web_scrape

Jonathan Cheon

10/24/2019

Loading in packages

```
include <- function(library_name){
  if( !(library_name %in% installed.packages()) )
    install.packages(library_name)
  library(library_name, character.only=TRUE)
}

include("rvest")
```

```
## Loading required package: xml2
```

```
include("tidyr")
include("dplyr")
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Scarping data from website and making a tibble.

```
sp19 <- "http://ems.csuchico.edu/APSS/schedule/spr2019/CSCI.shtml"


sp19_html <- read_html(sp19)

data_sp19 <- sp19_html %>%
              html_nodes(".classrow")

sp19_subj <- data_sp19 %>%
              html_nodes("td.subj") %>%
              html_text()

sp19_cat_num <- data_sp19 %>%
                html_nodes("td.cat_num") %>%
                html_text()

sp19_title <- data_sp19 %>%
```

```
              html_nodes("td.title") %>%
              html_text()

sp19_instructor <- data_sp19 %>%
                    html_nodes("td.Instructor") %>%
                    html_text()

sp19_enrtot <- data_sp19 %>%
                html_nodes("td.enrtot") %>%
                html_text()

sp19_table <- tibble(subj= sp19_subj,
              cat_num= sp19_cat_num,
              title= sp19_title,
              instructor= sp19_instructor,
              enrtot= sp19_enrtot)
```

New and improved with scarping data from each of the website with a function. This function will take in a url and return a tibble with data we need.

```
read_class_schedule <- function(url)
{

  website <- read_html(url)

  data <- website %>%
            html_nodes(".classrow")

  subj <- data %>%
            html_nodes("td.subj") %>%
            html_text()

  cat_num <- data %>%
              html_nodes("td.cat_num") %>%
              html_text()

  title <- data %>%
            html_nodes("td.title") %>%
            html_text()

  instructor <- data %>%
                  html_nodes("td.Instructor") %>%
                  html_text()

  enrtot <- data %>%
              html_nodes("td.enrtot") %>%
              html_text()

  table <- tibble(subj= subj,
              cat_num= cat_num,
              title= title,
              instructor= instructor,
              enrtot= enrtot)
```

```
  return(table)

}
```

Calls read_class_schedule and creates a tibble for each of them.

```
csci_spr2019 <- read_class_schedule("http://ems.csuchico.edu/APSS/schedule/spr2019/CSCI.shtml")

csci_spr2020 <- read_class_schedule("http://ems.csuchico.edu/APSS/schedule/spr2020/CSCI.shtml")

math_spr2019 <- read_class_schedule("http://ems.csuchico.edu/APSS/schedule/spr2019/MATH.shtml")

math_spr2020 <- read_class_schedule("http://ems.csuchico.edu/APSS/schedule/spr2020/MATH.shtml")
```

Now we will combine those into one final tibble.

```
final <- rbind(csci_spr2019, csci_spr2020, math_spr2019, math_spr2020)
```