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

CS636 DATA ANALYTICS WITH R

PROJECT 1

JOURNAL : MOBILE DNA

GROUP : 13

CONTRIBUTIONS OF GROUP MEMBERS

MEMBER NAME	CONTRIBUTIONS
Chandni Mandaviya	• Created the function to extract required information, worked on commands, and store data in excel file summary.xlsx .Also worked on reading the lines of data. • Readme document. • Testing and debugging.
Komaldeep Kaur Bhatia	• Created the function to get the year from the user, to generate all the article URL's and extracting required information also worked on the After input year results. • Presentation Document • Testing and debugging.
Twinkle Chaurasia	• Looked into the crawled output binding storing and extracting data using a for loop for the years after the input year specified by the user. • Testing and debugging.

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

CHALLENGES FACED

UNFAMILIARITY WITH R PACKAGES LIKE RVEST, XML, WRITEXL

EXTRACTING AND FORMATTING DATA FROM HTML PAGES.

STORING THE DATA IN THE EXCEL FILE.

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

SOURCE CODE MAIN FUNCTION

```
for(x in 1:length(articleURLs))
{
  currentpage <- articleURLs[x]   #1 was used to test the first article
  webpage <- readLines(currentpage)
  htmlpage <- htmlParse(webpage, asText=TRUE)
  test1 <- xpathSApply(htmlpage, "//*[@id=\"main-content\"]/main/article/div/h1", xmlValue)
  title <- c(title, paste(test1, collapse = ", "))
  test2 <- xpathSApply(htmlpage, "//*[@id=\"article-info-content\"]/div/div/ul[2]/li", xmlValue)
  keywords <- c(keywords, paste(test2, collapse = ", "))
  test3 <- xpathSApply(htmlpage, "//*[@id=\"main-content\"]/main/article/div/ul/li/span/a", xmlValue)
  author <- c(author, paste(test3, collapse = ", "))
  test4 <- xpathSApply(htmlpage, "//*[@id=\"main-content\"]/main/article/div/ul/li[3]/a/time", xmlValue)
  pubDate <- c(pubDate, paste(test4, collapse = ", "))
  test5 <- xpathSApply(htmlpage, "//*[@id=\"Abs1-content\"]/p", xmlValue)
  abstract <- c(abstract, paste(test5, collapse = ", "))
  test6 <- xpathSApply(htmlpage, "//*[@id=\"author-information-content\"]/ol", xmlValue)
  affiliations <- c(affiliations, paste(test6, collapse = ", "))
  text <- unlist(xpathApply(htmlpage, '//p | ///h1 | ///h2 | ///h3 | //*[@class="c-author-list js-list-authors js-etal-collapsed"]', xmlValue))
  text <- gsub('\\n', '', text)
  text <- paste(text, collapse = ': ')
  text <- str_remove(text, "Advertisement:")
  fulltext <- c(fulltext, text)
  htmlwebpage <- read_html(articleURLs[x])
  coauth <- c(coauth, html_text(html_nodes(htmlwebpage, 'a#corresp-c1')))
}

for(i in 1:length(title))
{
  DF <- data.frame(title[i], author[i], coauth[i], pubDate[i], abstract[i], keywords[i], fulltext[i])
  names <- c('Title', 'Authors', 'Corresponding Author', 'Publication Date', 'Abstract', 'Keywords', 'Full Text')
  if(i == 1){
    write.table(DF, file="Summary.txt", sep="\t", col.names = names ,row.names = F)
  }else{
    write.table(DF, file="Summary.txt", sep="\t", row.names = FALSE, col.names = FALSE, append = TRUE)
  }
}
DF <- data.frame(title, author, coauth, pubDate, abstract,keywords)
write_xlsx(DF,"Summary.xlsx")
```

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

THANK YOU