Group_Activity

Garcia, Piano, Talaban, Ticot

2025-12-01

```r
library(rvest)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(stringr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```r
library(ggplot2)
library(httr)
library(tinytex)
```

```r
titles <- character(0)
authors <- character(0)
submission_dates <- character(0)
originally_announced <- character(0)
doi <- character(0)
```

```r
base_url <- "https://arxiv.org/search/?query=physics&searchtype=all&abstracts=show&order=-announced_date

all_papers<-list()

starts <-seq(from =0, to =150, by=50)

for (i in starts) {


  url <- paste0(base_url, i)
  print(paste("Scraping:", url))

  tryCatch({
    page <- read_html(url)


    papers_html <- page %>% html_nodes("li.arxiv-result")


    titles <- papers_html %>%
      html_node("p.title.is-5.mathjax") %>%
      html_text(trim = TRUE)

    authors <- papers_html %>%
      html_node("p.authors") %>%
      html_text(trim = TRUE) %>%
      str_remove("Authors:\n")

    abstracts <- papers_html %>%
      html_node("span.abstract-full") %>%
      html_text(trim = TRUE) %>%
      str_remove(" Less")

    meta_raw <- papers_html %>%
      html_node("p.is-size-7") %>%
      html_text(trim = TRUE)

    temp_df <- data.frame(
      title = titles,
      author = authors,
      abstract = abstracts,
      meta_raw = meta_raw,
      stringsAsFactors = FALSE
    )

    all_papers[[length(all_papers) + 1]] <- temp_df

  }, error = function(e) {
    print(paste("Error on page starting at", i))
  })


  Sys.sleep(3)
}
```

```
## [1] "Scraping: https://arxiv.org/search/?query=physics&searchtype=all&abstracts=show&order=-announced
## [1] "Scraping: https://arxiv.org/search/?query=physics&searchtype=all&abstracts=show&order=-announced
## [1] "Scraping: https://arxiv.org/search/?query=physics&searchtype=all&abstracts=show&order=-announced
## [1] "Scraping: https://arxiv.org/search/?query=physics&searchtype=all&abstracts=show&order=-announced
```

```r
df_papers <- bind_rows(all_papers)


print(paste("Total papers extracted:", nrow(df_papers)))
```

```
## [1] "Total papers extracted: 200"
```

```r
df_clean <- df_papers %>%
  mutate(

    submission_date_text = str_extract(meta_raw, "Submitted.*?(=?;)"),
    submission_date_text = str_remove_all(submission_date_text, "Submitted |;"),
    submission_date = dmy(submission_date_text),


    doi = str_extract(meta_raw, "doi:.*"),
    doi = str_remove(doi, "doi:"),


    announced_date_text = str_extract(meta_raw, "originally announced [A-Za-z]+ [0-9]{4}"),
    announced_date_text = str_remove(announced_date_text, "originally announced "),
    originally_announced = my(announced_date_text)
  )


df_clean <- df_clean %>% filter(!is.na(submission_date))

head(df_clean %>% select(title, submission_date, doi))
```

