project-web-scrapping-website-1

Vishal

2025-02-14

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.4     ✓ tidyr     1.3.1
## ✓ purrr     1.0.4
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to be
come errors
```

```
library(rvest)
```

```
##
## Attaching package: 'rvest'
##
## The following object is masked from 'package:readr':
##
##     guess_encoding
```

```
library(dplyr)
library(httr)

url <- "https://collegedunia.com/university/25914-vellore-institute-of-technology-vit-univers
ity-vellore/reviews"

# Use a User-Agent to mimic a real browser
page <- read_html(GET(url, user_agent("Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/
537.36 (KHTML, like Gecko) Chrome/114.0.0.0 Safari/537.36")))
```

```
review_dates <- page %>% html_elements(xpath="//span[@class='jsx-3091098665' and not(ancesto
r::div[contains(@class, 'mb-1')])]") %>% html_text()
review_dates <- gsub("Reviewed on ", "", review_dates)
review_dates
```

```
##  [1] "Feb 13, 2025" "Feb 13, 2025" "Feb 12, 2025" "Feb 12, 2025" "Feb 12, 2025"
##  [6] "Feb 11, 2025" "Feb 10, 2025" "Feb 9, 2025"  "Feb 9, 2025"  "Feb 9, 2025"
```

```
Dept <- page %>% html_elements(xpath="//div[contains(@class, 'mb-1')]//a//span[@class='jsx-30
91098665']") %>% html_text()
Dept
```

```
##  [1] "B.Sc, Animation & Multimedia"
##  [2] "B.Sc, Animation & Multimedia"
##  [3] "B.Tech, Electrical And Electronics Engineering"
##  [4] "B.Tech, Computer Science and Engineering"
##  [5] "B.Tech + M.Tech, Software Engineering"
##  [6] "B.Tech, Computer Science and Engineering"
##  [7] "B.Tech, Electronics And Instrumentation Engineering"
##  [8] "B.Tech + M.Tech, Computer Science and Engineering + Data Science"
##  [9] "B.Tech, Mechanical Engineering"
## [10] "B.Tech + M.Tech, Computer Science & Engineering"
```

```
ratings <- page %>% html_elements(xpath="//span[contains(@class, 'f-16 font-weight-semi text-
dark-grey')]") %>% html_text()
print(ratings)
```

```
##  [1] "3.4" "3.6" "4.0" "5.0" "3.3" "3.9" "3.5" "3.9" "3.6" "4.0"
```

```
student_names <- page %>% html_elements(xpath="//span[contains(@class, 'font-weight-semi text
-primary-black')]") %>% html_text()

# Print extracted student names
print(student_names)
```

```
##  [1] "Mukesh Yashvanth"     "Saamya Kunhibi"       "Punya Oswal"
##  [4] "Soumyajit Ghosal"     "GDivya"               "Review guru"
##  [7] "Srikar Tandulwadikar" "Mohammed Shahid"      "Avinash"
## [10] "jyotinadh"
```

```
reviews <- page %>%
  html_elements(xpath="//ul[contains(@class, 'mt-2 mb-0 pl-4 fs-16 font-weight-normal text-gr
ay-10')]/li") %>%
  html_text()

length(reviews)
```

```
## [1] 55
```

```
last_page <- 300
```

```
review_date_all=c()
review_rating_all=c()
review_dept_all=c()
review_name_all=c()
review_review_all=c()
```

```
pagesequence=seq(1,last_page)
pagesequence
```

```
##   [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
##  [19]  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
##  [37]  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54
##  [55]  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72
##  [73]  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
##  [91]  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107 108
## [109] 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126
## [127] 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144
## [145] 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162
## [163] 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180
## [181] 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198
## [199] 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216
## [217] 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234
## [235] 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252
## [253] 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270
## [271] 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288
## [289] 289 290 291 292 293 294 295 296 297 298 299 300
```

```
# Function to ensure equal list lengths by padding with NA
lengthen <- function(vec, target_length) {
  if (length(vec) < target_length) {
    vec <- c(vec, rep(NA, target_length - length(vec)))  # Pad with NA
  }
  return(vec)
}

for (i in pagesequence) {
  page_url <- ifelse(i == 1, url, paste0(url, "/page-", i, "?sort=3"))

  # Handle Read Errors Gracefully
  page <- tryCatch({
    read_html(GET(page_url, user_agent("Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKi
t/537.36 (KHTML, like Gecko) Chrome/114.0.0.0 Safari/537.36")))
  }, error = function(e) return(NULL))  # Returns NULL if error occurs

