VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering

CSE3046-Programming for Data Science

Process: Obtain Data from Various Resources

Assignment-1

MyAnimeList-Obtaining Data with Web Scrapping using R

NAME :D VASANTH KUMAR

REG NO: 19BDS0083

Submitted Date: 28/08/201

Course Instructor: Dr. Anthoniraj A

DESCRIPTION:

Web scraping:

Web scraping is a process of using automated bots to crawl through the internet and extract data. The bots collect information by first breaking down the targeted site to its most basic form, HTML text, then scan through to gather data according to some preset parameters. After that, the collected data is delivered in CSV or Excel format, so it is readable for whoever wants to use it. Web scrapers are among the most efficient methods you can employ.

The main languages used to build web pages are called Hypertext Markup Language (HTML), Cascasing Style Sheets (CSS) and Javascript. HTML gives a web page its actual structure and content. CSS gives a web page its style and look, including details like fonts and colors. Javascript gives a webpage functionality.

In this, Data is collected from a website where more sensitive information's and tons of information's are shared each hour dynamically. Using some useful inbuilt libraries like rvest we are eligible to retrieve resourceful data and store them in proper format .The website which is used in this assignment is a dynamic website where people share their views and thoughts on the resources(anime) found on the website. We acquire their information (particularly used in the website to depict themselves) using web scraping method.

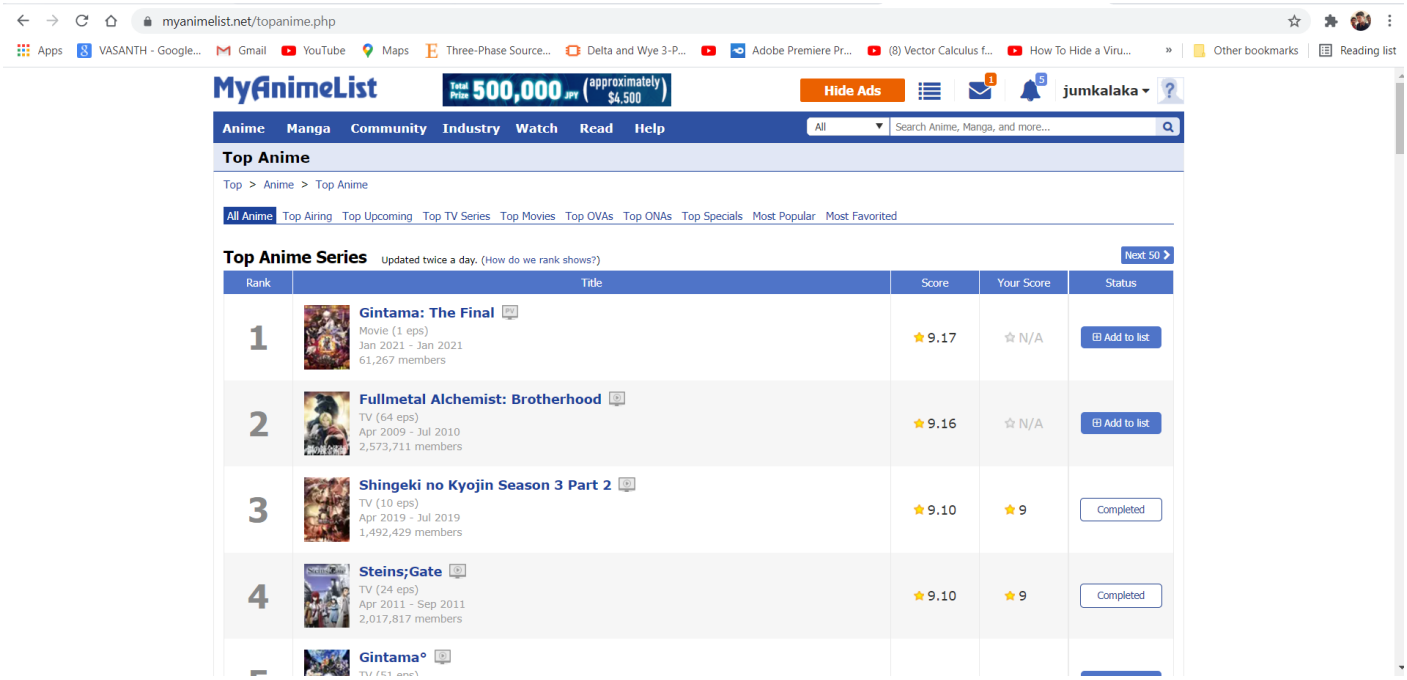
The main objective of this Data collection is to understand the working of web scrapping and how information's are circulated among such vast network. Also to gain more knowledge on how to gather information that is used in further enhancement of future knowledge.

DATA SOURCE DETAILS:

MyAnimeList.net is the data source used in this data collection assignment **MyAnimeList**.net is the world's largest anime and manga database and community created by an anime fan, for anime fans. We provide users with a quick and no-hassle way to catalog their anime and manga collections, as well as a platform to communicate with like minded fans, and keep up to date with important industry news.

Each month it display's 150 million page views to over 12 million unique visitors.

