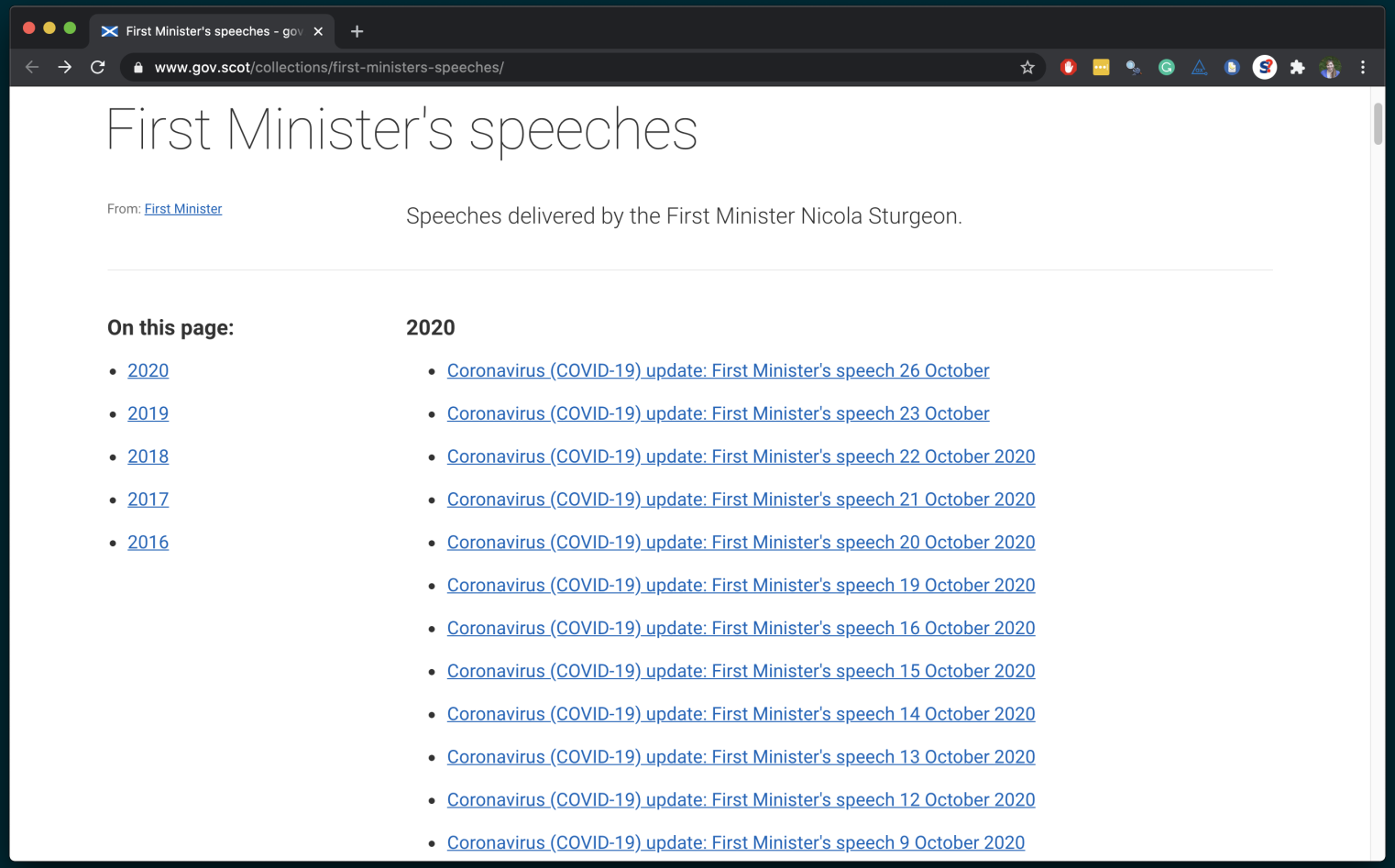
Iteration

Data Science in a Box

datasciencebox.org

First Minister's COVID speeches

🏁 Start with


First Minister's speeches - gov
www.gov.scot/collections/first-ministers-speeches/
First Minister's speeches
From: First Minister
Speeches delivered by the First Minister Nicola Sturgeon.
On this page:
2020
2019
2018
2017
2016
2020
Coronavirus (COVID-19) update: First Minister's speech 26 October
Coronavirus (COVID-19) update: First Minister's speech 23 October
Coronavirus (COVID-19) update: First Minister's speech 22 October 2020
Coronavirus (COVID-19) update: First Minister's speech 21 October 2020
Coronavirus (COVID-19) update: First Minister's speech 20 October 2020
Coronavirus (COVID-19) update: First Minister's speech 19 October 2020
Coronavirus (COVID-19) update: First Minister's speech 16 October 2020
Coronavirus (COVID-19) update: First Minister's speech 15 October 2020
Coronavirus (COVID-19) update: First Minister's speech 14 October 2020
Coronavirus (COVID-19) update: First Minister's speech 13 October 2020
Coronavirus (COVID-19) update: First Minister's speech 12 October 2020
Coronavirus (COVID-19) update: First Minister's speech 9 October 2020

End with 🛑

```
## # A tibble: 218 × 6
##    title                   date       location abstract text  url
##    <chr>                   <date>     <chr>    <chr>    <chr> <chr>
##  1 Coronavirus (COVID-1… 2021-04-20 St Andr… Stateme… "Goo… http…
##  2 Coronavirus (COVID-1… 2021-04-13 St Andr… Stateme… "Tha… http…
##  3 Coronavirus (COVID-1… 2021-04-06 St Andr… Stateme… "Goo… http…
