Functions

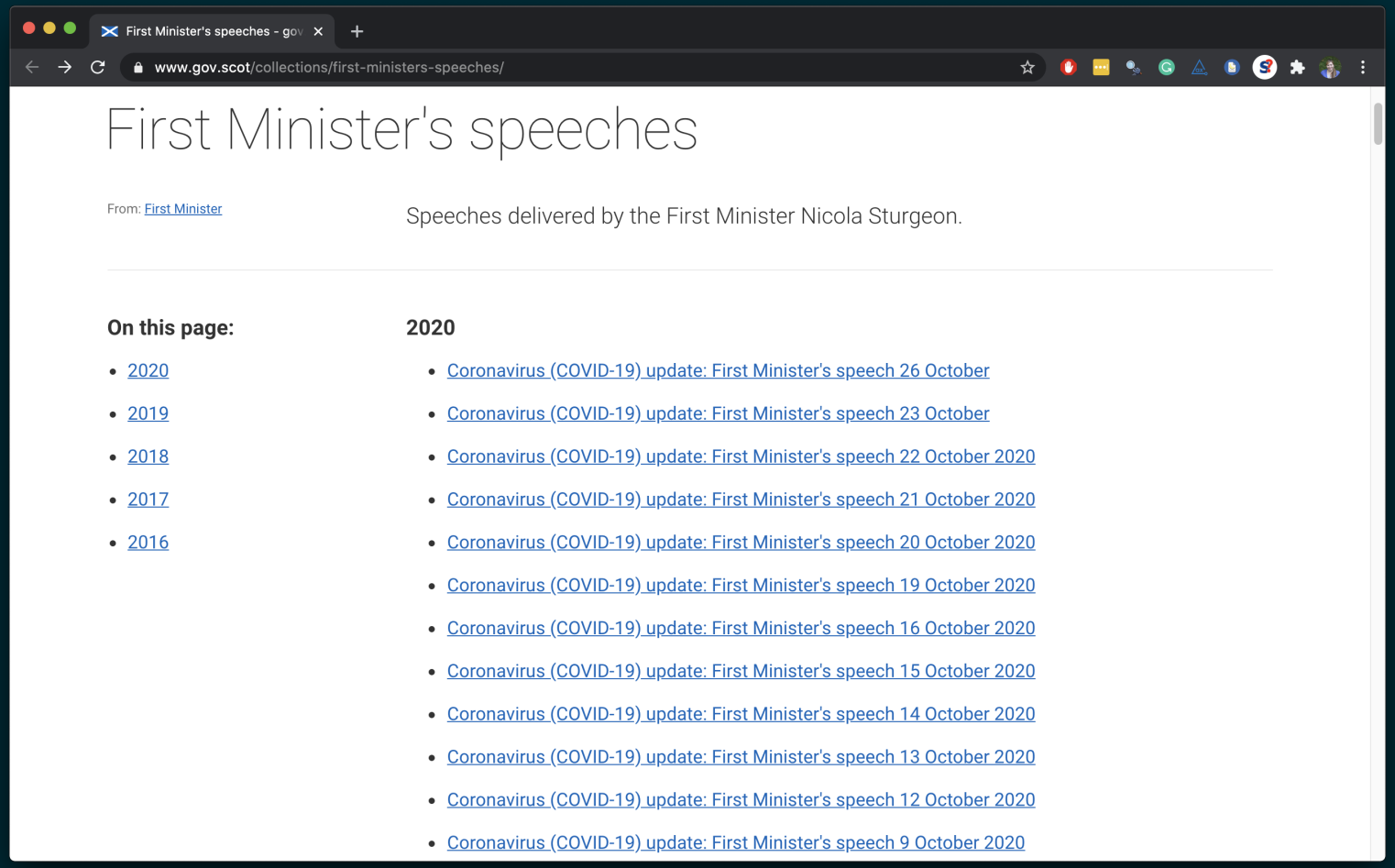
Data Science in a Box

datasciencebox.org

Modified by Tyler George

First Minister's COVID speeches

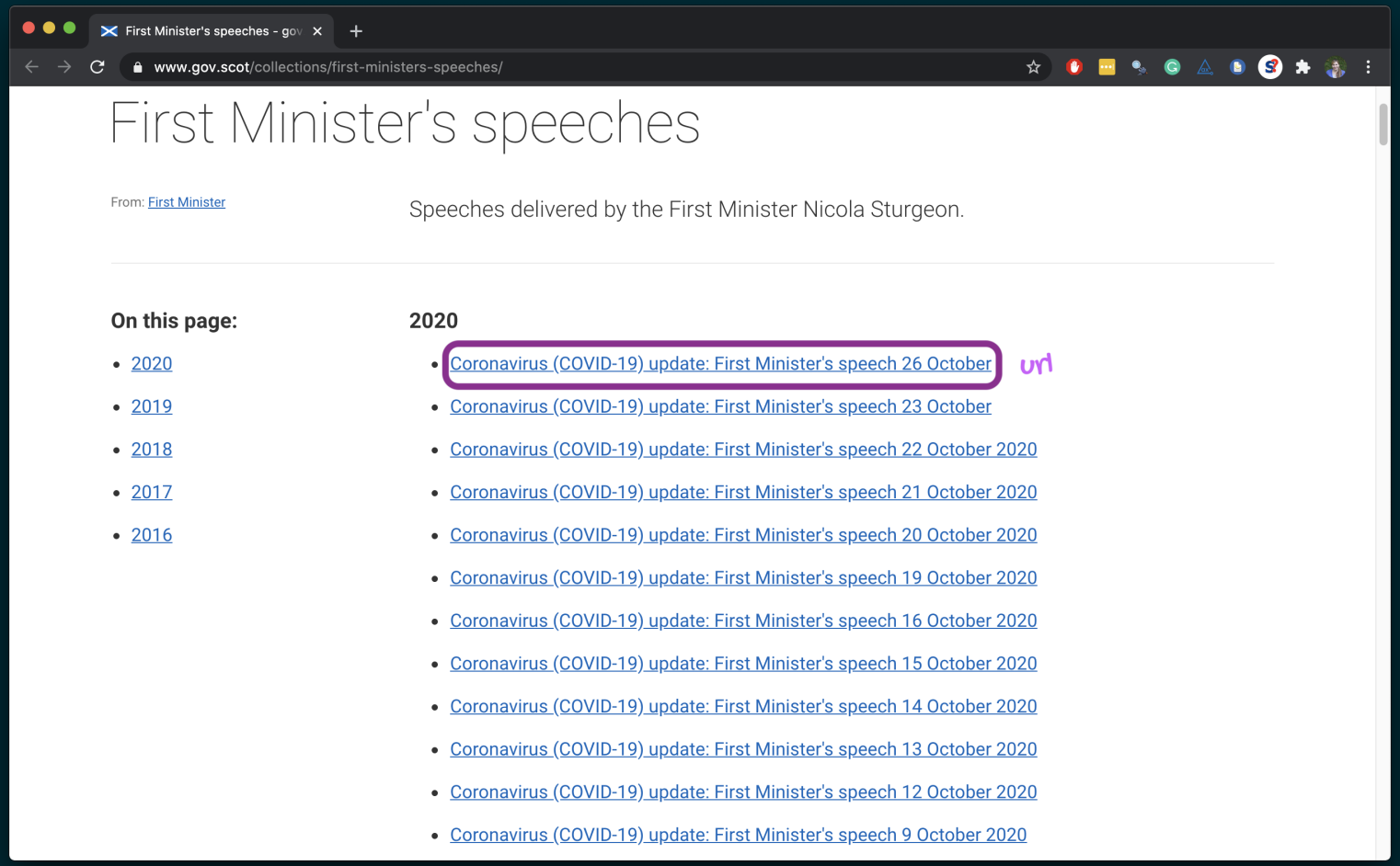
Start with

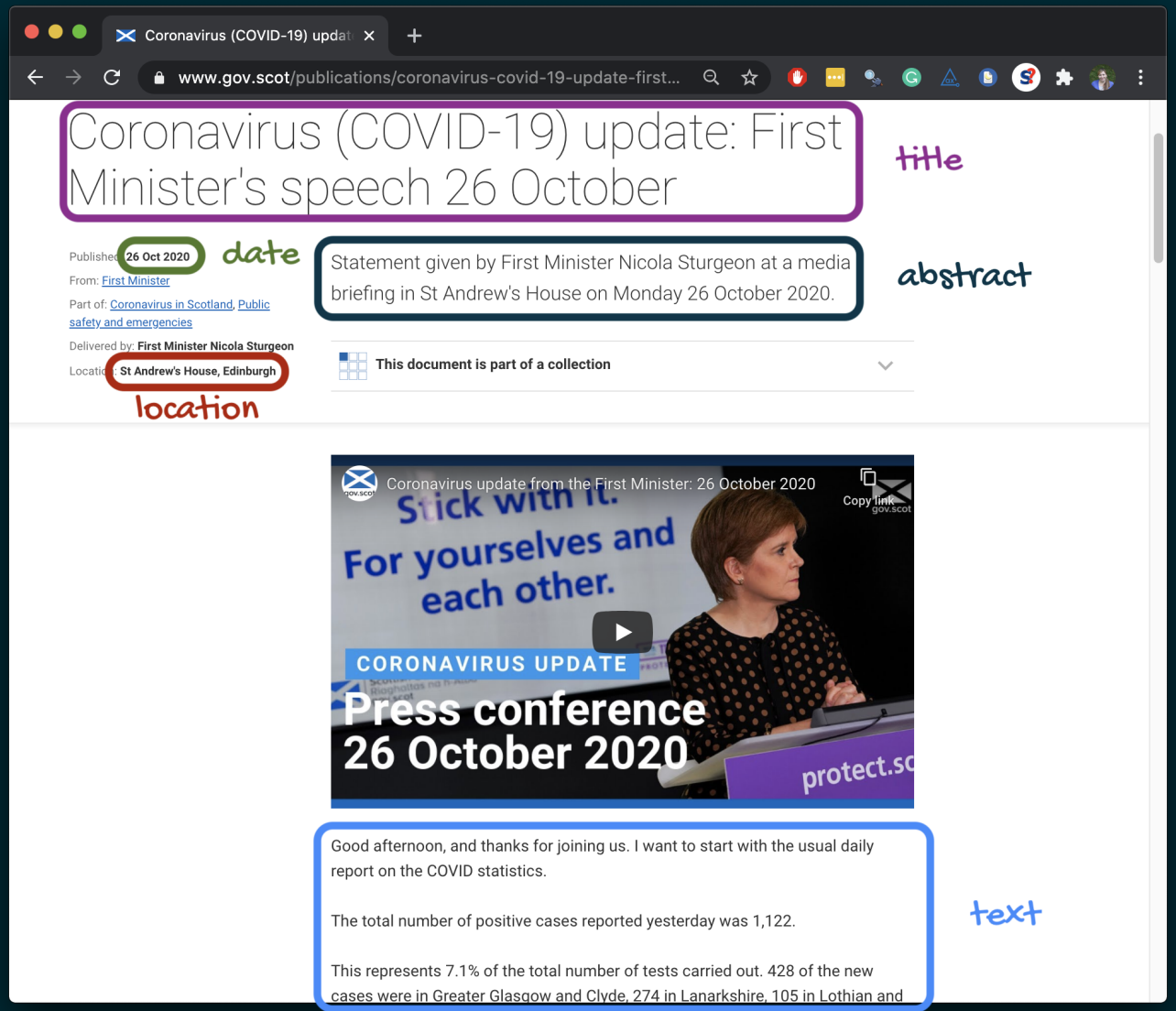

First Minister's speeches - gov
www.gov.scot/collections/first-ministers-speeches/
First Minister's speeches
From: First Minister
Speeches delivered by the First Minister Nicola Sturgeon.
On this page:
2020
2019
2018
2017
2016
2020
Coronavirus (COVID-19) update: First Minister's speech 26 October
Coronavirus (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 x 6
##    title     date       location   abstract     text     url
##    <chr>     <date>     <chr>      <chr>        <chr>    <chr>
##  1 Coronavi~ 2021-04-20 St Andrew~ Statement g~ "Good a~ https:/~
##  2 Coronavi~ 2021-04-13 St Andrew~ Statement g~ "Thanks~ https:/~
##  3 Coronavi~ 2021-04-06 St Andrew~ Statement g~ "Good a~ https:/~
##  4 Coronavi~ 2021-03-30 St Andrew~ Statement