URL: https://myanimelist.net/



DATA RETRIVEL PROCESS:

LIBRARIES USED:

LIBRARY(RVEST):

The **rvest** library is a library that lets users easily scrape ("harvest") data from web pages.
In order to use the **rvest** library, we first need to install it and import it with the library() function.

LIBRARY(DPLYR):

The **dplyr** package in R is a structure of data manipulation that provides a uniform set of verbs, helping to resolve the most frequent data manipulation hurdles.

LIBRARY("WRITEXL):

This library is used to convert Dataframe objects to xlsx file.

KEYWORDS:

- DATA FRAME
- READ_HTML
- SCRAPING

STEPS INVOLVED:

- First of all in order to collect data from a webpage, the data must be dynamic in order to observe gradual changes in data and helps in studying the data over a period of time.
- First loading two libraries in Rstudio, they are
  Library(rvest)
  Library(dplyr)

- Selecting a webpage where data is shared dynamically. In **MyAnimeList** the reviews are the dynamic value. People share their comments, username, friend request and rating of an anime daily in order to promote or de-promote an anime's market.

- Secondly, installing necessary libraies helps in proper run of the code to collect data in R. The above mentioned libraries are necessary for few inbuilt functions to run in the code

- Install the mentioned libraries

- Create a variable which store the URL of the website where the dynamic data is available

- Then the URL read as data frame objects using **read_html** and stored in another variable

- Set of variables are created to represent the data in column wise. Create several rows to scrap particular details from the webpage.

- Essentially CSS tags are needed to scrap details from the website

- There are tools like scanner gadget, anypicker to get the css of the page which reduces the work in searching particular CSS tag in the inspect element of the html page of the data.

- Then the Data is columned using data frame where it is the most common Structured API and simply represents a table of data with rows and columns.

- Using writexl package the obtained data frame set in converted into xlsx or csv file .

NOTE:

Four reasons why you should be using pipes(%>%) in R:

- You'll structure the sequence of your data operations from left to right, as apposed to from inside and out;
- You'll avoid nested function calls;
- You'll minimize the need for local variables and function definitions; And
- You'll make it easy to add steps anywhere in the sequence of operations.
-

R Script for Obtaining data:

19BDS0083_anime_review.R:(Script file)

```
library(rvest)

library(dplyr)

require(rvest)

#install.packages("writexl")

library("writexl")
```

```
#link="https://myanimelist.net/reviews.php?t=anime"

anime_review=data.frame()

for(page_result in seq(from =1,to=15,by=1)){

  link=paste0("https://myanimelist.net/reviews.php?t=anime&p=",page_result)

  page=read_html(link)

  anime_name=page %>% html_nodes(".hoverinfo_trigger") %>% html_text()

  user_id=page %>% html_nodes("td > a") %>% html_text()

  overall_rating=page %>% html_nodes(".mb8 .spaceit+ div") %>% html_text()

  review_date=page %>% html_nodes(".mb8 div:nth-child(1)") %>% html_text()


anime_review=rbind(anime_review,data.frame(anime_name,user_id,overall_rati
ng,review_date))

}

View(anime_review)

write_xlsx(anime_review,"V:/vit/sem5/rvest/19BDS0083_anime_review_AUG18_
28.xlsx")

write.csv(anime_review,"V:/vit/sem5/rvest/19BDS0083_anime_review_AUG18_2
8.csv")
```

Rscript for task schedule:

```
install.packages("taskscheduleR")

library(taskscheduleR)

#install.packages('miniUI')

#install.packages('shiny')

#install.packages('shinyFiles')

taskscheduler_create(

  taskname = "r_web_scraping_anime_1",

  rscript="V:/vit/sem5/rvest/anime_review_19BDS0083.r",

  schedule="HOURLY",

  starttime=format(Sys.time() +62,"%H:%M"),

  Rexe = file.path(Sys.getenv("R_HOME"),"bin","Rscript.exe")

)

taskscheduler_stop("r_web_scraping_anime")

taskscheduler_delete("r_web_scraping_anime")
```

DATA SET DESCRIPTION:

The Data set of MyAnimeList consists of four columns. Each column represent different data acquired from same webpage with different information in it.

The column **anime_name** consists of the title of the anime on which reviews are being published in the website. Users mention the title of the anime and share their opinion on the anime that is mentioned. Providing reviews for such titled anime enhances popularity of the title or some bad reviews notices the quality of the anime and makes user know about it.

The column **user_id** is important data set among all other columns. In this columns the user_id of the user using this website is mentioned along with the title and review of an anime. So this information helps in identifying few people and approach them to acquire different opinions regarding the reviewed anime. This enhances interaction among the webusers of MyAnimeList.

The column **overall_rating** tells about the quality and strong base for an anime to become popular. this rating is calculated out of 10 which is the best score for depict that the anime is good at all conditions. Users provide multiple rating depending on their personal opinion and this rating is averaged and provides overall rating for an anime. Users prefer seeing rating of an anime to watch it which is also important data.

The final column **review_date** reperesents the date when the review of particular anime is published.This helps us to find number of reviews is being published per day.

