GROUP ACTIVITY - BSIT 2C

PINEDA, PEREZ, SAPAN, JADULOS

2025-12-04

```
library(rvest)
library(dplyr)
library(lubridate)
library(stringr)
library(purrr)
library(httr)
library(ggplot2)

extractArxivPapers <- function(url) {
  page <- GET(url, add_headers("User-Agent" = "Mozilla/5.0"))
  html <- read_html(page)
  dtNodes <- html %>% html_nodes("dt")
  if (length(dtNodes) == 0) return(data.frame())
  paperIds <- character(length(dtNodes))
  titles <- character(length(dtNodes))
  authors <- character(length(dtNodes))
  abstracts <- character(length(dtNodes))
  submissionDates <- character(length(dtNodes))
  originallyAnnounced <- character(length(dtNodes))
  ddNodes <- html %>% html_nodes("dd")
  for (i in 1:min(length(dtNodes), length(ddNodes))) {
    paperIdRaw <- dtNodes[i] %>% html_text() %>%
↪   str_extract("arXiv:\\d{4}\\.\\d{4,5}(v\\d+)?")
    paperIds[i] <- ifelse(is.na(paperIdRaw), "", paperIdRaw)
    if (paperIds[i] == "") next
    dd <- ddNodes[i]
    titleNode <- dd %>% html_node("div.list-title")
    titles[i] <- if (!is.null(titleNode)) titleNode %>% html_text(trim = TRUE) %>%
↪   str_remove("^Title:\\s*") else ""
    authorNode <- dd %>% html_node("div.list-authors")
    authors[i] <- if (!is.null(authorNode)) authorNode %>% html_text(trim = TRUE) %>%
↪   str_remove("^Authors:\\s*") else ""
    if (paperIds[i] != "") {
      paperUrl <- paste0("https://arxiv.org/abs/", str_remove(paperIds[i], "arXiv:"))
      tryCatch({
        Sys.sleep(0.5)
        paperPage <- GET(paperUrl, add_headers("User-Agent" = "Mozilla/5.0"))
        paperHtml <- read_html(paperPage)
        abstractNode <- paperHtml %>% html_node("blockquote.abstract")
        if (!is.null(abstractNode)) {
          abstracts[i] <- abstractNode %>% html_text(trim = TRUE) %>%
↪   str_remove("^Abstract:\\s*") %>% str_remove_all("\\n") %>% str_squish()
```

```
        }
        submissionNode <- paperHtml %>% html_node("div.submission-history")
        if (!is.null(submissionNode)) {
          submissionText <- submissionNode %>% html_text(trim = TRUE)
          dates <- str_extract_all(submissionText, "\\[v\\d+\\]\\s*(\\w{3},
↪  \\s*\\d{1,2}\\s*\\w{3}\\s*\\d{4})")[[1]]
          if (length(dates) > 0) {
            cleanDates <- dates %>% str_remove("\\[v\\d+\\]\\s*") %>% str_trim()
            parsedDates <- parse_date_time(cleanDates, orders = c("%a, %d %b %Y", "%a,%d
↪  %b %Y"))
            if (length(parsedDates) > 0) {
              originallyAnnounced[i] <- as.character(as.Date(min(parsedDates)))
              submissionDates[i] <- as.character(as.Date(max(parsedDates)))
            }
          }
        }
        if (originallyAnnounced[i] == "" || submissionDates[i] == "") {
          datelineNode <- paperHtml %>% html_node("div.dateline")
          if (!is.null(datelineNode)) {
            datelineText <- datelineNode %>% html_text(trim = TRUE)
            dateMatch <- str_extract(datelineText, "\\d{1,2}\\s+\\w{3}\\s+\\d{4}")
            if (!is.na(dateMatch)) {
              parsedDate <- parse_date_time(dateMatch, orders = "%d %b %Y")
              if (!is.na(parsedDate)) {
                dateStr <- as.character(as.Date(parsedDate))
                submissionDates[i] <- dateStr
                originallyAnnounced[i] <- dateStr
              }
            }
          }
        }
