project-web-scrapping-website-2

Vishal

2025-02-14

```
library(rvest)
library(httr)
url="https://www.vidyavision.com/college-reviews/vellore-institute-of-technology-(vit)"
page=read_html(url)
```

```
names <- page %>%
  html_elements("div.post_content") %>%  # Select the div with class "post_content"
  html_text(trim = TRUE)                 # Extract the text and trim whitespace

names
```

```
##  [1] "Name: Mahesh | Batch of 2026 | Rating 4.5Course: B.E / B.Tech (Computer Science and
Engineering)"
##  [2] "Anonymous | Rating  4.5"
##  [3] "Name: Atharv Porwal | Batch of 2028 | Rating 3.5Course: B.E / B.Tech (Electronics an
d Communication Engineering)"
##  [4] "Anonymous | Rating  3.5"
##  [5] "Name: Arpit Rajpurohit | Batch of 2027 | Rating 3.5Course: B.E / B.Tech (Computer Sc
ience and Engineering)"
##  [6] "Anonymous | Rating  3.5"
##  [7] "Name: Meda Venkata Naga Hemanth Kumar | Batch of 2026 | Rating 5Course: B.E / B.Tech
(Computer Science and Engineering)"
##  [8] "Anonymous | Rating  5"
##  [9] "Name: Raju varre | Batch of 2027 | Rating 5Course: B.E / B.Tech (Computer Science an
d Engineering)"
## [10] "Anonymous | Rating  5"
## [11] "Name: RITIK ROUSHAN RANA | Batch of 2027 | Rating 4.5Course: B.E / B.Tech (Computer
Science and Engineering)"
## [12] "Anonymous | Rating  4.5"
## [13] "Name: Adtya Tripathi | Batch of 2027 | Rating 4Course: B.E / B.Tech (CSE - Specializ
ation in Artificial Intelligence and Machine Learning)"
## [14] "Anonymous | Rating  4"
## [15] "Name: Deepika Reddy | Batch of 2025 | Rating 3.5Course: B.E / B.Tech (Computer Scien
ce and Engineering)"
## [16] "Anonymous | Rating  3.5"
## [17] "Name: Tanmay Admuthe | Batch of 2025 | Rating 3.5Course: B.E / B.Tech (Electronics a
nd Computer Engineering)"
## [18] "Anonymous | Rating  3.5"
## [19] "Name: Shyam.V | Batch of 2026 | Rating 3.5Course: Integrated (CSE - Data Science)"
## [20] "Anonymous | Rating  3.5"
## [21] "Name: Aashish Lalwani | Batch of 2022 | Rating 3.5Course: B.E / B.Tech (Electronics
and Communication Engineering)"
## [22] "Anonymous | Rating  3.5"
## [23] "Name: Vinayak  | Batch of 2025 | Rating 4Course: B.E / B.Tech (Computer Science and
Engineering)"
## [24] "Anonymous | Rating  4"
## [25] "Name: Aayush Kudalkar | Batch of 2025 | Rating 3.5Course: B.E / B.Tech (Computer Sci
ence and Engineering)"
## [26] "Anonymous | Rating  3.5"
## [27] "Name: Vikrant pandey | Batch of 2016 | Rating 5Course: B.E / B.Tech (Electrical and
Electronics Engineering)"
## [28] "Anonymous | Rating  5"
## [29] "Name: talha | Batch of 2020 | Rating 3Course:"
## [30] "Anonymous | Rating  3"
```

```r
library(stringr)
names_extracted <- str_extract(names, "(?<=Name: )[^|]+")

# Print extracted names
print(names_extracted)
```

```
##  [1] "Mahesh "                           NA
##  [3] "Atharv Porwal "                    NA
##  [5] "Arpit Rajpurohit "                 NA
##  [7] "Meda Venkata Naga Hemanth Kumar " NA
##  [9] "Raju varre "                       NA
## [11] "RITIK ROUSHAN RANA "               NA
## [13] "Adtya Tripathi "                   NA
## [15] "Deepika Reddy "                    NA
## [17] "Tanmay Admuthe "                   NA
## [19] "Shyam.V "                          NA
## [21] "Aashish Lalwani "                  NA
## [23] "Vinayak  "                         NA
## [25] "Aayush Kudalkar "                  NA
## [27] "Vikrant pandey "                   NA
## [29] "talha "                            NA
```

```
# Extract the full text from "post_content" div
post_content <- page %>%
  html_elements("div.post_content") %>%  # Select the div with class "post_content"
  html_text(trim = TRUE)

# Extract course names from the text using regex
course_names <- str_extract(post_content, "(?<=Course: )[^\n]+")

