Markdown_Week6_Week7

Aymeric Collart

2025-10-02

1. Prepare the environment

```
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
```

```
library(xml2)
library(openxlsx)
```

2. Scraping

2.1 List the URLs of the webpages by date and by category

```
ETTodayPageUrl <- "https://www.ettoday.net/news/news-list"

Year <- 2024
Month <- 1:12
Day <- 1:31

Dates_total <- data.frame(matrix(ncol = 1, nrow = 0))

for (a in 1:length(Day)){
  for (b in 1:length(Month)){
    for (c in min(Year):max(Year)){
      Dates <- paste0(c, "-", b, "-", a)
```

```
      Dates <- as.data.frame(Dates)
      Dates_total <- rbind(Dates_total, Dates)
    }
  }
}

Dates_total <- as.character(Dates_total$Dates)

# "6.htm" = "Society"
# "1.htm" = "Politics"
# "2.htm" = "International"
# "5.htm" = "Life"
Category_total <- c("6", "1")

URL_total <- data.frame(matrix(ncol = 1, nrow = 0))

for (a in 1:length(Category_total)){
  URL_temp <- paste0(ETTodayPageUrl, "-", Dates_total, "-", Category_total[a], ".htm")
  URL_temp <- as.data.frame(URL_temp)
  URL_total <- rbind(URL_total, URL_temp)
}

URL_total <- as.character(URL_total$URL_temp)

head(URL_total)
```

```
## [1] "https://www.ettoday.net/news/news-list-2024-1-1-6.htm"
## [2] "https://www.ettoday.net/news/news-list-2024-2-1-6.htm"
## [3] "https://www.ettoday.net/news/news-list-2024-3-1-6.htm"
## [4] "https://www.ettoday.net/news/news-list-2024-4-1-6.htm"
## [5] "https://www.ettoday.net/news/news-list-2024-5-1-6.htm"
## [6] "https://www.ettoday.net/news/news-list-2024-6-1-6.htm"
```

2.2. Retrieve the URLs of the individual articles

2.2.1 One page to test

```
IndPage <- read_html(URL_total[1])

news_url <- IndPage %>%
  html_nodes("div.part_list_2") %>%
  html_nodes("h3") %>%
  html_nodes("a") %>%
  html_attr("href")

head(news_url)
```

```
## [1] "https://www.ettoday.net/news/20240101/2655300.htm"
## [2] "https://www.ettoday.net/news/20240101/2655313.htm"
## [3] "https://www.ettoday.net/news/20240101/2655305.htm"
## [4] "https://www.ettoday.net/news/20240101/2655303.htm"
```

```
## [5] "https://www.ettoday.net/news/20240101/2655277.htm"
## [6] "https://www.ettoday.net/news/20240101/2655264.htm"
```

2.2.2 Find the URLs of the articles from all the pages

```
## 3 to 4 minutes
URL_IndArticles <- data.frame(matrix(ncol = 1, nrow = 0))

for (i in 1:length(URL_total)){
  IndPage <- read_html(URL_total[i])
  news_url <- IndPage %>%
    html_nodes("div.part_list_2") %>%
    html_nodes("h3") %>%
    html_nodes("a") %>%
    html_attr("href")

  news_url <- as.data.frame(news_url)
  URL_IndArticles <- rbind(URL_IndArticles, news_url)
}
```

2.2.3 Check for repeated URLs

```
n_occur <- data.frame(table(URL_IndArticles$news_url))
Duplicates <- n_occur[n_occur$Freq > 1,]

article_overview_unique <- URL_IndArticles %>%
  group_by(news_url) %>%
  slice_sample(n = 1)

head(article_overview_unique)
```

```
## # A tibble: 6 x 1
## # Groups:   news_url [6]
##   news_url
##   <chr>
## 1 https://www.ettoday.net/news/20231231/2652951.htm
## 2 https://www.ettoday.net/news/20231231/2653322.htm
## 3 https://www.ettoday.net/news/20231231/2653406.htm
## 4 https://www.ettoday.net/news/20231231/2653538.htm
## 5 https://www.ettoday.net/news/20231231/2654400.htm
## 6 https://www.ettoday.net/news/20231231/2654433.htm
```

2.3 Scrape the content of the articles

2.3.1 Test with one page

```r
OneArticle <- read_html(article_overview_unique$news_url[1])

# time of news
news_time <- OneArticle %>%
  html_nodes("time") %>%
  #html_nodes(".date") %>%
  html_text(trim = TRUE)

# class of news
news_class <- OneArticle %>%
  html_nodes(".part_menu_5") %>%
  html_nodes("strong") %>%
  html_text(trim = TRUE)

# title
news_title <- OneArticle %>%
  html_nodes("h1") %>%
  html_text(trim = TRUE)

# content
news_body <- OneArticle %>%
  html_nodes('div[itemprop="articleBody"]') %>%
  html_text(trim = TRUE)

article <- (data.frame(time = news_time,
                       class = news_class,
                       title = news_title,
                       body = news_body,
                       url = article_overview_unique$news_url[1]))

head(article)
```

```
##                   time class
## 1 2023年12月31日 14:02  社會
##                                                          title
## 1 新北男腦溢血倒警所值班台　警衝百米求助消防分隊！急送醫救回
##
## 1 鄭男疑腦溢血跑到派出所求助，警消協力將鄭男送醫急救　　。（圖／記者陳以🄵翻攝）\r\n記者陳以
##                                                  url
## 1 https://www.ettoday.net/news/20231231/2652951.htm
```