      }, error = function(e) {
        message(paste("Error fetching", paperUrl, ":", e$message))
      })
    }
  }
  dois <- ifelse(paperIds != "", paste0("https://doi.org/10.48550/", paperIds), "")
  df <- data.frame(
    PaperID = paperIds,
    Title = titles,
    Authors = authors,
    Abstract = abstracts,
    SubmissionDate = submissionDates,
    OriginallyAnnounced = originallyAnnounced,
    DOI = dois,
    stringsAsFactors = FALSE
  )
  df <- df %>% filter(PaperID != "" & Abstract != "")
  df
}

urls <- c(
  "https://arxiv.org/list/q-bio/2025-01",
```

```
  "https://arxiv.org/list/q-bio/2025-02",
  "https://arxiv.org/list/q-bio/2025-03",
  "https://arxiv.org/list/q-bio/2025-04",
  "https://arxiv.org/list/q-bio/2025-05",
  "https://arxiv.org/list/q-bio/2025-06",
  "https://arxiv.org/list/q-bio/2025-07",
  "https://arxiv.org/list/q-bio/2025-08",
  "https://arxiv.org/list/q-bio/2025-09"
)

allPapers <- map_df(urls, function(u) {
  message(paste("Scraping:", u))
  Sys.sleep(2)
  extractArxivPapers(u)
})

finalPapers <- allPapers %>%
  distinct(PaperID, .keep_all = TRUE) %>%
  head(200)

finalPapers <- finalPapers %>%
  mutate(
    SubmissionDate = ifelse(SubmissionDate == "", NA, SubmissionDate),
    OriginallyAnnounced = ifelse(OriginallyAnnounced == "", NA, OriginallyAnnounced),
    SubmissionDate = as.Date(SubmissionDate),
    OriginallyAnnounced = as.Date(OriginallyAnnounced),
    OriginallyAnnounced = if_else(is.na(OriginallyAnnounced) & !is.na(SubmissionDate),
    ↪  SubmissionDate, OriginallyAnnounced),
    monthApprox = case_when(
      !is.na(SubmissionDate) ~ as.Date(SubmissionDate),
      !is.na(OriginallyAnnounced) ~ as.Date(OriginallyAnnounced),
      TRUE ~ as.Date("2025-01-01")
    ),
    SubmissionDate = if_else(is.na(SubmissionDate), monthApprox, SubmissionDate),
    OriginallyAnnounced = if_else(is.na(OriginallyAnnounced), monthApprox,
    ↪  OriginallyAnnounced)
  ) %>%
  select(-monthApprox)

write.csv(finalPapers, "arxivQbio2025Papers.csv", row.names = FALSE)

if (nrow(finalPapers) > 0) View(finalPapers)

if (nrow(finalPapers) > 0) {
  plotData <- finalPapers %>%
    mutate(month = floor_date(OriginallyAnnounced, "month")) %>%
    filter(!is.na(month)) %>%
    group_by(month) %>%
    summarise(count = n()) %>%
    arrange(month) %>%
    complete(month = seq.Date(min(month), max(month), by = "month"), fill = list(count =
    ↪  0))
```

```
  if (nrow(plotData) > 0) {
    p <- ggplot(plotData, aes(x = month, y = count)) +
      geom_line(color = "steelblue", size = 1) +
      geom_point(color = "darkred", size = 2) +
      geom_text(aes(label = count), vjust = -0.5, size = 3) +
      scale_x_date(date_breaks = "1 month", date_labels = "%b %Y") +
      labs(
        title = "arXiv q-bio Papers Submission Timeline (2025)",
        subtitle = paste("Total papers:", nrow(finalPapers)),
        x = "Month",
        y = "Number of Papers"
      ) +
      theme_minimal() +
      theme(plot.title = element_text(hjust = 0.5, face = "bold"),
            plot.subtitle = element_text(hjust = 0.5),
            axis.text.x = element_text(angle = 45, hjust = 1))
    print(p)
    ggsave("arxivQbioTimeseries.png", p, width = 12, height = 6, dpi = 300)
  } else cat("No valid dates for plotting.\n")
} else cat("No papers collected for plotting.\n")

print(p)
```