# Print extracted course names
print(course_names)
```

```
##  [1] "B.E / B.Tech (Computer Science and Engineering)"
##  [2] NA
##  [3] "B.E / B.Tech (Electronics and Communication Engineering)"
##  [4] NA
##  [5] "B.E / B.Tech (Computer Science and Engineering)"
##  [6] NA
##  [7] "B.E / B.Tech (Computer Science and Engineering)"
##  [8] NA
##  [9] "B.E / B.Tech (Computer Science and Engineering)"
## [10] NA
## [11] "B.E / B.Tech (Computer Science and Engineering)"
## [12] NA
## [13] "B.E / B.Tech (CSE - Specialization in Artificial Intelligence and Machine Learning)"
## [14] NA
## [15] "B.E / B.Tech (Computer Science and Engineering)"
## [16] NA
## [17] "B.E / B.Tech (Electronics and Computer Engineering)"
## [18] NA
## [19] "Integrated (CSE - Data Science)"
## [20] NA
## [21] "B.E / B.Tech (Electronics and Communication Engineering)"
## [22] NA
## [23] "B.E / B.Tech (Computer Science and Engineering)"
## [24] NA
## [25] "B.E / B.Tech (Computer Science and Engineering)"
## [26] NA
## [27] "B.E / B.Tech (Electrical and Electronics Engineering)"
## [28] NA
## [29] NA
## [30] NA
```

```
# Extract ratings from <span> tag with class "badge"
ratings <- page %>%
  html_elements("span.badge") %>%  # Select the span with class "badge"
  html_text(trim = TRUE)

# Print extracted ratings
print(ratings)
```

```
##  [1] "4.5" "4.5" "3.5" "3.5" "3.5" "3.5" "5"   "5"   "5"   "5"   "4.5" "4.5"
## [13] "4"   "4"   "3.5" "3.5" "3.5" "3.5" "3.5" "3.5" "3.5" "3.5" "4"   "4"
## [25] "3.5" "3.5" "5"   "5"   "3"   "3"
```

```
dates_raw <- page %>%
  html_elements("div.pull-right") %>%
  html_text(trim = TRUE)

# Remove "Posted On: " from the extracted text
dates_clean <- str_replace(dates_raw, "Posted On:", "")

# Print extracted dates
print(dates_clean)
```

```
##  [1] " 18-Dec-2024" " 11-Dec-2024" " 28-Nov-2024" " 17-Aug-2024" " 23-Jul-2024"
##  [6] " 20-May-2024" " 09-Apr-2024" " 17-Dec-2023" " 20-Apr-2023" " 20-Apr-2022"
## [11] " 26-Feb-2022" " 16-Feb-2022" " 07-Jan-2022" " 30-Nov-2021" " 07-Oct-2021"
```

```
reviews <- page %>%
  html_elements("div.margBtm10") %>%  # Select div with class "margBtm10"
  html_text(trim = TRUE)              # Extract and clean text

# Print extracted reviews
print(reviews[1:2])
```

```
## [1] "Vellore Institute of Technology    Vellore, Tamil Nadu\r\n\t\t\t\t\t\t\t\t    \"Vit v
ellore genuine review ( good academics but mass crowd)\" \r\n\t\t\t\t\t\t\t\t    \r\n\t\t\t\t
\t\t\t\t\t    Name: Mahesh | Batch of 2026 | Rating 4.5Course: B.E / B.Tech (Computer Science
and Engineering)\r\n\t\t\t\t\t\t\t\t    \r\n\t\t\t\t\t\t\t\t    \r\n\t\t\t\t\t\t\t\t\t    Ano
nymous | Rating  4.5 \r\n\t\t\t\t\t\t\t\t    \r\n\t\t\t\t\t\t\t    \r\n\t\t\t\t\t\t\t    \r\n
\t\t\t\t\t\t\t\t    \r\n\t\t\t\t\t\t\t\t\t    Doesn't Recommend this college  \r\n\t\t\t\t\t
\t\t\t    \r\n\t\t\t\t\t\t\t\t    \r\n\t\t\t\t\t\t\t\t\t    Recommends this college"
## [2] "Vit vellore has both positive and negatives, coming to positives... Vit vellore is ca
mpus where you can pursue any degree you want except medical.The environment in and around th
e campus is so beautiful and peaceful, campus is flooded with large trees. Vit has its own pl
acement cell in the campus. More than 250+ companies visit the campus to recruit the final ye
ar students. Previous year (2023) the Highest package was grabbed by 2 students from CSE with
1.2 crore package annually from MOTARQ company. This year ongoing final year students grabbed
the highest package of 1 crore from the same company. Pre final years are getting internship
offers from many companies with stipends. Coming to hostels and mess ... There are two types