SAMPLE DATA SET:

SAMPLE CODE:

anime_review_AUG23.R | scheduler.r

Source on Save | Run | Source

```r
library(rvest)
library(dplyr)

require(rvest)
#install.packages("writexl")
library("writexl")

#link="https://myanimelist.net/reviews.php?t=anime"
anime_review=data.frame()
for(page_result in seq(from =1,to=15,by=1)){
  link=paste0("https://myanimelist.net/reviews.php?t=anime&p=",page_result)
  page=read_html(link)

  anime_name=page %>% html_nodes(".hoverinfo_trigger") %>% html_text()
  user_id=page %>% html_nodes("td > a") %>% html_text()
  overall_rating=page %>% html_nodes(".mb8 .spaceit+ div") %>% html_text()
  review_date=page %>% html_nodes(".mb8 div:nth-child(1)") %>% html_text()

  anime_review=rbind(anime_review,data.frame(anime_name,user_id,overall_rating,review_date))
}

view(anime_review)

write_xlsx(anime_review,"V:/vit/sem5/rvest/19BDS0083_anime_review_AUG18_28.xlsx")
write.csv(anime_review,"V:/vit/sem5/rvest/19BDS0083_anime_review_AUG18_28.csv")
```

DATA SET:

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

anime_review | anime_review_AUG23.R | scheduler.r

Filter

	anime_name	user_id	overall_rating	review_date
1	Shigatsu wa Kimi no Uso	Nibir_Kujo	Overall Rating: 8	Aug 25, 2021
2	Boku wa Tomodachi ga Sukunai Next	_hsiW	Overall Rating: 10	Aug 25, 2021
3	Tonari no Kaibutsu-kun	titogonzalez	Overall Rating: 6	Aug 25, 2021
4	Meikyuu Black Company	Circle-kun	Overall Rating: 8	Aug 25, 2021
5	Hai to Gensou no Grimgar	NewStarLabs	Overall Rating: 10	Aug 25, 2021
6	Boku no Hero Academia	AniManga_Boy07	Overall Rating: 7	Aug 25, 2021
7	Kono Sekai no Katasumi ni	Samimito	Overall Rating: 8	Aug 25, 2021
8	School Days	jossreio	Overall Rating: 10	Aug 25, 2021
9	Tsuki ga Michibiku Isekai Douchuu	olandir	Overall Rating: 7	Aug 25, 2021
10	Re:Zero kara Hajimeru Isekai Seikatsu 2nd Season Part 2	oknem	Overall Rating: 7	Aug 25, 2021
11	Re:Zero kara Hajimeru Isekai Seikatsu 2nd Season	oknem	Overall Rating: 9	Aug 25, 2021
12	Re:Zero kara Hajimeru Isekai Seikatsu	oknem	Overall Rating: 8	Aug 25, 2021
13	86	oknem	Overall Rating: 9	Aug 25, 2021
14	Black Lagoon	FakeAccount1	Overall Rating: 6	Aug 25, 2021
15	Boruto: Naruto Next Generations	BosniaMan	Overall Rating: 4	Aug 25, 2021
16	Shuumatsu no Walküre	Toshito123	Overall Rating: 10	Aug 25, 2021
17	Ergo Proxy	DeltaWDunn	Overall Rating: 7	Aug 25, 2021
18	Watashi ga Motete Dousunda	giraffenanime	Overall Rating: 6	Aug 25, 2021
19	Yubisaki kara Honki no Netsujou 2: Koibito wa Shouboushi	SeaLia	Overall Rating: 1	Aug 25, 2021
20	Initial D First Stage	Patrxaa	Overall Rating: 8	Aug 25, 2021
21	Karigurashi no Arrietty	Anterb	Overall Rating: 9	Aug 25, 2021
22	Hige wo Soru. Soshite Joshikousei wo Hirou.	RenFlexx	Overall Rating: 10	Aug 25, 2021
23	Shiguang Dailiren	Zeooooo	Overall Rating: 9	Aug 25, 2021
24	Vinland Saga	Ali_Hayki	Overall Rating: 9	Aug 25, 2021
25	Tokyo Revengers	muvinco	Overall Rating: 6	Aug 25, 2021
26	Madonna (Movie)	Remjy	Overall Rating: 10	Aug 25, 2021

Showing 1 to 27 of 1,100 entries, 4 total columns

EXTERNAL LINKS:

FULL DATA SET(IN CSV AND XLSX) URL:
XLSX:
https://docs.google.com/spreadsheets/d/15LfZyQSJkvDDqGz85Z8MQULDLv_B6oms/edit?usp=sharing&ouid=103452182630991993013&rtpof=true&sd=true

CSV:
https://drive.google.com/file/d/19-wAu5XsTbipxVdBtE3Sg2idUi0s9Ans/view?usp=sharing

R CODE:
https://drive.google.com/file/d/14v0yXdB4I3DM1iXw47E0evJ6wjsQHdgN/view?usp=sharing

CODE FOR taskscheduler:
https://drive.google.com/file/d/1-tiTz34VrsDBc7Zh1JjwLIVy26-YvrWI/view?usp=sharing