Lab 06 - Regular Expressions and Web Scraping

Learning goals

- Use a real world API to make queries and process the data.
- Use regular expressions to parse the information.
- Practice your GitHub skills.

Lab description

In this lab, we will be working with the NCBI API to make queries and extract information using XML and regular expressions. For this lab, we will be using the `httr`, `xml2`, and `stringr` R packages.

This markdown document should be rendered using `github_document` document ONLY and pushed to your *JSC370-labs* repository in `lab06/README.md`.

Question 1: How many sars-cov-2 papers?

Build an automatic counter of sars-cov-2 papers using PubMed. You will need to apply XPath as we did during the lecture to extract the number of results returned by PubMed in the following web address:

```
https://pubmed.ncbi.nlm.nih.gov/?term=sars-cov-2
```

Complete the lines of code:

```
# Downloading the website
website <- xml2::read_html("https://pubmed.ncbi.nlm.nih.gov/?term=sars-cov-2")

# Finding the counts
counts <- xml2::xml_find_first(website, "//div[@class='results-amount']/span")

# Turning it into text
counts <- as.character(counts)

# Extracting the data using regex
stringr::str_extract(counts, "[0-9]+,?[0-9]+")
```

```
## [1] "192,677"
```

- How many sars-cov-2 papers are there?

Answer here. 192,667 (when I searched it.)

Don't forget to commit your work!

Question 2: Academic publications on COVID19 and Hawaii

Use the function `httr::GET()` to make the following query:

1. Baseline URL: https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi
2. Query parameters:
    - db: pubmed
    - term: covid19 hawaii
    - retmax: 1000

The parameters passed to the query are documented here.

```
library(httr)
query_ids <- GET(
  url   = "https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi",
  query = list(retmax=1000,
               db="pubmed",
               term="covid19 hawaii")
)

# Extracting the content of the response of GET
ids <- httr::content(query_ids)
```

The query will return an XML object, we can turn it into a character list to analyze the text directly with `as.character()`. Another way of processing the data could be using lists with the function `xml2::as_list()`. We will skip the latter for now.

Take a look at the data, and continue with the next question (don't forget to commit and push your results to your GitHub repo!).

```
# as.character(ids)
# I took a look at the data.
```

Question 3: Get details about the articles

The Ids are wrapped around text in the following way: `<Id>... id number ...</Id>`. we can use a regular expression that extract that information. Fill out the following lines of code:

```
# Turn the result into a character vector
ids <- as.character(ids)

# Find all the ids
ids <- stringr::str_extract_all(ids, "<Id>.*</Id>")[[1]]

# Remove all the leading and trailing <Id> </Id>. Make use of "|"
ids <- stringr::str_remove_all(ids, "(<Id>)|(</Id>)")
```

With the ids in hand, we can now try to get the abstracts of the papers. As before, we will need to coerce the contents (results) to a list using:

1. Baseline url: https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi

2. Query parameters:

    - db: pubmed
    - id: A character with all the ids separated by comma, e.g., "1232131,546464,13131"
    - retmax: 1000
    - rettype: abstract

Pro-tip: If you want `GET()` to take some element literal, wrap it around `I()` (as you would do in a formula in R). For example, the text `"123,456"` is replaced with `"123%2C456"`. If you don't want that behavior, you would need to do the following `I("123,456")`.

```
# I will take the abstracts of the first 100 entries or else the dataset takes too long to load.
publications <- GET(
  url   = "https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi",
  query = list(
    db=I("pubmed"),
    id=paste(ids, collapse=","),
    retmax=1000,
    rettype=I("abstract")
    )
)
# paste use sep if you have args of length 1 eg. paste("1st", "2nd", "3rd", sep = ", ")
# else use collapse

# Turning the output into character vector
publications <- httr::content(publications)
publications_txt <- as.character(publications)
```

With this in hand, we can now analyze the data. This is also a good time for committing and pushing your work!