of hostels for both men and women...AC and NONAC .. there are like more than 10-15 blocks of
hostels only for men like A, B ,C .... Blocks .. coming to mess of 3 types , special, veg , n
on veg .. and also paid mess from private caterings. Dhobi is Available in every block.There
are various types of sports facilities like tennis , cricket, volleyball, basketball..and the
re's also a stadium inside men's hostel. There are 3 Gyms and 2 gyms for women. The placement
s in VIT are quite good as far I know 90% of students gets placed regardless of their dream c
ompany but they will place. There are terms like super dream offers and dream offers for the
students provided by academics. \r\n\r\n  \r\n                                \r\n
More..."
```

```
select_odd_values <- function(values) {
  odd_values <- values[seq(1, min(length(values), 29), by = 2)]  # Select odd indices
  limited_values <- head(odd_values, 15)  # Store only first 15 values
  return(limited_values)
}

select_even_values <- function(values) {
  even_values <- values[seq(0, min(length(values), 30), by = 2)]  # Select odd indices
  limited_values <- head(even_values, 15)  # Store only first 15 values
  return(limited_values)
}

names_selected  <- select_odd_values(names_extracted)
ratings_selected <- select_odd_values(ratings)
depts_selected <- select_odd_values(course_names)
reviews_selected <- select_even_values(reviews)
```

```
# Print the length of each selected vector
cat("Total Names Selected:", length(names_selected), "\n")
```

```
## Total Names Selected: 15
```

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

```
## Total Ratings Selected: 15
```

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

```
## Total Departments Selected: 15
```

```
cat("Total Reviews Selected:", length(reviews_selected), "\n")
```

```
## Total Reviews Selected: 15
```

```
cat("Total Dates:", length(dates_clean), "\n\n")
```

```
## Total Dates: 15
```

```
# Print two sample values from each vector
cat(names_selected[1:2], "\n")
```

```
## Mahesh  Atharv Porwal
```

```
cat(ratings_selected[1:2], "\n")
```

```
## 4.5 3.5
```

```
cat(depts_selected[1:2], "\n")
```

```
## B.E / B.Tech (Computer Science and Engineering) B.E / B.Tech (Electronics and Communicatio
n Engineering)
```

```
cat(reviews_selected[1:2], "\n")
```

```
## Vit vellore has both positive and negatives, coming to positives... Vit vellore is campus
where you can pursue any degree you want except medical.The environment in and around the cam
pus is so beautiful and peaceful, campus is flooded with large trees. Vit has its own placeme
nt cell in the campus. More than 250+ companies visit the campus to recruit the final year st
udents. Previous year (2023) the Highest package was grabbed by 2 students from CSE with 1.2
crore package annually from MOTARQ company. This year ongoing final year students grabbed the
highest package of 1 crore from the same company. Pre final years are getting internship offe
rs from many companies with stipends. Coming to hostels and mess ... There are two types of h
ostels for both men and women...AC and NONAC .. there are like more than 10-15 blocks of host
els only for men like A, B ,C .... Blocks .. coming to mess of 3 types , special, veg , non v
eg .. and also paid mess from private caterings. Dhobi is Available in every block.There are
various types of sports facilities like tennis , cricket, volleyball, basketball..and there's
also a stadium inside men's hostel. There are 3 Gyms and 2 gyms for women. The placements in
VIT are quite good as far I know 90% of students gets placed regardless of their dream compan
y but they will place. There are terms like super dream offers and dream offers for the stude
nts provided by academics.
