HW3

Anya Conti

February 25, 2017

1. First the library stringr is imported. Then the command readLines() stores the file into the character vector nhl1415.
a. length() is used to to find the number of lines the file has which is 3141.
b. nchar() is used to find the number of characters in each line individually, and then this is summed to find the total number of characters in the file which is 2482700.
c. nchar() is used to find the number of characters in each line individually, and then the max of these is found with max(), which is 16047.

```
library(stringr)
nhl1415 <- readLines("http://people.math.umass.edu/~wei/Teaching/STAT597_Spring17/NHLHockeySchedule2.htm

length(nhl1415)
```

```
## [1] 3141
```

```
sum(nchar(nhl1415))
```

```
## [1] 2482700
```

```
max(nchar(nhl1415))
```

```
## [1] 16047
```

2. The first game is being played by Montreal and Toronto. The final game is being played by Edmonton and Vancouver.

3. The line which corresponds to the first game is 480 and the line which corresponds to the last game is 2950. Each of these lines begins with HTML tags, which for most lines except the first game is

“, value = TRUE) “‘

4. The corresponding URL for each game is actually at the start of the line for the following game (usually 2 lines after), and for the last game is 2 lines after.

5. First grep is used to get the lines that have the date pattern modeled by the regular expression below. The length is 1230 as it’s supposed to be. The entries for the first and last games are also shown below. The first game is Wed Oct 8th being played by Montreal, and the last is Sat Apr 11th being played by Edmonton, so these do match.

```
dates <- grep(nhl1415, pattern = "[A-Z][a-z]{2} [0-9]*, [0-9]{4}", value = TRUE)
length(dates)
```

```
## [1] 1230
```

```
dates[1]
```

```
## [1] "</h3><div style=\"float: right; margin-top: -26px;\" class=\"newsTools\"><a shape=\"rect\" href=
```

```
dates[1230]
```

```
## [1] "</td><td colspan=\"1\" rowspan=\"1\" class=\"skedLinks\"><!-- Button Links --><a class=\"btn\" s
```

6. Using str_match on Lines (which contains all the lines that have games in them), the formatting pattern for the dates was used to extract all the dates and stored it in a vector called dates. Then as.Date() was used to turn that vector into a date vector, and that was stored in dates.

```
dates <- str_match(dates, "[A-Z][a-z]{2} [0-9]*, [0-9]{4}")
dates <- as.Date(dates,'%b %d, %Y')
```

7. Using strsplit(), the Home teams were extracted from the Lines vector and stored in a character vector called HTeam, and the Away teams were extracted and stored in a character vector called ATeam.

```
HTeam <- unlist(strsplit(Lines,split = "<!-- Home -->"))[seq(from=2, to=(2*1230), by=2 )]
HTeam <- unlist(strsplit(HTeam,split = "<!-- Time -->"))[seq(from=1, to=(2*1230-1), by=2 )]
HTeam <- unlist(strsplit(HTeam,split = "<!-- Black Background 50% Opacity Overlay -->"))[seq(from=2, to=
HTeam <- unlist(strsplit(HTeam,split = "void"))[seq(from=2, to=(2*1230), by=2 )]
HTeam <- unlist(strsplit(HTeam,split = "</a>"))[seq(from=1, to=(2*1230-1), by=2 )]
HTeam <- unlist(strsplit(HTeam,split = ">"))[seq(from=2, to=(2*1230), by=2 )]

ATeam <- unlist(strsplit(Lines,split = "<!-- Away -->"))[seq(from=2, to=(2*1230), by=2 )]
ATeam <- unlist(strsplit(ATeam,split = "<!-- Home -->"))[seq(from=1, to=(2*1230-1), by=2 )]
ATeam <- unlist(strsplit(ATeam,split = "<!-- Black Background 50% Opacity Overlay -->"))[seq(from=2, to=
ATeam <- unlist(strsplit(ATeam,split = "void"))[seq(from=2, to=(2*1230), by=2 )]
ATeam <- unlist(strsplit(ATeam,split = "</a>"))[seq(from=1, to=(2*1230-1), by=2 )]
ATeam <- unlist(strsplit(ATeam,split = ">"))[seq(from=2, to=(2*1230), by=2 )]
```

