymd_h
  # i.e., (2020,1,26,3) -> "2020-01-26 03"
  if(nchar(m)==1){m=str_c("0",m)}
  if(nchar(d)==1){d=str_c("0",d)}
  if(nchar(h)==1){h=str_c("0",h)}
  date = str_c(y,m,d,sep="-")
  datetime = str_c(date,h,sep = " ")

  # Match time, then match Status, then combine/append vectors
```

```
  for (i in seq_along(df2$Time)) {
    if (str_detect(df2$Time[[i]], datetime)){
      if (df2$Status[[i]]) {
        ret$start <- c(ret$start, df2$Routes[[i]])
      } else {
        ret$stop <- c(ret$stop, df2$Routes[[i]])
      }
    }
  }


  # Remove all the Space in the Routes column
  ret$start <- gsub('\\s+', '', ret$start)
  ret$stop <- gsub('\\s+', '', ret$stop)

  # Delete duplicates
  ret$start = unique(ret$start)
  ret$stop = unique(ret$stop)

  # If no match, append warning message
  if (length(ret$start)==0) {ret$start <- c(ret$start,"No detour had started")}
  if (length(ret$stop)==0) {ret$stop <- c(ret$stop,"No detour had ended")}

  return(ret)
}
```

Test cases:

(1) On March 2, 2022 at 16:00, there are three tweets. We expect to find routes 301, 340, 28, 130, and 222 stopped detour and 28, 130, and 222 started detour.

```
df2$Post[22:24]
```

```
## [1] "#RiderAlert 301/340 delays have eased. Resuming regular schedules. ^RR"
## [2] "#RiderAlert  28/130/222 Phibbs Exchange detour has ended. resuming regular route. ^RR"
## [3] "#RiderAlert 28/130/222 Phibbs Exchange detour. Regular route to Hastings &amp;amp; Hwy 1 then conti
```

```
translink(2022,3,2,16)
```

```
## $start
## [1] "28"  "130" "222"
##
## $stop
## [1] "301" "340" "28"  "130" "222"
```

(2) On March 1, 2022 at 13 o'clock, there are two tweets. These are used to test the extracting functionality of numbers with different formats.

```
df2$Post[53:54]
```

```
## [1] "#RiderAlert Update. 4 UBC/7Dunbar/10 Granville/14 UBC/16 Arbutus/50 False Creek detours have cl
## [2] "#RiderAlert 4 UBC/7Dunbar/10 Granville/14 UBC/16 Arbutus/50 False Creek detour. Regular route t
```

```
translink(2022,3,1,13)
```

```
## $start
## [1] "4"  "7"  "10" "14" "16" "50"
##
## $stop
## [1] "4"  "7"  "10" "14" "16" "50"
```

(3) On January 31 at 16 o'clock, there are 9 tweets. This test is used to test the extracting functionality of numbers with different suffix/prefix.

```
df2$Post[585:593]
```

```
## [1] "#RiderAlert UPDATE: 257 Horseshoe Bay detour has cleared, back to regular route ^LA"
## [2] "#RiderAlert 257 Horseshoe Bay detour. Regular route to Marine Drive and 15 St, then continue Ma
## [3] "#RiderAlert UPDATE. 342 Newton Exchange detour due to fire has cleared, back to regular route ^
## [4] "#RiderAlert UPDATE:. 342 Newton Exchange detour. Now regular route to 56 Ave and 200 St, then 2
## [5] "#RiderAlert UPDATE: 28 Joyce Stn/Phibbs Ex, 130 Phibbs Ex/Metrotown Stn, 222 Phibbs Ex/Metrotow
## [6] "#RiderAlert 222 Metrotown Station/Phibbs Exchange detour. Regular route to Willingdon and Pende
## [7] "#RiderAlert 130 Metrotown Station/Phibbs Exchange detour. Regular route to Willingdon and Pende
## [8] "#RiderAlert 28 Joyce Station/Phibbs Exchange detour. Regular route to Kootenay Loop as 28 Phibb
## [9] "#RiderAlert 342 Newton Exchange detour. Regular route to 56 Ave and 200 St, then 200 St, Fraser
```

```
translink(2022,1,31,16)
```

```
## $start
## [1] "257" "342" "28"  "130" "222"
##
## $stop
## [1] "342" "210" "211" "28"  "130" "222"
```

```
# rows such as 594 do not include its status; should have been filtered out
# but the default assignment of status goes "FALSE"
```