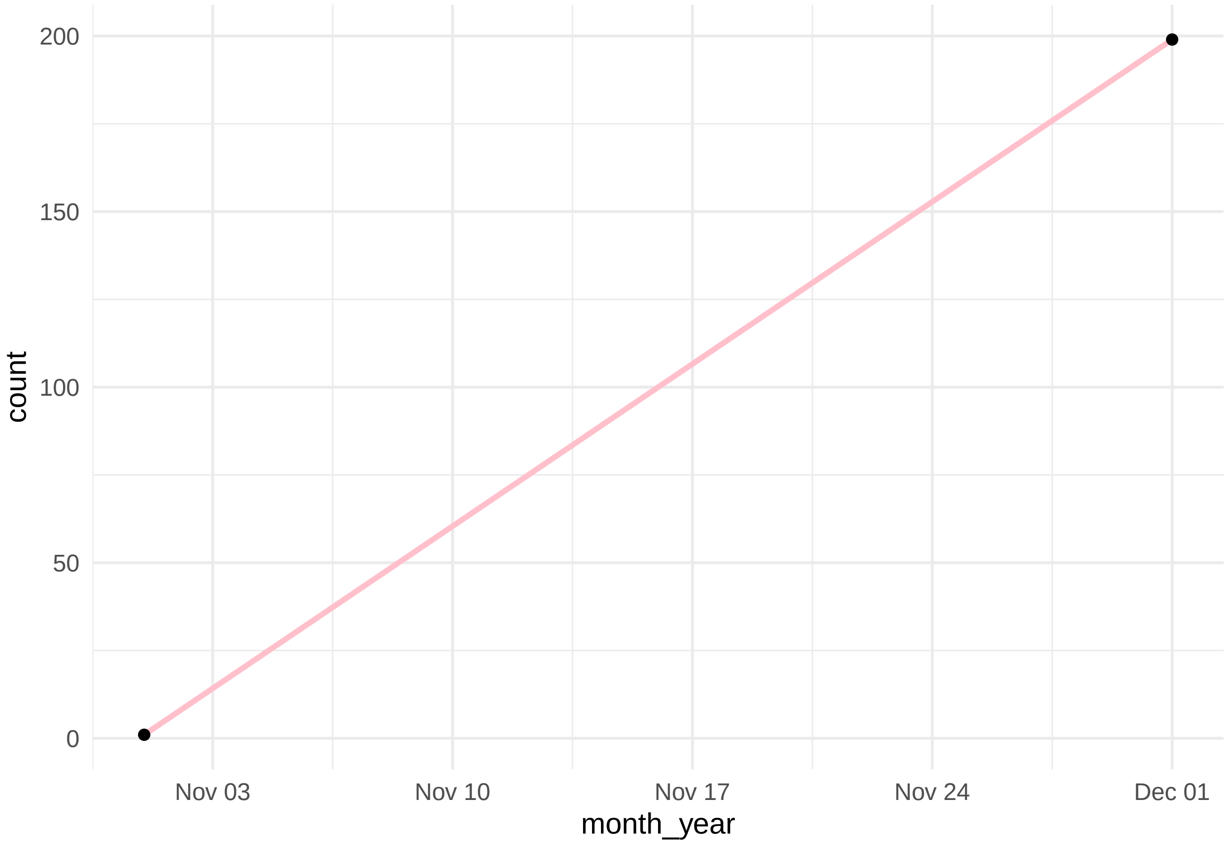
```
##
## 1 Objects in Generated Videos Are Slower Than They Appear: Models Suffer Sub-Earth Gravity and Don't
## 2                                               LLM-Driven Corrective Robot Operation Code Generatio
## 3       JWST & the Waz Arc I: Spatially Resolving the Physical Conditions within a Post-Starburst Gal
## 4                                           Parametric processes in nonlinear structures with reflec
## 5                                                 Constraining Dark Acoustic Oscillations with the H
## 6                                                 Dimensionality and confinement reshape competition
##   submission_date  doi
## 1      2025-12-01 <NA>
## 2      2025-12-01 <NA>
## 3      2025-12-01 <NA>
## 4      2025-12-01 <NA>
## 5      2025-12-01 <NA>
## 6      2025-12-01 <NA>
```

```r
df_sorted <- df_clean %>%
  arrange(submission_date)
```

```r
papers_per_month <- df_sorted %>%
  mutate(month_year = floor_date(submission_date, "month")) %>%
  group_by(month_year) %>%
  summarise(count = n())

ggplot(papers_per_month, aes(x = month_year, y = count)) +
  geom_line(color = "pink", size = 1) +
  geom_point(color = "black") +
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```


200
150
100
50
0
count
Nov 03
Nov 10
Nov 17
Nov 24
Dec 01
month_year