Data Analytics with R

Project – 01 - Web Crawling and data extraction.

Group – 01

Members – Albert Appouh – aka39, Shrey Sharma – ss4399, Gayathri Murugesan – gm382

Project – Journal no- 1

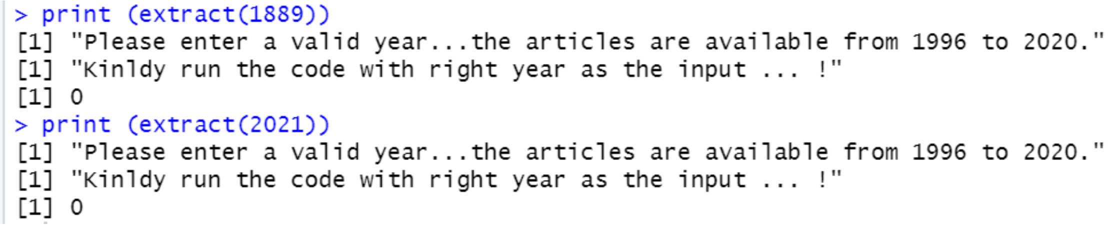
Journal Name – Abstract and Applied Analysis

Project – Implementation description.

Libraries used

library(bitops)
library(RCurl)
library(XML)
library(stringr)

1. extract is the function, which gets the year as the parameter.
2. This function has the code, which crawls the journal page, fetches the article links and crawls each article link to extract the required fields.
3. The site.url is the base url, from where the actual crawling starts.
4. The received parameter – year is checked for the right range of years. The chosen article has the publication from 1996 – 2020. The code prints a warning message to enter the right year and exists, for any year which are not falling within this range.



```
> print (extract(1889))
[1] "Please enter a valid year...the articles are available from 1996 to 2020."
[1] "Kinldy run the code with right year as the input ... !"
[1] 0
> print (extract(2021))
[1] "Please enter a valid year...the articles are available from 1996 to 2020."
[1] "Kinldy run the code with right year as the input ... !"
[1] 0
```

5. The readLines function and htmlParse functions are used to get the html content of the URL.
6. From the base URL, the list of article IDs are greped and a vector with all the article IDs is created.
7. This article ID list is used to loop through each and every article.
8. Here, we have used XPaths, regular expressions, stringR functions to find the required parameters from the article.
9. For the corresponding Authors Emails, there is not any email in the website for that.
10. Viewer can contact the Corresponding Author by clicking on the href of the webpage; it will open a separate goggle form/mailing application to send mail to him. So, regarding this we are adding NA in the corresponding Authors email column.

11. There is no keywords for the articles in our journal. Hence, the Keywords field is not included in the output file.
12. For the articles in 1990’s the corresponding authors are not mentioned. Hence the code returns NULL for the corresponding authors from those years.
13. The extracted fields (DOI, Article Title, Published Date, Author, Abstract, Author Affiliation, Corresponding author and Full Text Link) are made into a matrix and are written into a CSV file – output_data.csv
14. This project and code is developed based on the reference from the sample project Project1_ParseHTMLExample, from the canvas.