8. Using the str_match() command on Lines, the pattern for times was used to extract all the times and then store them in a vector called times. The ET was added at the end of the pattern to ensure the Eastern standard time was obtained.

```
times <- str_match(Lines,"[0-9]*:[0-9]{2} [A-Z]{2} ET")
```

9. First using the grep() command I extracted the all the lines from nhl1415 that had "TICKETS" in them, because that is the tag associated with the URL. Then I used strsplit() to remove all the text on either side of the URL's, and stored the URL's in a string vectore called Tix.

```
webLines <- grep(nhl1415, pattern = "<span>TICKETS", value = TRUE)

Tix <- unlist(strsplit(webLines, split="<span>TICKETS></span></a><a class=\"btn\" shape=\"rect\" href=\
Tix <- unlist(strsplit(Tix, split="href=\""))[seq(from=2, to=(2*1230), by=2 )]
Tix <- unlist(strsplit(Tix, split="\\\">"))
```

10. A data frame with a column for each of the extracted variable vectors is created with the data.fram() command, and that is stored in GameData. Then the rows 1221 through 1230 are shown below, and they do match the last 10 columns of the table in the web browser.

```
GameData <- data.frame("Date" = dates, "Visiting Team" = ATeam, "Home Team" = HTeam, "Time" = times, "T
GameData[1221:1230,]
```

```
##            Date Visiting.Team    Home.Team        Time
## 1221 2015-04-11      Montreal      Toronto  7:00 PM ET
## 1222 2015-04-11    New Jersey      Florida  7:00 PM ET
## 1223 2015-04-11      Columbus NY Islanders  7:00 PM ET
## 1224 2015-04-11       Detroit     Carolina  7:00 PM ET
## 1225 2015-04-11        Boston    Tampa Bay  7:30 PM ET
## 1226 2015-04-11     Minnesota    St. Louis  7:30 PM ET
## 1227 2015-04-11     Nashville       Dallas  8:00 PM ET
## 1228 2015-04-11       Chicago     Colorado  9:00 PM ET
## 1229 2015-04-11       Anaheim      Arizona  9:00 PM ET
## 1230 2015-04-11      Edmonton    Vancouver 10:00 PM ET
##
## 1221                    http://www.ticketmaster.ca/artist/806033?intcmp=tm204081&wt.mc_id=NHL_LEAGUE_
## 1222 http://www.ticketmaster.com/artist/805945?brand=nhlgeneric&intcmp=tm204066&wt.mc_id=NHL_LEAGUE_
```

```
## 1223 http://www.ticketmaster.com/artist/805986?brand=nhlgeneric&intcmp=tm204072&wt.mc_id=NHL_LEAGUE_
## 1224 http://www.ticketmaster.com/artist/805908?brand=nhlgeneric&intcmp=tm204059&wt.mc_id=NHL_LEAGUE_
## 1225                   http://www.ticketmaster.com/artist/806028?intcmp=tm204080&wt.mc_id=NHL_LEAGUE
## 1226                                                 http://www.ticketmaster.com/St-Louis-Blues-ticket
## 1227                  http://www.ticketmaster.com/artist/805933?intcmp=tm204063&wt.mc_id=NHL_LEAGUE_
## 1228                                                      http://www.tickethorse.com/sports/hockey/col
## 1229                  http://www.ticketmaster.com/artist/806002?intcmp=tm204076&wt.mc_id=NHL_LEAGUE_
## 1230  http://www.ticketmaster.ca/artist/806037?brand=nhlgeneric&intcmp=tm204082&wt.mc_id=NHL_LEAGUE_
```