PROJECT 1

Chandni Mandaviya(csm45)
Komaldeep Kaur Bhatia(kb488)
Twinkle Chaurasia(tc449)

Readme

Our project contains two Scripts. One of them is the Scrapping.R file, which contains the bulk of our code, and requires five packages: stringr, xml2, XML,writexl and rvest. This RScript will first retrieve the URLs for every article published in Mobile DNA in a particular year (The input will be given by the user). Following, the program will go to each one of those URLs and crawl the site to record the following information about each article:
Title, Authors, Author Affiliations, Correspondence Author, Publish Date, Abstract, Keywords, Full Paper (Text format).

The correspondence author's email is not included because the field information was unavailable in the journal. After all this information has been recorded, the program will generate a data frame, and from it will generate a plain text file called Summary.txt and write all the gathered information into the summary.txt text file. Additionally, it also generates an excel file summary.xlsx which includes all the fields except "Full-text". The reason it doesn't include a full-text field is that it exceeds Excel's character limit of 32,767 characters.

Scraping.R only gets the desired information for one particular year. In order to get that information for all the years after the input year, we use the After.R file. This is the RScript with which the user directly interacts. It asks the user for an input year, and it will call the Scraping.R program for every year between the input year and 2020. Hence after After.R has finished running, there will be Summary.txt file which contains the aforementioned information for every article published by Mobile DNA in and after the input year.