Web Scraping using R: Extracting and Analyzing Journal Article Data

Project Overview:

In this project, you will learn how to use R for web scraping by extracting article data from a journal. You will build a specialized R program to crawl, parse, and extract useful information from your selected journal. This hands-on experience will not only enhance your web scraping skills but also prepare you for real-world data extraction and analysis tasks.

Learning Objectives:

By the end of this project, you should be able to:

1. Understand the basics of web scraping.
2. Use R to scrape article data from the Journal.
3. Clean and preprocess the scraped data.
4. Perform basic data analysis on the collected article data.
5. Create visualizations and insights based on the scraped data.

Project Tasks:

Task 1: Set Up Your R Environment

Ensure you have R and RStudio installed on your computer. Install the necessary R packages for web scraping, such as `rvest`, `httr`, and `xml2`.

```
install.packages("rvest")
install.packages("httr")
install.packages("xml2")
```

Task 2: Scraping Article Data

Your goal is to scrape article data from the Journal. Specifically, Given an input year, the objective is to extract all articles published in that year from a journal by extracting the following 7 fields for each article:

Title, Authors, Correspondence Author, Correspondence Author's Email, Publish Date, Abstract, Keywords.

Task 3: Data Cleaning and Preprocessing

After scraping the data, clean and preprocess it to remove any irrelevant information, handle missing values, and format the data for analysis.

Task 4: Data Analysis and Visualization

Conduct basic data analysis on the scraped data. Create only one or two meaningful visualizations (e.g., histograms, scatter plots) to illustrate key findings.

For example, create a bar chart that will visually display which keywords appear most frequently in the articles you've scraped.

Create very simple charts. No need to make it difficult.

Task 5: Report and Insights

Prepare a PowerPoint presentation(approx. 10 slides) summarizing your findings from the article data. Include the following in your presentation:

- Introduction to the project and its objectives.
- Your web scraping code, explaining just the key steps.
- Results of your data analysis, including visualizations.
- Challenges faced while doing your project.
- How each member of your team contributed to the project

Project Submission:

Submit your project as a well-documented **R script**, the **CSV file** of the scraped data along with your **presentation**. Ensure that your code is well-organized and includes comments explaining your thought process.

Evaluation Criteria:

You will be evaluated based on the following criteria:

- Completion of all project tasks.
- Quality and accuracy of the web scraping code.
- Clarity of Visualization Graphs.
- Clarity and professionalism of the presentation.

List of Journals:

1. Human Genomics
2. Immunity & Ageing
3. Malaria Journal
4. Microbiome
5. Mobile DNA
6. Molecular Brain
7. Molecular Cancer
8. Neural Development
9. Parasites & Vectors
10. Particle and Fibre Toxicology
11. Radiation Oncology
12. Retrovirology
13. The Journal of Physiological Sciences
14. Translational Neurodegeneration
15. Virology Journal