##
##
##
##                                 More... VIT is a good university with great diversity all
over india and all teachers are doctorate with good experience in teaching, the teaching meth
odology is updated with update syllabus the con is the high student intake of almost 5 thousa
nd students including all courses.
##
##                                 More...
```

```
cat(dates_clean[1:2], "\n")
```

```
##  18-Dec-2024  11-Dec-2024
```

```
# Create a dataframe with the selected values
reviews_df <- data.frame(
  Name = names_selected,
  Rating = ratings_selected,
  Department = depts_selected,
  Review_Date = dates_clean,
  Review_Text = reviews_selected,

  stringsAsFactors = FALSE  # Prevent conversion to factors
)

# Print the first two rows of the dataframe
head(reviews_df[1:2, ])
```

```
##             Name Rating
## 1        Mahesh     4.5
## 2 Atharv Porwal     3.5
##                                                  Department  Review_Date
## 1          B.E / B.Tech (Computer Science and Engineering)  18-Dec-2024
## 2 B.E / B.Tech (Electronics and Communication Engineering)  11-Dec-2024
##
Review_Text
## 1 Vit vellore has both positive and negatives, coming to positives... Vit vellore is campu
s where you can pursue any degree you want except medical.The environment in and around the c
ampus is so beautiful and peaceful, campus is flooded with large trees. Vit has its own place
ment cell in the campus. More than 250+ companies visit the campus to recruit the final year 
students. Previous year (2023) the Highest package was grabbed by 2 students from CSE with 1.
2 crore package annually from MOTARQ company. This year ongoing final year students grabbed t
he highest package of 1 crore from the same company. Pre final years are getting internship o
ffers from many companies with stipends. Coming to hostels and mess ... There are two types o
f hostels for both men and women...AC and NONAC .. there are like more than 10-15 blocks of h
ostels only for men like A, B ,C .... Blocks .. coming to mess of 3 types , special, veg , no
n veg .. and also paid mess from private caterings. Dhobi is Available in every block.There a
re various types of sports facilities like tennis , cricket, volleyball, basketball..and ther
e's also a stadium inside men's hostel. There are 3 Gyms and 2 gyms for women. The placements
in VIT are quite good as far I know 90% of students gets placed regardless of their dream com
pany but they will place. There are terms like super dream offers and dream offers for the st
udents provided by academics. \r\n\r\n  \r\n                                \r\n
More...
## 2
VIT is a good university with great diversity all over india and all teachers are doctorate w
ith good experience in teaching, the teaching methodology is updated with update syllabus the
con is the high student intake of almost 5 thousand students including all courses.  \r\n
\r\n                                More...
```

```
dim(reviews_df)
```

```
## [1] 15  5
```

```
df <- read.csv("D:\\Prog for DS\\Vit_reviews_data2.csv")
```

```
head(df)
```

```
##                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
```

```
# Print dimensions before merging
cat("Before merging: Rows =", nrow(df), "Columns =", ncol(df), "\n")
```

```
## Before merging: Rows = 3253 Columns = 5
```

```
# Combine the existing data with the new data
df_combined <- rbind(df, reviews_df)

# Print dimensions after merging
cat("After merging: Rows =", nrow(df_combined), "Columns =", ncol(df_combined), "\n")
```

```
## After merging: Rows = 3268 Columns = 5
```

```
head(df_combined)
```

```
##               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 B.Tech, Electrical And Electronics Engineering
## 4 Soumyajit Ghosal      5        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
```

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