

Project 1

In this project, each group is expected to develop a specialized R program to crawl, parse and extract all articles published in a specific journal. Please refer to **Summary.xlsx** for assigned journals.

Given a journal, your R code should be capable of fetching html pages of all articles automatically. For each article, you are required to extract the following **10 fields**:

DOI, Title, Authors, Author Affiliations, Corresponding Author, Corresponding Author's Email, Publication Date, Abstract, Keywords, Full Text (Textual format).

Extracted information should be written into a plain text file (one row per article and one column per field). If any columns are not available, please mark them as **NA** (don't leave them blank).

Your final submission should be a compressed file including **4** folders:

1. all related R scripts and a file readme.txt specifying the functionality of each R script
2. crawled html pages of all articles, the name of each article is DOI.html (e.g., 10.1371/journal.pgen.1005958.html)
3. one plain text file with the aforementioned **10** fields, its name should be JOURNAL_NAME.txt (e.g., PLOS Genetics.txt). **One R script to read the delivered plain text file.**
4. one PDF file for respective contributions of group members and major challenges you have addressed.

REMAKRS:

1. Please focus on the journal **assigned to your group**, otherwise no credits.
2. GroupID and group members are given as follow:

GroupID	S1	S2	S3
1	Shao Yu Chen	Akash Asli	Tushar Asnani
2	Sneha Awasthi	Abhishek Baidyaroy	Ankit Chaudhary
3	Ashvinder Singh	Shau-dong Chou	Prithwish Ganguly
4	Tihtena Kebede	Chandani Patel	Sindur Sarangi
5	Renchao Liu	Devkiran Narayana	Atish Nayak
6	Ami Patel	Shih Chuan Weng	Charmi Ranjitbhai Patel
7	Shivam Patel	Sravani Atmakuri	Huiyuan Yu
8	Hao Peng	Anvesh Kumar Perugu	Raj Ponnusamy
9	Priyadharshini Rajbabu	Shubhangi Rakhe	Tejas Ramani
10	Neil Samant	Yufei Zhang	Ashwin Venkataraman
11	Hassan Khan	Abdullah Kheruwala	

3. Splitting the whole project into two procedures might be more feasible as per my experience. The first phase: crawling all articles and saving them as html files. The second phase: parsing, extracting 10 different fields and delivering the final plain text file. This is only my personal suggestion rather than requirement.
4. A sample code for parsing html pages is also given as your starting point.