  # Skip this iteration if page failed
  if (is.null(page)) next

  # Extract Data
  review_dates <- page %>% html_elements(xpath="//span[@class='jsx-3091098665' and not(ancest
or::div[contains(@class, 'mb-1')])]") %>% html_text()
  review_dates <- gsub("Reviewed on ", "", review_dates)

  student_names <- page %>% html_elements(xpath="//span[contains(@class, 'font-weight-semi te
xt-primary-black')]") %>% html_text()
  ratings <- page %>% html_elements(xpath="//span[contains(@class, 'f-16 font-weight-semi tex
t-dark-grey')]") %>% html_text()
  reviews <- page %>% html_elements(xpath="//ul[contains(@class, 'mt-2 mb-0 pl-4 fs-16 font-w
eight-normal text-gray-10')]/li") %>% html_text()
  Dept <- page %>% html_elements(xpath="//div[contains(@class, 'mb-1')]//a//span[@class='jsx-
3091098665']") %>% html_text()


  # Append Data to Lists
  review_date_all <- append(review_date_all, review_dates)
  review_rating_all <- append(review_rating_all,ratings)
  review_dept_all <- append(review_dept_all,Dept)
  review_name_all <- append(review_name_all,student_names)
  review_review_all <- append(review_review_all,reviews)
}
```

```
# Check lengths of extracted vectors
cat("Total Review Dates:", length(review_date_all), "\n")
```

```
## Total Review Dates: 3000
```

```
cat("Total Ratings:", length(review_rating_all), "\n")
```

```
## Total Ratings: 3000
```

```
cat("Total Departments:", length(review_dept_all), "\n")
```

```
## Total Departments: 2750
```

```
cat("Total Student Names:", length(review_name_all), "\n")
```

```
## Total Student Names: 3000
```

```
cat("Total Reviews (Likes Section):", length(review_review_all), "\n")
```

```
## Total Reviews (Likes Section): 9758
```

```
# Flatten the lists into single vectors
review_date_all <- unlist(review_date_all)
review_rating_all <- unlist(review_rating_all)
review_dept_all <- unlist(review_dept_all)
review_name_all <- unlist(review_name_all)
review_review_all <- unlist(review_review_all)

length(review_date_all)
```

```
## [1] 3000
```

```
length(review_rating_all)
```

```
## [1] 3000
```

```
length(review_dept_all)
```

```
## [1] 2750
```

```
length(review_name_all)
```

```
## [1] 3000
```

```
length(review_review_all)
```

```
## [1] 9758
```

```
print(review_date_all[1:2])
```

```
## [1] "Feb 13, 2025" "Feb 13, 2025"
```

```
print(review_rating_all[1:2])
```

```
## [1] "3.4" "3.6"
```

```
print(review_dept_all[1:2])
```

```
## [1] "B.Sc, Animation & Multimedia" "B.Sc, Animation & Multimedia"
```

```
print(review_name_all[1:2])
```

```
## [1] "Mukesh Yashvanth" "Saamya Kunhibi"
```

```
print(review_review_all[1:2])
```

```
## [1] "One of my senior got placed in Couch Base for 34LPA CTC another got placed in Unileve
r for 42LPA but this was for CSE and related branches while few of my seniors got decent pack
age from mechanical branch too 8-12L in companies like Mercedes a Euler  ~ Anonymous, B.Tech"
## [2] "FFCS. A study who is facing difficulty to study more, can take less number of credits
and it will reduce the burden of the students. And the student can also Complete the entire c
ourse in 4 years instead of 5years  by taking maximum no of credits.  ~ Supriya R Patil, B.Sc
+ M.Sc"
```

```
combine_reviews_fixed <- function(reviews, group_size = 4) {
  if (length(reviews) == 0) {
    return(character(0))  # Return empty character vector if no reviews
  }

  sapply(seq(1, length(reviews), by = group_size), function(i) {
    paste(reviews[i:min(i+group_size-1, length(reviews))], collapse = " ")
  })
}
k=round(length(review_review_all)/length(review_name_all))
# Store the rearranged reviews
review_review_all_rearranged <- combine_reviews_fixed(review_review_all,k)

# Check new length
length(review_review_all_rearranged)
```

```
## [1] 3253
```

```
print(review_review_all_rearranged[1:2])
```

```
## [1] "One of my senior got placed in Couch Base for 34LPA CTC another got placed in Unileve
r for 42LPA but this was for CSE and related branches while few of my seniors got decent pack
age from mechanical branch too 8-12L in companies like Mercedes a Euler  ~ Anonymous, B.Tech
FFCS. A study who is facing difficulty to study more, can take less number of credits and it
will reduce the burden of the students. And the student can also Complete the entire course i
n 4 years instead of 5years  by taking maximum no of credits.  ~ Supriya R Patil, B.Sc + M.Sc
Campus Life in Vit is simply magnificent.All types of facilities are provided inside the camp
us including sport complex and restaurants.The girls to boys ratio in Mechanical engineering
is surprisinly high of about 1:7.  ~ Mithilesh Angal, B.Tech"
## [2] "I don’t like how the student and faculty ratio is 60-40  ~ Praveen Kumar, BCA Lot of
exams yearly 3 exams in a semester 2-3 quiz and 1-2 DA  ~ Kartik Navnath Narare, B.Tech Water
is really bad. 750+ TDS. It makes you go bald. Really.  ~ Piyush Prajapati, B.Des"
```

