hw_4

Andrew Mikolinski

2023-10-31

library(“xml2”)

Question 1

```
suppressWarnings({
library("rvest")
library("tidyverse")
})
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter()         masks stats::filter()
## x readr::guess_encoding() masks rvest::guess_encoding()
## x dplyr::lag()            masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
url <- "https://introdatasci.dlilab.com/schedule_materials/"
page <- read_html(url)
schedule_table <- page %>%
  html_nodes("table") %>%
  .[[1]] %>%
  html_table()
materials_table <- page %>%
  html_nodes("table") %>%
  .[[2]] %>%
  html_table()
schedule_table
```

```
## # A tibble: 31 x 5
##    Date   Topic                                                 Notes    HW Reading
##    <chr>  <chr>                                                 <chr> <int> <chr>
##  1 Aug 22 About the course                                      "\U0~    NA "Leek ~
##  2 Aug 24 Data science project cycle                            "\U0~    NA "Mason~
##  3 Aug 29 Introduction and install tools                 "\U0~    NA "Coope~
##  4 Aug 31 Version control with Git                       "\U0~    NA "Blisc~
##  5 Sep 05 Introduction to GitHub                         "\U0~    NA ""
##  6 Sep 07 RStudio project and dynamic documents with R Mark~ "\U0~     1 "Xie e~
##  7 Sep 12 R basics: data types, vectors, matrix, data frame~ "\U0~    NA ""
##  8 Sep 14 More R basics: lists, dates, etc.              "\U0~    NA "Hadle~
##  9 Sep 19 R programming basics: conditional statements   "\U0~     2 ""
## 10 Sep 21 R programming basics: loops, apply             "\U0~    NA ""
## # i 21 more rows
```

```
materials_table
```

```
## # A tibble: 2 x 4
##   Topic                                           Notes         HW    Reading
##   <chr>                                           <chr>         <lgl> <chr>
## 1 The file system and basic unix shell            "\U0001f4d9" NA    "Allesi~
## 2 Open Science, Makefile, and Ethics in data science "\U0001f4d9" NA    ""
```

Question 2

```
schedule_table$month <- str_extract(schedule_table$Date, "[A-Za-z]{3}")
schedule_table$day <- as.numeric(str_extract(schedule_table$Date, "\\d+"))
schedule_table
```

```
## # A tibble: 31 x 7
##    Date   Topic                                 Notes    HW Reading month   day
##    <chr>  <chr>                                 <chr> <int> <chr>   <chr> <dbl>
##  1 Aug 22 About the course                      "\U0~    NA "Leek ~ Aug      22
##  2 Aug 24 Data science project cycle            "\U0~    NA "Mason~ Aug      24
##  3 Aug 29 Introduction and install tools        "\U0~    NA "Coope~ Aug      29
##  4 Aug 31 Version control with Git              "\U0~    NA "Blisc~ Aug      31
##  5 Sep 05 Introduction to GitHub                "\U0~    NA ""      Sep       5
##  6 Sep 07 RStudio project and dynamic documents~ "\U0~     1 "Xie e~ Sep       7
##  7 Sep 12 R basics: data types, vectors, matrix~ "\U0~    NA ""      Sep      12
##  8 Sep 14 More R basics: lists, dates, etc.     "\U0~    NA "Hadle~ Sep      14
##  9 Sep 19 R programming basics: conditional sta~ "\U0~     2 ""      Sep      19
## 10 Sep 21 R programming basics: loops, apply    "\U0~    NA ""      Sep      21
## # i 21 more rows
```

Question 3

```
lecture_counts <- schedule_table %>%
  group_by(month) %>%
  summarise(lecture_count = n())
lecture_counts <- schedule_table %>%
  group_by(month) %>%
  summarise(lecture_count = n())
lecture_counts
```

```
## # A tibble: 5 x 2
##   month lecture_count
##   <chr>         <int>
## 1 Aug               4
## 2 Dec               1
## 3 Nov               9
## 4 Oct               9
## 5 Sep               8
```

Question 4

```
topic_words <- schedule_table %>%
  mutate(words = str_split(Topic, "\\s+")) %>%
  pull(words) %>%
  unlist()
word_freq <- table(topic_words)
sorted_word_freq <- sort(word_freq, decreasing = TRUE)
top_5_words <- head(names(sorted_word_freq), 5)
data.frame(Word = top_5_words, Frequency = sorted_word_freq[top_5_words])
```

```
##   Word Frequency.topic_words Frequency.Freq
## 1  and                   and              9
## 2    R                     R              8
## 3 data                  data              7
## 4 with                  with              5
## 5 (Dr.                  (Dr.              4
```