g~ "Thanks~ https:/~
##  5 Coronavi~ 2021-03-24 Scottish ~ Statement g~ "Thank ~ https:/~
##  6 Coronavi~ 2021-03-23 The Scott~ Statement g~ "Presid~ https:/~
##  7 Coronavi~ 2021-03-18 Scottish ~ Statement g~ "Thank ~ https:/~
##  8 Coronavi~ 2021-03-17 St Andrew~ Statement g~ "\nGood~ https:/~
##  9 Coronavi~ 2021-03-16 Scottish ~ Statement g~ "Presid~ https:/~
## 10 Coronavi~ 2021-03-15 St Andrew~ Statement g~ "\nGood~ https:/~
## 11 Coronavi~ 2021-03-11 Scottish ~ Statement g~ "I can ~ https:/~
## 12 Coronavi~ 2021-03-09 Scottish ~ Statement g~ "Presid~ https:/~
## 13 Coronavi~ 2021-03-05 Scottish ~ Parliamenta~ "Hello.~ https:/~
## 14 Coronavi~ 2021-03-04 Scottish ~ Parliamenta~ "I will~ https:/~
## 15 Coronavi~ 2021-03-02 Scottish ~ Statement g~ "Presid~ https:/~
## # ... with 203 more rows
```

www.gov.scot/collections/first-ministers-speeches




Coronavirus (COVID-19) upda
www.gov.scot/publications/coronavirus-covid-19-update-first...
Coronavirus (COVID-19) update: First Minister's speech 26 October