##  4 Coronavirus (COVID-1… 2021-03-30 St Andr… Stateme… "Tha… http…
##  5 Coronavirus (COVID-1… 2021-03-24 Scottis… Stateme… "Tha… http…
##  6 Coronavirus (Covid-1… 2021-03-23 The Sco… Stateme… "Pre… http…
##  7 Coronavirus (COVID-1… 2021-03-18 Scottis… Stateme… "Tha… http…
##  8 Coronavirus (COVID-1… 2021-03-17 St Andr… Stateme… "\nG… http…
##  9 Coronavirus (COVID-1… 2021-03-16 Scottis… Stateme… "Pre… http…
## 10 Coronavirus (COVID-1… 2021-03-15 St Andr… Stateme… "\nG… http…
## 11 Coronavirus (COVID-1… 2021-03-11 Scottis… Stateme… "I c… http…
## 12 Coronavirus (COVID-1… 2021-03-09 Scottis… Stateme… "Pre… http…
## 13 Coronavirus (COVID-1… 2021-03-05 Scottis… Parliam… "Hel… http…
## 14 Coronavirus (COVID-1… 2021-03-04 Scottis… Parliam… "I w… http…
## 15 Coronavirus (COVID-1… 2021-03-02 Scottis… Stateme… "Pre… http…
## # … with 203 more rows
```

Define `scrape_speech()`

```
scrape_speech <- function(url) {

  speech_page <- read_html(url)

  title <- speech_page %>%
    html_node(".article-header__title") %>%
    html_text()

  date <- speech_page %>%
    html_node(".content-data__list:nth-child(1) strong") %>%
    html_text() %>%
    dmy()

  location <- speech_page %>%
    html_node(".content-data__list+ .content-data__list strong") %>%
    html_text()

  abstract <- speech_page %>%
    html_node(".leader--first-para p") %>%
    html_text()

  text <- speech_page %>%
    html_nodes("#preamble p") %>%
    html_text() %>%
    list()

  tibble(
    title = title, date = date, location = location,
    abstract = abstract, text = text, url = url
  )
}
```