2.3.2 Scrape all the pages

```r
Article_total <- data.frame()

for (j in 1:length(article_overview_unique$news_url)){
  tryCatch({
    temp <- read_html(article_overview_unique$news_url[j])

    # time of news
    news_time <- temp %>%
```

```
      html_nodes("time") %>%
      html_text(trim = TRUE)

    # class of news
    news_class <- temp %>%
      html_nodes(".part_menu_5") %>%
      html_nodes("strong") %>%
      html_text(trim = TRUE)

    # title
    news_title <- temp %>%
      html_nodes("h1") %>%
      html_text(trim = TRUE)

    # content
    news_body <- temp %>%
      html_nodes('div[itemprop="articleBody"]') %>%
      html_text(trim = TRUE)

    # url
    news_url <- temp %>%
      html_nodes("div.block div.block_content div.part_list_2 h3") %>%
      html_nodes("a") %>%
      html_attr("href")

    Article <- (data.frame(time = news_time,
                           class = news_class,
                           title = news_title,
                           body = news_body,
                           url = article_overview_unique$news_url[1]))
    Article_total <- rbind(Article_total, Article)
  }, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
}
```

```
## ERROR : arguments imply differing number of rows: 1, 0
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : arguments imply differing number of rows: 1, 0
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : arguments imply differing number of rows: 1, 0
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : arguments imply differing number of rows: 1, 0
## ERROR : arguments imply differing number of rows: 1, 0
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : cannot open the connection
```

```
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : arguments imply differing number of rows: 1, 0
## ERROR : arguments imply differing number of rows: 1, 0
## ERROR : arguments imply differing number of rows: 1, 0
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : arguments imply differing number of rows: 1, 0
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : arguments imply differing number of rows: 1, 0
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : cannot open the connection
## ERROR : arguments imply differing number of rows: 1, 0
## ERROR : cannot open the connection
## ERROR : arguments imply differing number of rows: 1, 0
## ERROR : cannot open the connection
## ERROR : arguments imply differing number of rows: 1, 0
## ERROR : arguments imply differing number of rows: 1, 0
## ERROR : cannot open the connection
## ERROR : cannot open the connection
```

```
head(Article_total)
```

```
##                   time class
## 1 2023年12月31日 14:02  社會
## 2 2023年12月31日 12:30  社會
## 3 2023年12月31日 15:30  社會
## 4 2023年12月31日 17:00  社會
## 5 2023年12月31日 12:19  社會
## 6 2023年12月31日 12:14  社會
##                                                      title
```

```
## 1  新北男腦溢血倒警所值班台　警衝百米求助消防分隊！急送醫救回
## 2      新北23歲女家教超車擦撞機車　騎士頭部重創慘死！判🄵出爐
## 3    🄵穿🄵衣看電影　舞蹈系嫩妹遭🄵吻胸部！色男辯「以🄵她懂」
## 4            台北市熱水器補助元旦開跑　正確安裝居家安全有保障
## 5 🄵吻藥廠女助理！台大醫辯「愛的抱抱」　拿百萬和解不成遭判刑
## 6   一句「看三小」引爆！15人墾丁大街亂鬥　賓士、機車競逐釀1死
##
## 1
## 2
## 3
## 4
## 5 台大骨科名醫猥🄵藥廠研究助理，遭判刑8月。（示意圖，非當事人／取自免費圖庫123RF）\r\n記者郭🄵潔／台
## 6
##                                                  url
## 1 https://www.ettoday.net/news/20231231/2652951.htm
## 2 https://www.ettoday.net/news/20231231/2652951.htm
## 3 https://www.ettoday.net/news/20231231/2652951.htm
## 4 https://www.ettoday.net/news/20231231/2652951.htm
## 5 https://www.ettoday.net/news/20231231/2652951.htm
## 6 https://www.ettoday.net/news/20231231/2652951.htm
```

3. Save the data

3.1 Save as an Excel file

```
write.xlsx(Article_total, "ArticleETToday_CorpusCourse.xlsx")
```

```
## Warning in wb$writeData(df = x, colNames = colNames, sheet = sheet, startCol = startCol, : icy&quot;]
## &lt;/html&gt; is truncated.
## Number of characters exeed the limit of 32767.
```

3.2 Save as an RData file

```
save(Article_total, file = "ArticleETToday_CorpusCourse.Rdata")
```