```
length(review_date_all)
```

```
## [1] 3000
```

```
length(review_rating_all)
```

```
## [1] 3000
```

```
length(review_dept_all)
```

```
## [1] 2750
```

```
length(review_name_all)
```

```
## [1] 3000
```

```
length(review_review_all_rearranged)
```

```
## [1] 3253
```

```
# Set target length
target_length <- max(length(review_date_all),
length(review_rating_all),
length(review_dept_all),
length(review_name_all),
length(review_review_all_rearranged))

# Function to extend vectors to target length by filling with NA if needed
extend_to_length <- function(vec, target_length) {
  length(vec) <- target_length  # This automatically fills missing values with NA
  return(vec)
}

# Extend all columns to 3880 rows
review_date_all_trimmed <- extend_to_length(review_date_all, target_length)
review_rating_all_trimmed <- extend_to_length(review_rating_all, target_length)
review_dept_all_trimmed <- extend_to_length(review_dept_all, target_length)
review_name_all_trimmed <- extend_to_length(review_name_all, target_length)
review_review_all_trimmed <- extend_to_length(review_review_all_rearranged, target_length)

# Check if all columns have 3880 rows
length(review_date_all_trimmed) == target_length  # TRUE
```

```
## [1] TRUE
```

```
length(review_rating_all_trimmed) == target_length  # TRUE
```

```
## [1] TRUE
```

```
length(review_dept_all_trimmed) == target_length  # TRUE
```

```
## [1] TRUE
```

```
length(review_name_all_trimmed) == target_length  # TRUE
```

```
## [1] TRUE
```

```
length(review_review_all_trimmed) == target_length  # TRUE
```

```
## [1] TRUE
```

```
reviews_df_final <- data.frame(
  Name = review_name_all_trimmed,
  Rating = review_rating_all_trimmed,
  Department = review_dept_all_trimmed,
  Review_Date = review_date_all_trimmed,
  Review_Text = review_review_all_trimmed,
  stringsAsFactors = FALSE
)

# Check results
head(reviews_df_final)
```

```
##               Name Rating                                      Department
## 1 Mukesh Yashvanth    3.4                   B.Sc, Animation & Multimedia
## 2   Saamya Kunhibi    3.6                   B.Sc, Animation & Multimedia
## 3      Punya Oswal    4.0 B.Tech, Electrical And Electronics Engineering
## 4 Soumyajit Ghosal    5.0       B.Tech, Computer Science and Engineering
## 5           GDivya    3.3          B.Tech + M.Tech, Software Engineering
## 6      Review guru    3.9       B.Tech, Computer Science and Engineering
##    Review_Date
## 1 Feb 13, 2025
## 2 Feb 13, 2025
## 3 Feb 12, 2025
## 4 Feb 12, 2025
## 5 Feb 12, 2025
## 6 Feb 11, 2025
##
Review_Text
## 1 One of my senior got placed in Couch Base for 34LPA CTC another got placed in Unilever f
or 42LPA but this was for CSE and related branches while few of my seniors got decent package
from mechanical branch too 8-12L in companies like Mercedes a Euler  ~ Anonymous, B.Tech FFC
S. A study who is facing difficulty to study more, can take less number of credits and it wil
l reduce the burden of the students. And the student can also Complete the entire course in 4
years instead of 5years  by taking maximum no of credits.  ~ Supriya R Patil, B.Sc + M.Sc Cam
pus Life in Vit is simply magnificent.All types of facilities are provided inside the campus
including sport complex and restaurants.The girls to boys ratio in Mechanical engineering is
surprisinly high of about 1:7.  ~ Mithilesh Angal, B.Tech
## 2
I don’t like how the student and faculty ratio is 60-40  ~ Praveen Kumar, BCA Lot of exams ye
arly 3 exams in a semester 2-3 quiz and 1-2 DA  ~ Kartik Navnath Narare, B.Tech Water is real
ly bad. 750+ TDS. It makes you go bald. Really.  ~ Piyush Prajapati, B.Des
## 3
The university is vast to have best experience  Best faculty crew providing best source of ed
ucation  Security guards of universities are bad.
## 4
Transportation is very bad making it difficult for students  The infrastructure of the univer
sity gives you a perfect learning environment  The faculties are highly qualified to train th
e students
## 5
The library is the best part out of the university  Too many rules and regulations to be foll
owed  Arrear fees is 6000 per subject which is too much
## 6
Attendance minimum criteria is seventify percentage  The College has many hostel buildings an
d enough space  The nature of college and greenery is the best part
```

```
dim(reviews_df_final)
```

```
## [1] 3253    5
```

```
write.csv(reviews_df_final, "D:\\Prog for DS\\Vit_reviews_data2.csv", row.names = FALSE)
```

``