Project 1

In this project, you will build a specialized R program to crawl, parse and extract useful information from online websites. Here, we select 10 similar journals for this project. Each group could randomly choose one of them for their projects. But different groups must choose different journals. Please let me know once you and your group select a journal, I will post it on Moodle. (First come, first served.)

Given an input year, your objective is to extract all articles published in/after that year from your selected journal. As a start point, you are required to extract the following **9 fields** for each article:
Title, Authors, Author Affiliations, Correspondence Author, Correspondence Author's Email, Publish Date, Abstract, Keywords, Full Paper (Text format).
The Project_Info_1.xlsx file lists the details about each field in each journal. You could ignore one or two fields in your implementation, if they are not available (marked as NO) for your selected journal. Given an input year, your program is expected to crawl the journal's website automatically, and parse and extract useful fields for each crawled article. The program is expected to store the extracted information into a plain file elegantly. (**One column for one field.**)

In the final submission, please encapsulate your program into a function which will take the year as a parameter. Please submit both the runnable code and the crawled data. Please make sure your stored data could be easily read into R again, as the extracted text may contain some special characters.

Since it may be new to you and your group, there are some useful links in the Project_Info_1.xlsx file. You and your group are highly encouraged to search, study and finish this project independently.

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.