Question 4: Distribution of universities, schools, and departments

Using the function `stringr::str_extract_all()` applied on `publications_txt`, capture all the terms of the form:

1. University of ...
2. ... Institute of ...

Write a regular expression that captures all such instances

```
# s1 = "(University|Institute) of [a-zA-Z ]*[,]"
# this one works better, although it wouldn't incorporate ... University
s2 = "(University|Institute) of ([A-Z][a-z ]*)+"
library(stringr)
institution <- str_extract_all(
  publications_txt,
  s2
  )
institution <- unlist(institution)
inst_frame <- as.data.frame(table(institution))
library(dplyr)
head(inst_frame %>% arrange(desc(Freq)))
```

```
##                            institution Freq
## 1                   University of Hawai  362
## 2         University of Hawaii at Manoa  212
## 3                  University of Hawaii  111
## 4            University of Pennsylvania   92
## 5              University of California   69
## 6 University of Science and Technology   34
```

Note that if an abstract includes the name of a university/institute multiple times, it will be counted multiple times. Also note that universities with other formats eg. Harvard University, won't be included as they are not captured by the regex.

Repeat the exercise and this time focus on schools and departments in the form of

1. School of ...
2. Department of ...

And tabulate the results

```
schools_and_deps <- str_extract_all(
  publications_txt,
  "(School|Department) of ([A-Z][a-z ]*)+"
  )
school_frame <- as.data.frame(table(schools_and_deps))
head(school_frame %>% arrange(desc(Freq)))
```

```
##                                      schools_and_deps Freq
## 1                                  School of Medicine  639
## 2                               Department of Medicine  202
## 3 Department of Preventive Medicine and Biostatistics  159
## 4                                 Department of Health  135
## 5                              School of Public Health   74
## 6                      Department of Tropical Medicine   62
```

Note that the same problem occurs again for this problem.

For both of these questions. The Regex could be improved to accomodate other cases, but I am a lazy man and I didn't do it.

Question 5: Form a database

We want to build a dataset which includes the title and the abstract of the paper. The title of all records is enclosed by the HTML tag `ArticleTitle`, and the abstract by `Abstract`.

Before applying the functions to extract text directly, it will help to process the XML a bit. We will use the `xml2::xml_children()` function to keep one element per id. This way, if a paper is missing the abstract, or something else, we will be able to properly match PUBMED IDS with their corresponding records.

```
pub_char_list <- xml2::xml_children(publications)
pub_char_list <- sapply(pub_char_list, as.character)
```

Now, extract the abstract and article title for each one of the elements of `pub_char_list`. You can either use `sapply()` as we just did, or simply take advantage of vectorization of `stringr::str_extract`

```
# it seems like abstracts are within <AbstractText> tags, and the tags can have attributes.
abstracts <- str_extract(pub_char_list, "Abstract.*Abstract")
abstracts <- str_remove_all(abstracts, "(Abstract.*>)|(</Abstract)")
abstracts <- str_replace_all(abstracts, "\\s+", " ")
sum(is.na(abstracts))
```

```
## [1] 50
```

- How many of these don't have an abstract?

Answer here.

50 papers.

Now, the title

```
titles <- str_extract(pub_char_list, "ArticleTitle.*ArticleTitle")
titles <- str_remove_all(titles, "(ArticleTitle.*>)|(</ArticleTitle)")
sum(is.na(titles))
```

```
## [1] 0
```

- How many of these don't have a title ?

Answer here. None. All articles have a title

Finally, put everything together into a single `data.frame` and use `knitr::kable` to print the results

I will only look at the first few rows of the database, because I don't want to print out 338 article titles and abstracts.

```
database <- data.frame(
  title=titles,
  abstract=abstracts
)
knitr::kable(head(database))
```

title	abstract
A machine learning approach identifies distinct early-symptom cluster phenotypes which correlate with hospitalization, failure to return to activities, and prolonged COVID-19 symptoms.	Accurate COVID-19 prognosis is a critical aspect of acute and long-term clinical management. We identified discrete clusters of early stage-symptoms which may delineate groups with distinct disease severity phenotypes, including risk of developing long-term symptoms and associated inflammatory profiles.

