#Project : R program to crawl, parse and extract all articles published in a specific journal
#Group 4: Chandani Patel, Tihtena Taye Kebede , Sindur Sarangi
#2018 Fall

Readme.txt

How to run?

- copy all the R scripts in the wd(working directory
- execute crawled_webpages_GSE.R only

crawled_webpages_GSE.R

- step1: extract main page url and extract total page information store it in main.page.html.

-step2: extract article page url from the main page and extract total page information store it in different files as number of article webpages crawled.

-step3: save all article page with *.html extension. article__of_page61with_serail_number14.html is the 14th article on page 61 of articles webpage.

-Step4: extract all the fields that occur in every article page. DOI, Title, Authors, Author Affiliation, Corresponding Author, Corresponding Author Email, Abstract are extracted and stored in a data frame.

-Step5 – write the data into a text file called Genetics_Selection_Evolution.txt

-Step6 – read the Genetics_Selection_Evolution.txt file in R and store it as a dataframe , named reading.output, with DOI, Title, Authors, Author Affiliation, Corresponding Author, Corresponding Author Email, Abstract, Keywords and FullText as column names. The Keywords and Column names are marked as NA as they do not occur in my article pages.

-Step7 – Proivde the summary of the fields in a excel sheet of all the webpage crawled. Write the dataframe data onto a excel sheet named, GSE_summary_of_fileds.xls.

Library Used:

library(bitops)
library(RCurl)
library(XML)
library(xml2)
library(httr)
library(stringr)
library(Rcrawler)
library(rvest)
library(rlist)

library(openxlsx)

Function Used:

LinkExtractor() – to extract articles url from webpages

Grep()- search for matches to argument `pattern` within each element of a character vector

List.append()- to append a data to a existing list.

Read_html()- Read in the content from a .html file.

xpathSApply()-xpathSApply is a version of xpathApply which attempts to simplify the result if it can be converted to a vector or matrix rather than left as a list. It uses XPath of the html pages to get required content of the page.

htmlParse()-Parses an XML or HTML file or string containing XML/HTML content, and generates an R structure representing the XML/HTML tree

Gsub()-perform replacement of the first and all matches respectively.

Trimws()-trims white spaces in a string. It can be leading or trailing spaces or both.