lab6Q1.R

mirrien

2022-03-15

```
library(stringr)
course_url ="https://www.sfu.ca/outlines.html?2022/spring/stat/240/d100"
course_page = readLines(course_url)



##################################################
# Q1.2 Course name (e.g., STAT 240)

# extract course name
cn = grep('<h1\\sid=\"name\"',course_page, v=T)

# remove section number
cn_1 = gsub("<span>.*</span>","",cn)

# remove html formatting pieces and trim white spaces
cn_1 = trimws(gsub("<[^<>]+>"," ",cn_1))

# remove Term-Year, e.g., Spring 2022
cn_2 = trimws(str_extract(cn_1,'\\s[A-Z]+.*'))

# print result
print(cn_2)
```

```
## [1] "STAT 240"
```

```
##################################################
# Q1.3 Course title (e.g., Introduction to Data Science)

# locate, extract, and tidy course title
ct = trimws(course_page[grep('<h2\\sid=\"title\"',course_page)+1])

# print result
print(ct)
```

```
## [1] "Introduction to Data Science"
```

```
##################################################
# Q1.4 Instructor
```

```
# Locate, extract, and tidy course instructor
ci = trimws(course_page[grep('<h4>Ins', course_page)+1])

# Remove html formatting pieces
ci_1 = gsub("<[^<>]+>","",ci)

# print result
print(ci_1)
```

```
## [1] "Lloyd Elliott"
```

```
################################################
# Q1.5 Course Times + Location

# Locate, extract, and tidy course times and location
ctl = trimws(course_page[grep('<h4>Course', course_page)+1])

# Replace UTF-8 punctuation
ctl = str_replace_all(ctl,"\\&ndash;","-")

# Replace the HTML line break tag with space
ctl = str_replace_all(ctl,"<br>", " ")

# Remove HTML formatting pieces
ctl_1 = gsub("<[^<>]+>","",ctl)

# print result
print(ctl_1)
```

```
## [1] "Mo 12:30 PM - 2:20 PM AQ 3182, Burnaby"
```