Use `scrape_speech()`

```
url_26_oct <- "https://www.gov.scot/publications/coronavirus-covid-19-update-first-minist
scrape_speech(url = url_26_oct)
```

```
## # A tibble: 1 × 6
##   title                  date       location abstract text  url
##   <chr>                  <date>     <chr>    <chr>    <lis> <chr>
## 1 Coronavirus (COVID-19… 2020-10-26 St Andr… <NA>     <chr> http…
```

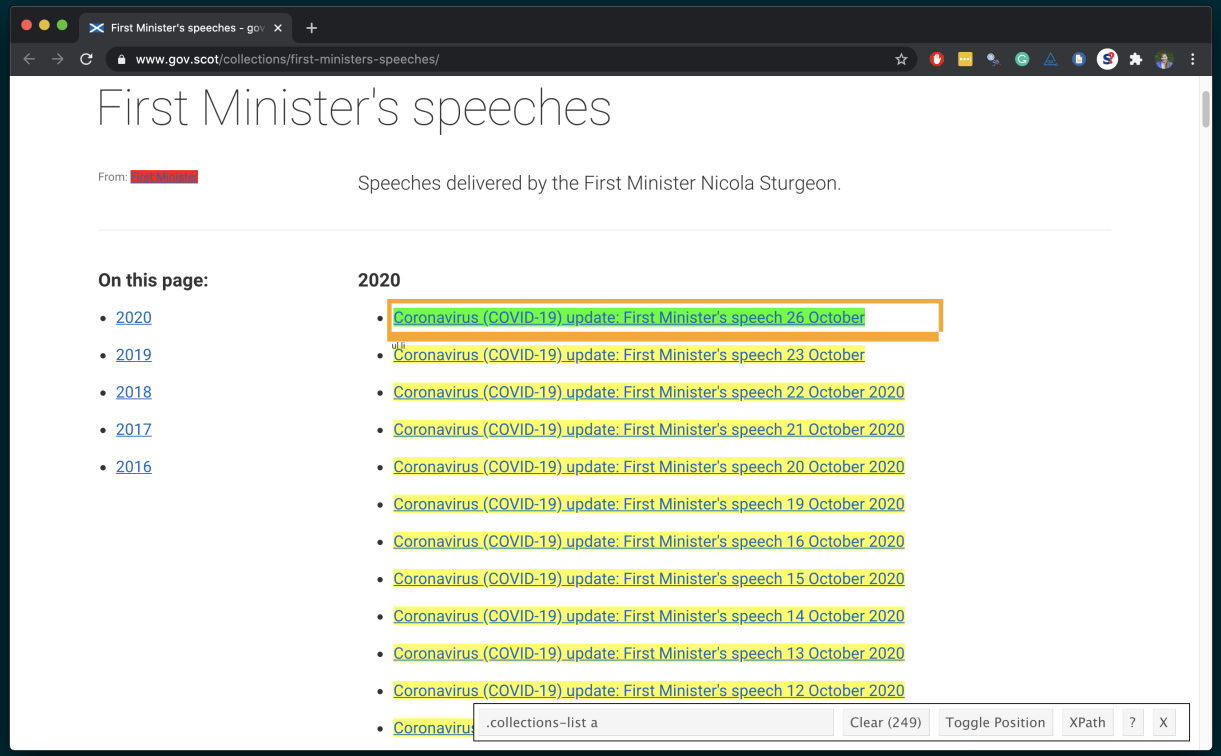
```
url_23_oct <- "https://www.gov.scot/publications/coronavirus-covid-19-update-first-minist
scrape_speech(url = url_23_oct)
```

```
## # A tibble: 1 × 6
##   title                  date       location abstract text  url
##   <chr>                  <date>     <chr>    <chr>    <lis> <chr>
## 1 Coronavirus (COVID-19… 2020-10-23 St Andr… <NA>     <chr> http…
```

Inputs

Inputs

You now have a function that will scrape the relevant info on speeches given the URL of the page of the speech. Where can we get a list of URLs of each of the speeches?



