Sourcing html data with R

Contents

1 Task description

Parse course information from FH Technikum's CIS web site.

Steps:

- Explore html code
- Start parsing information for one course
- Create loop for getting several course infos

2 Required packages

```
library(tidyverse)
library(magrittr)
library(rvest)
```

3 One course page

Example: Accounting in BWI1bb

URL: https://cis.technikum-wien.at/addons/lvinfo/cis/view.php?lehrveranstaltung_id=34425&studiensemester_kurzbz=WS2018

Load one page:

```
page <- read_html("https://cis.technikum-wien.at/addons/lvinfo/cis/view.php?lehrveranstaltung_id=34425&s
```

Explore content:

```
page
```

```
## {html_document}
## <html>
## [1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-8 ...
## [2] <body>\n<h1>Lehrveranstaltungsinformationen</h1>Verfügbare Sprachen: <a h ...
```

```
html_nodes(page, "*")
```

```
## {xml_nodeset (79)}
##  [1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF- ...
##  [2] <meta http-equiv="Content-Type" content="text/html; charset=UTF-8">\n
##  [3] <title>LV-Informationen</title>\n
##  [4] <link rel="stylesheet" href="../../../skin/style.css.php" type="text/css ...
##  [5] <link rel="stylesheet" href="../../../skin/tablesort.css" type="text/css ...
##  [6] <link rel="stylesheet" href="../skin/lvinfo.css" type="text/css">\n
##  [7] <script type="text/javascript" src="../../../vendor/jquery/jqueryV1/jque ...
##  [8] <script type="text/javascript" src="../../../vendor/christianbach/tables ...
##  [9] <body>\n<h1>Lehrveranstaltungsinformationen</h1>Verfügbare Sprachen: <a  ...
## [10] <h1>Lehrveranstaltungsinformationen</h1>
## [11] <a href="view.php?lehrveranstaltung_id=34425&amp;studiensemester_kurzbz= ...
## [12] <a href="view.php?lehrveranstaltung_id=34425&amp;studiensemester_kurzbz= ...
## [13] <table class="tablesorter">\n<tr>\n<td>Lehrveranstaltung:</td>\n\t\t\t<t ...
## [14] <tr>\n<td>Lehrveranstaltung:</td>\n\t\t\t<td>Buchhaltung</td>\n\t\t</tr>\n
## [15] <td>Lehrveranstaltung:</td>
## [16] <td>Buchhaltung</td>
## [17] <tr>\n<td>Studiengang:</td>\n\t\t\t<td>BWI</td>\n\t\t</tr>\n
## [18] <td>Studiengang:</td>
## [19] <td>BWI</td>
## [20] <tr>\n<td>Semester:</td>\n\t\t\t<td>1</td>\n\t\t</tr>\n
## ...
```

3.1 Upper table

Idea:

- extract table using html_table()
- transpose table to get data column-wise

```
tmp <-
  page %>%
  html_nodes("table") %>%
  first() %>%
  html_table()
tmp
```

```
## # A tibble: 10 x 2
##    X1                    X2
##    <chr>                 <chr>
##  1 Lehrveranstaltung:    "Buchhaltung"
##  2 Studiengang:          "BWI"
##  3 Semester:             "1"
##  4 Studiensemester:      "WS2018"
##  5 Organisationsform:    "BB"
##  6 Lehrbeauftragte(r):   "FH-Prof. Ing. Dipl.-Ing. Mag. Dr. Daniel F. Leutgeb M~
##  7 Sprache:              "Deutsch"
##  8 ECTS:                 "3.00"
##  9 Incomingplätze:       "0"
## 10 Organisationseinheit: "Kompetenzfeld Wirtschaft-Recht\n\t\t\t \t\n\t\t\t\t(\~
```

Transpose table:

```
course <- tmp %>% select(X2) %>% t()
course
```

```
##    [,1]          [,2]  [,3] [,4]     [,5]
## X2 "Buchhaltung" "BWI" "1"  "WS2018" "BB"
##    [,6]                                                                  [,7]
## X2 "FH-Prof. Ing. Dipl.-Ing. Mag. Dr. Daniel F. Leutgeb MBA CMC CSE" "Deutsch"
##    [,8]   [,9]
## X2 "3.00" "0"
##    [,10]
## X2 "Kompetenzfeld Wirtschaft-Recht\n\t\t\t \t\n\t\t\t\t\t(\n\t\t\t\t\t\tLeitung: Mag. Dr. Thomas Wala MB
```

Fix column names:

```
colnames(course) <- t(tmp[,1] %>% mutate_at("X1", str_replace, ":", ""))
course <- as.tibble(course, .name_repair = make.names)
```

```
## Warning: `as.tibble()` was deprecated in tibble 2.0.0.
## Please use `as_tibble()` instead.
## The signature and semantics have changed, see `?as_tibble`.
```

Modify data types of numeric variables:

```
course %<>%
  mutate(
    Semester = parse_integer(Semester),
    ECTS = parse_double(ECTS),
    Incomingplätze = parse_integer(Incomingplätze)
  )
course
```

```
## # A tibble: 1 x 10
##   Lehrveranstaltung Studiengang Semester Studiensemester Organisationsform
##   <chr>             <chr>          <int> <chr>           <chr>
## 1 Buchhaltung       BWI                1 WS2018          BB
## # ... with 5 more variables: Lehrbeauftragte.r. <chr>, Sprache <chr>,
## #   ECTS <dbl>, Incomingplätze <int>, Organisationseinheit <chr>
```

