**OPTIONAL JOURNAL SCRAPING PROJECT REPORT**

**Subject-** CS636 (Data Analytics with R)

**Professor –** Yao Shen

**TEAM:**

Sirisha Bojjireddy- sb2423

Xuwen Zhao – xz455

Vimal Patel – vp223

Aleksander M. Zaloga - amz35

**CONTRIBUTION OF TEAM MEMBERS:**

Sirisha & Xuwen -> worked on the crawling

Vimal & Aleksander -> worked om the parsing

**WORKFLOW**

CRAWL

PARSE

READ

**ALGORITHM:**

We developed 3 R scripts for crawling, parsing and reading the articles from html pages mentioned below respectively

Crawler.R

1) We created a function “enter” with year as parameter.

2) We used a paste command to take the

3) switch for the year

4) Create a folder “articles” to download all the html files

5) Loop for the html in articles folder and download each file

parser.R

1) Split the data and analyze according to the attributes required by us

read.R

1) Read the following attributes from the html files downloaded

"Title", "DOI", "Authors", "Date", "Abstract", "Keywords", "Text"

**HTML files crawled:**

10.1186\_s13100-018-0143-2.html

10.1186\_s13100-018-0144-1.html

10.1186\_s13100-019-0145-8.html

10.1186\_s13100-019-0146-7.html

10.1186\_s13100-019-0147-6.html

10.1186\_s13100-019-0148-5.html

10.1186\_s13100-019-0149-4.html

10.1186\_s13100-019-0150-y.html

10.1186\_s13100-019-0151-x.html

10.1186\_s13100-019-0152-9.html

10.1186\_s13100-019-0153-8.html

10.1186\_s13100-019-0154-7.html

10.1186\_s13100-019-0155-6.html

10.1186\_s13100-019-0156-5.html

10.1186\_s13100-019-0157-4.html

10.1186\_s13100-019-0158-3.html

10.1186\_s13100-019-0159-2.html

10.1186\_s13100-019-0160-9.html

10.1186\_s13100-019-0161-8.html

10.1186\_s13100-019-0162-7.html