All URLs

```
all_speeches_page <- read_html("https://www.gov.scot/collections/first-ministers-speeches

all_speeches_page %>%
  html_nodes(".collections-list a") %>%
  html_attr("href")
```

```
##   [1] "/publications/fm-statement-parliament-march-30-2022/"
##   [2] "/publications/first-minister-speech-national-economic-forum-2022/"
##   [3] "/publications/ukraine-update-first-ministers-statement-16-march-2022/"
##   [4] "/publications/fm-statement-parliament-march-15-2022/"
##   [5] "/publications/international-womens-day-2022-first-ministers-statement-8-march-2022/"
##   [6] "/publications/solidarity-ukraine-debate-first-ministers-statement/"
##   [7] "/publications/cosla-conference-2022-first-ministers-speech/"
##   [8] "/publications/coronavirus-covid-19-update-first-ministers-speech-tuesday-22-february-
##   [9] "/publications/coronavirus-covid-19-update-first-ministers-statement-8-february-2022/"
##  [10] "/publications/coronavirus-covid-19-update-first-ministers-statement-1-february-2022/"
...
```

COVID-19 URLs *fragments*

```
all_speeches_page %>%
  html_nodes(".collections-list a") %>%
  html_attr("href") %>%
  str_subset("covid-19")
```

```
##   [1] "/publications/coronavirus-covid-19-update-first-ministers-speech-tuesday-22-february-
##   [2] "/publications/coronavirus-covid-19-update-first-ministers-statement-8-february-2022/"
##   [3] "/publications/coronavirus-covid-19-update-first-ministers-statement-1-february-2022/"
##   [4] "/publications/coronavirus-covid-19-update-first-ministers-statement-25-january-2022/"
##   [5] "/publications/coronavirus-covid-19-update-first-ministers-statement-18-january-2022/"
##   [6] "/publications/coronavirus-covid-19-update-first-ministers-statement-11-january-2022/"
##   [7] "/publications/coronavirus-covid-19-update-first-ministers-statement-5-january-2022/"
##   [8] "/publications/coronavirus-covid-19-update-first-ministers-statement-21-december-2021/
##   [9] "/publications/coronavirus-covid-19-update-first-ministers-speech-17-december-2021/"
##  [10] "/publications/coronavirus-covid-19-update-first-ministers-statement-14-december-2021/
...
```

COVID-19 URLs

```
all_speeches_page %>%
  html_nodes(".collections-list a") %>%
  html_attr("href") %>%
  str_subset("covid-19") %>%
  str_c("https://www.gov.scot", .)
```

```
##   [1] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-speech-
##   [2] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-stateme
##   [3] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-stateme
##   [4] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-stateme
##   [5] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-stateme
##   [6] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-stateme
##   [7] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-stateme
##   [8] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-stateme
##   [9] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-speech-
##  [10] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-stateme
...
```

Save COVID-19 URLs

```
covid_speech_urls <- all_speeches_page %>%
  html_nodes(".collections-list a") %>%
  html_attr("href") %>%
  str_subset("covid-19") %>%
  str_c("https://www.gov.scot", .)

covid_speech_urls
```

```
##   [1] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-speech-
##   [2] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-stateme
##   [3] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-stateme
##   [4] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-stateme
##   [5] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-stateme
##   [6] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-stateme
##   [7] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-stateme
##   [8] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-stateme
##   [9] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-speech-
##  [10] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-stateme
...
```

Iteration

Define the task

- Goal: Scrape info on all COVID-19 speeches of the First Minister
- So far:

```
scrape_speech(covid_speech_urls[1])
scrape_speech(covid_speech_urls[2])
scrape_speech(covid_speech_urls[3])
```

- What else do we need to do?
  - Run the `scrape_speech()` function on all COVID-19 speech links
  - Combine the resulting data frames from each run into one giant data frame

Iteration

> How can we tell R to apply the `scrape_speech()` function to each link in `covid_speech_urls`?