3.2 Lower table

Approach:

- Select section via `lvinfo` class
- Headings: all `<h2>` tags
- Content: all tags with `lv-info` class
- challenge: some sections have bullet points

```
heads <- page %>%
  html_nodes(".lvinfo") %>%
  html_nodes("h2") %>%
  html_text()
heads
```

```
## [1] "Kurzbeschreibung"     "Methodik"             "Lernergebnisse"
## [4] "Lehrinhalte"          "Vorkenntnisse"        "Literatur"
## [7] "Leistungsbeurteilung" "Anwesenheit"          "Anmerkungen"
```

```
FUN <- function(i) {
  if (length(html_children(i)))
      paste0(html_text(html_children(html_children(i))), collapse = "\n")
  else
     html_text(i)
}

values <- page %>%
  html_nodes(".lvinfo_data") %>%
  map(FUN)

values
```

```
## [[1]]
## [1] "Im Rahmen dieser Lehrveranstaltung erwerben die Studierenden Kenntnisse über theoretische und r
##
## [[2]]
## [1] "Fragen, Übungen"
##
## [[3]]
## [1] "die Richtlinien des externen Rechnungswesens wiederzugeben\nGeschäftsfälle auf Konten zu verbuc
##
## [[4]]
## [1] "Elemente und Funktionen der Finanzbuchhaltung, rechtliche Vorschriften, Grundsätze ordnungsmäßi
##
## [[5]]
## [1] "Keine besonderen Vorkenntnisse erforderlich"
##
## [[6]]
## [1] "Bertl, Romuald / Deutsch-Goldoni, Eva / Hirschler, Klaus (2013): \"Buchhaltungs- und Bilanzieru
##
## [[7]]
## [1] "LV-Immanente Leistungsbeurteilung (100%)"
##
```

```
## [[8]]
## [1] "Es besteht Anwesenheitspflicht."
##
## [[9]]
## [1] "Arbeiten Sie bitte bereits vor der ersten LV das Skriptum durch.  "
```

create tibble and add to course:

```
names(values) <- heads
course <- cbind(course, as.tibble(values))
```

3.3 Combine to function

```
parse_one_course <-
  function(id = 34425, term = "WS2018")
  {
    ## get html
    url <- paste0("https://cis.technikum-wien.at/addons/lvinfo/cis/view.php?lehrveranstaltung_id=", id,
    page <- read_html(url)

    ## upper table
    tmp <-
      page %>%
      html_nodes("table") %>%
      first() %>%
      html_table()

    course <- tmp %>% select(X2) %>% t()

    #Diese Zeilen:
    #colnames(course) <- str_replace(tmp[,1], ":", "")
    #course <- as.tibble(course)

    #Durch diese ersetzt:
    colnames(course) <- t(tmp[,1] %>% mutate_at("X1", str_replace, ":", ""))
    course <- as.tibble(course, .name_repair = make.names)


    course %<>%
      mutate(
        Semester = parse_integer(Semester),
        ECTS = parse_double(ECTS),
        Incomingplätze = parse_integer(Incomingplätze)
      )

    ## lower table
    heads <- page %>%
      html_nodes(".lvinfo") %>%
      html_nodes("h2") %>%
      html_text()

    FUN <- function(i) {
```

```
      if (length(html_children(i)))
        paste0(html_text(html_children(html_children(i))),
               collapse = "\n")
      else
        html_text(i)
    }

    values <- page %>%
      html_nodes(".lvinfo_data") %>%
      map(FUN)

    names(values) <- heads
    course <- cbind(course, as.tibble(values))

    course
  }
```

4 Loop over several courses

```
parse_courses <-
function(ids = 1, term = "WS2018")
{
    lapply(ids, parse_one_course, term)

}

ids = c(34509, 34473, 34496)
result <-
  ids %>%
  parse_courses %>%
  bind_rows()
summary(result)
```

```
##  Lehrveranstaltung  Studiengang           Semester Studiensemester
##  Length:3           Length:3           Min.   :1   Length:3
##  Class :character   Class :character   1st Qu.:1   Class :character
##  Mode  :character   Mode  :character   Median :1   Mode  :character
##                                        Mean   :1
##                                        3rd Qu.:1
##                                        Max.   :1
##  Organisationsform  Lehrbeauftragte.r.   Sprache              ECTS
##  Length:3           Length:3           Length:3           Min.   :4.000
##  Class :character   Class :character   Class :character   1st Qu.:4.500
##  Mode  :character   Mode  :character   Mode  :character   Median :5.000
##                                                           Mean   :4.667
##                                                           3rd Qu.:5.000
##                                                           Max.   :5.000
##  Incomingplätze Organisationseinheit Kurzbeschreibung     Methodik
##  Min.   :0      Length:3             Length:3           Length:3
##  1st Qu.:0      Class :character     Class :character   Class :character
```

```
##  Median :0      Mode  :character     Mode  :character   Mode  :character
##  Mean   :0
##  3rd Qu.:0
##  Max.   :0
##  Lernergebnisse     Lehrinhalte        Vorkenntnisse       Literatur
##  Length:3           Length:3           Length:3           Length:3
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Leistungsbeurteilung Anwesenheit        Anmerkungen
##  Length:3             Length:3           Length:3
##  Class :character     Class :character   Class :character
##  Mode  :character     Mode  :character   Mode  :character
##
##
##
```