hw3_yw3204

wyh

10/3/2018

i)

ii)

```
# read html file
nets1819 <- readLines("NetsSchedule1819.html", warn = FALSE)

# total lines
length(nets1819)
```

```
## [1] 104
```

```
# total characters
sum(nchar(nets1819))
```

```
## [1] 462979
```

```
# maxmimum # of characters in a single line
max(nchar(nets1819))
```

```
## [1] 249787
```

iii)

They were playing with Detroit first on Wed, Oct 17 and playing with Miami last on Wed, Apr 10.

iv)

Line 64.

v)

```
s64 <- nets1819[64]

# define starting regexp pattren
p_0 <- "\\[\\{\"date\":"

# define ending regexp pattern
p_1 <- "\"notes\":\\{\\}\\}\\]"

# find positions
gregexpr(p_0, s64)
```

```
## [[1]]
## [1] 99749
```

```
## attr(,"match.length")
## [1] 9
```

```
gregexpr(p_1, s64)
```

```
## [[1]]
## [1] 198675
## attr(,"match.length")
## [1] 12
```

```
# extract
s <- substr(s64, 99751, 198684)
```

vi)

```
# split into 82 substrings and unlist list
s_1 <- strsplit(s, split = "\\},\\{")
s_1 <- unlist(s_1)
```

vii)

```
# define date regexp pattern
p_2 <- "[0-9]{4}-[0-9]{2}-[0-9]{2}"
grep(p_2, s_1)
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
## [24] 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46
## [47] 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69
## [70] 70 71 72 73 74 75 76 77 78 79 80 81 82
```

```
grep("Detroit", s_1)
```

```
## [1]  1  8 69
```

```
grep("Miami", s_1)
```

```
## [1] 15 18 65 82
```

Yes, we have found 82 lines and the locations of the first and last games match with that in (ii).

viii)

```
# definedate pattern
p_4 <- "[A-z]+,\\s[A-z]*\\s[0-9]+(th|st|nd|rd)"
# grepl(p_4, s_1)
date <- regmatches(s_1, gregexpr(p_4, s_1))
date <- unlist(date)
```

ix)

```r
# define time pattern
p_5 <- "[0-9]+:[0-9]+\\sPM\\s(EDT|EST)"
# grepl(p_5, s_1)
time <- regmatches(s_1, regexpr(p_5, s_1))
time <- unlist(time)
```

x)

```r
# define home or away pattern
p_6 <- "\"homeAwaySymbol\":\"(@|vs)\""
# grepl(p_6, s_1)
home <- regmatches(s_1, regexpr(p_6, s_1))
home <- unlist(home)
home <- substr(home, 19, nchar(home)-1)
```

xi)

```r
# define opponent pattern
p_7 <- "\"displayName\":\"[A-z0-9\\ ]+\""
# grepl(p_7, s_1)
opponent <- regmatches(s_1, regexpr(p_7, s_1))
opponent <- unlist(opponent)
opponent <- substr(opponent, 16, nchar(opponent)-1)
```

xii)

```r
# create data frame based on the info we extrcted
nets_df <- data.frame(date, time, opponent, home)
head(nets_df, 10)
```

```
##                  date        time              opponent home
## 1  Wed, October 17th 7:00 PM EDT       Detroit Pistons    @
## 2  Fri, October 19th 7:30 PM EDT       New York Knicks   vs
## 3  Sat, October 20th 7:00 PM EDT        Indiana Pacers    @
## 4  Wed, October 24th 7:00 PM EDT   Cleveland Cavaliers    @
## 5  Fri, October 26th 8:00 PM EDT  New Orleans Pelicans    @
## 6  Sun, October 28th 5:00 PM EDT Golden State Warriors   vs
## 7  Mon, October 29th 7:30 PM EDT       New York Knicks    @
## 8  Wed, October 31st 7:30 PM EDT       Detroit Pistons   vs
## 9  Fri, November 2nd 7:30 PM EDT       Houston Rockets   vs
## 10 Sun, November 4th 6:00 PM EST    Philadelphia 76ers   vs
```

Yes, it matches.