- Option 1: Write a **for loop**, i.e. explicitly tell R to visit a link, apply the function, store the result, then visit the next link, apply the function, append the result to the stored result from the previous link, and so on and so forth.
- Option 2: **Map** the function to each element in the list of links, and let R take care of the storing and appending of results.
- We'll go with Option 2!

How does mapping work?

Suppose we have exam 1 and exam 2 scores of 4 students stored in a list...

```
exam_scores <- list(
  exam1 <- c(80, 90, 70, 50),
  exam2 <- c(85, 83, 45, 60)
)
```

...and we find the mean score in each exam

```
map(exam_scores, mean)
```

```
## [[1]]
## [1] 72.5
##
## [[2]]
## [1] 68.25
```

...and suppose we want the results as a numeric (double) vector

```
map_dbl(exam_scores, mean)
```

```
## [1] 72.50 68.25
```

...or as a character string

```
map_chr(exam_scores, mean)
```

```
## [1] "72.500000" "68.250000"
```

`map_something`

Functions for looping over an object and returning a value (of a specific type):

- `map()` - returns a list
- `map_lgl()` - returns a logical vector
- `map_int()` - returns a integer vector
- `map_dbl()` - returns a double vector
- `map_chr()` - returns a character vector
- `map_df()` / `map_dfr()` - returns a data frame by row binding
- `map_dfc()` - returns a data frame by column binding
- ...

Go to each page, scrape speech

- Map the `scrape_speech()` function
- to each element of `covid_speech_urls`
- and return a data frame by row binding

```
covid_speeches <- map_dfr(covid_speech_urls, scrape_speech)
```

```
covid_speeches %>%
  print(n = 15)
```

```
## # A tibble: 218 × 6
##    title                 date       location abstract text  url
##    <chr>                 <date>     <chr>    <chr>    <chr> <chr>
##  1 Coronavirus (COVID-1… 2021-04-20 St Andr… Stateme… "Goo… http…
##  2 Coronavirus (COVID-1… 2021-04-13 St Andr… Stateme… "Tha… http…
##  3 Coronavirus (COVID-1… 2021-04-06 St Andr… Stateme… "Goo… http…
##  4 Coronavirus (COVID-1… 2021-03-30 St Andr… Stateme… "Tha… http…
##  5 Coronavirus (COVID-1… 2021-03-24 Scottis… Stateme… "Tha… http…
##  6 Coronavirus (Covid-1… 2021-03-23 The Sco… Stateme… "Pre… http…
##  7 Coronavirus (COVID-1… 2021-03-18 Scottis… Stateme… "Tha… http…
##  8 Coronavirus (COVID-1… 2021-03-17 St Andr… Stateme… "\nG… http…
##  9 Coronavirus (COVID-1… 2021-03-16 Scottis… Stateme… "Pre… http…
## 10 Coronavirus (COVID-1… 2021-03-15 St Andr… Stateme… "\nG… http…
## 11 Coronavirus (COVID-1… 2021-03-11 Scottis… Stateme… "I c… http…
## 12 Coronavirus (COVID-1… 2021-03-09 Scottis… Stateme… "Pre… http…
## 13 Coronavirus (COVID-1… 2021-03-05 Scottis… Parliam… "Hel… http…
## 14 Coronavirus (COVID-1… 2021-03-04 Scottis… Parliam… "I w… http…
## 15 Coronavirus (COVID-1… 2021-03-02 Scottis… Stateme… "Pre… http…
## # … with 203 more rows
```

What could go wrong?

```
covid_speeches <- map_dfr(covid_speech_urls, scrape_speech)
```

- This will take a while to run
- If you get `HTTP Error 429 (Too many requests)` you might want to slow down your hits by modifying your function to slow it down by adding a random wait (sleep) time between hitting each link

```
scrape_speech <- function(url){

  # Sleep for randomly generated number of seconds
  # Generated from a uniform distribution between 0 and 1
  Sys.sleep(runif(1))

  # Rest of your function code goes here...
}
```