title	abstract
Barriers and Challenges for Career Milestones Among Faculty Mentees.	'Critical' career milestones for faculty (e.g., tenure, securing grant funding) relate to career advancement, job satisfaction, service/leadership, scholarship/research, clinical or teaching activities, professionalism, compensation, and work-life balance. However, barriers and challenges to these milestones encountered by junior faculty have been inadequately studied, particularly those affecting underrepresented minorities in science (URM-S). Additionally, little is known about how barriers and challenges to career milestones have changed during the COVID-19 pandemic for URM-S and non-URM faculty mentees in science. In this study, we conducted semi-structured interviews with 31 faculty mentees from four academic institutions (located in New Mexico, Arizona, Idaho, and Hawaii), including 22 URM-S (women or racial/ethnic). Respondents were given examples of 'critical' career milestones and were asked to identify and discuss barriers and challenges that they have encountered or expect to encounter while working toward achieving these milestones. We performed thematic descriptive analysis using NVivo software in an iterative, team-based process. Our preliminary analysis identified five key themes that illustrate barriers and challenges encountered: Job and career development, Discrimination and a lack of workplace diversity; Lack of interpersonal relationships and inadequate social support at the workplace; Personal and family matters; and Unique COVID-19-related issues. COVID-19 barriers and challenges were related to online curriculum creation and administration, interpersonal relationship development, inadequate training/service/conference opportunities, and disruptions in childcare and schooling. Although COVID-19 helped create new barriers and challenges for junior faculty mentees, traditional barriers and challenges for 'critical' career milestones continue to be reported among our respondents. URM-S respondents also identified discrimination and diversity-related barriers and challenges. Subsequent interviews will focus on 12-month and 24-month follow-ups and provide additional insight into the unique challenges and barriers to 'critical' career milestones that URM and non-URM faculty in science have encountered during the unique historical context of the COVID-19 pandemic.
COVID-19 Information on YouTube: Analysis of Quality and Reliability of Videos in Eleven Widely Spoken Languages across Africa.	Whilst the coronavirus disease 2019 (COVID-19) vaccination rollout is well underway, there is a concern in Africa where less than 2% of global vaccinations have occurred. In the absence of herd immunity, health promotion remains essential. YouTube has been widely utilised as a source of medical information in previous outbreaks and pandemics. There are limited data on COVID-19 information on YouTube videos, especially in languages widely spoken in Africa. This study investigated the quality and reliability of such videos.

title	abstract
BNT162b2 against COVID-19-associated Emergency Department and Urgent Care Visits among Children 5-11 Years of Age: a Test Negative Design.	In a 1:1 matched test-negative design among 5-11-year-olds in the Kaiser Permanente Southern California health system (n=3984), BNT162b2 effectiveness against omicron-related emergency department or urgent care encounters was 60% [95%CI: 47-69] <3 months post-dose-two and 28% [8-43] after =3 months. A booster improved protection to 77% [53-88].
COVID-19 Information Seeking Behaviors of University of Hawai'i at Manoa Undergraduates: Information Channels, Sources, and Consumption.	This study explored how undergraduate students at the University of Hawai'i at Manoa sought and consumed information about the virus that causes COVID-19. This study also examined student perceptions of the severity of and their susceptibility to the virus and their main concerns about it. Four hundred fifty-six students completed online surveys between October and early December of 2020 and 2021. Students reported low to moderate levels of information seeking across four domains: (1) knowledge about COVID-19 and its symptoms; (2) preventing the spread of the virus; (3) the current state of the pandemic in Hawai'i; and (4) the likely future of the pandemic in Hawai'i. Overall, websites, television, and Instagram were the top 3 channels used by students to seek information for these domains. Students reported primarily paying attention to information from government and news organizations as sources. However, students' preferred channels and sources varied with the type of information they sought. Students also reported believing that COVID-19 is severe and that they are susceptible to being infected with it. The more time students reported seeking information, the greater their perceptions of COVID-19's severity across all domains. Students' primary concerns about COVID-19 centered on state regulations/policies, vaccines, tourism/travel, the economy, and pandemic/post-pandemic life. These findings can help public health practitioners in Hawai'i determine how best to reach an undergraduate student population with information related to COVID-19.

title	abstract
Analysis of mRNA COVID-19 Vaccine Uptake Among Immunocompromised Individuals in a Large US Health System.	Immunocompromised individuals are at increased risk for severe outcomes due to SARS-CoV-2 infection. Given the varying and complex nature of COVID-19 vaccination recommendations, it is important to understand COVID-19 vaccine uptake in this vulnerable population.

Done! Knit the document, commit, and push.

Final Pro Tip (optional)

You can still share the HTML document on github. You can include a link in your `README.md` file as the following:

```
View [here](https://cdn.jsdelivr.net/gh/:user/:repo@:tag/:file)
```

For example, if we wanted to add a direct link the HTML page of lecture 6, we could do something like the following:

```
View Week 6 Lecture [here]()
```