title
Published 26 Oct 2020
date
From: First Minister
Part of: Coronavirus in Scotland, Public safety and emergencies
Delivered by: First Minister Nicola Sturgeon
Location: St Andrew's House, Edinburgh
location
Statement given by First Minister Nicola Sturgeon at a media briefing in St Andrew's House on Monday 26 October 2020.
abstract
This document is part of a collection
Coronavirus update from the First Minister: 26 October 2020
Copy link
Stick with it.
For yourselves and each other.
CORONAVIRUS UPDATE
Press conference
26 October 2020
protect.sc
Good afternoon, and thanks for joining us. I want to start with the usual daily report on the COVID statistics.
The total number of positive cases reported yesterday was 1,122.
This represents 7.1% of the total number of tests carried out. 428 of the new cases were in Greater Glasgow and Clyde, 274 in Lanarkshire, 105 in Lothian and
text

Plan

1. Scrape `title`, `date`, `location`, `abstract`, and `text` from a few COVID-19 speech pages to develop the code

2. Write a function that scrapes `title`, `date`, `location`, `abstract`, and `text` from COVID-19 speech pages

3. Scrape the `urls` of COVID-19 speeches from the main page

4. Use this function to scrape from each individual COVID-19 speech from these `urls` and create a data frame with the columns `title`, `date`, `location`, `abstract`, `text`, and `url`

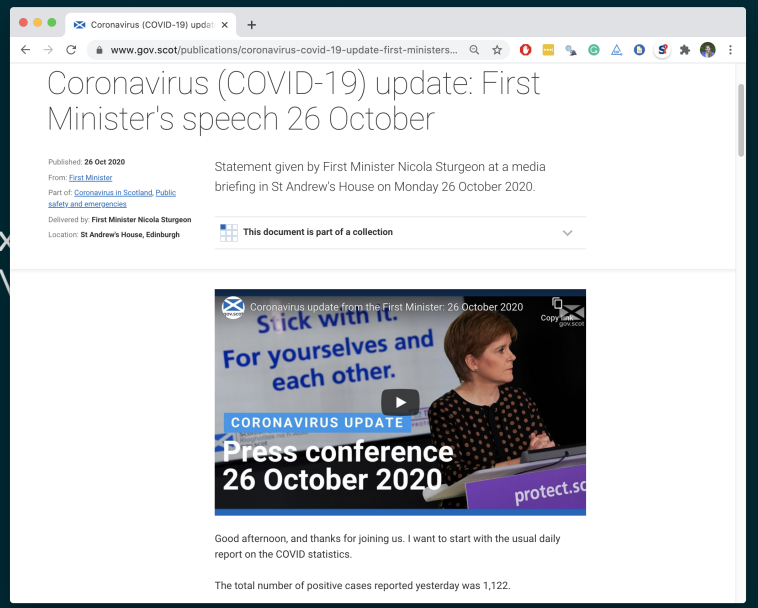
Scrape data from a few COVID-19 speech pages

Read page for 26 Oct speech

```
url <- "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-speech-26-oc
speech_page <- read_html(url)
```

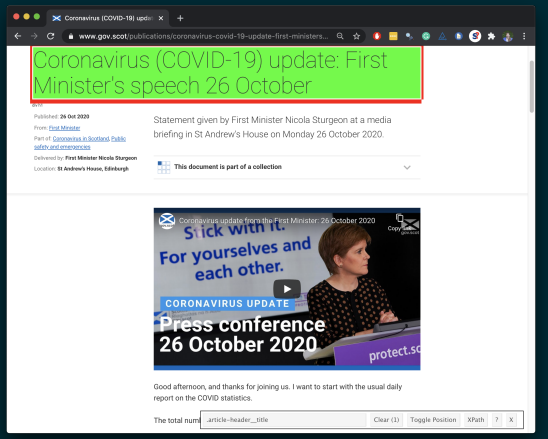
```
speech_page
```

```
## {html_document}
## <html dir="ltr" lang="en">
## [1] <head>\n<meta http-equiv="Content-Type" content="tex
## [2] <body class="fontawesome site-header__container">\n\
```



Extract title

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

title
```




```
## [1] "Coronavirus (COVID-19) update: First Minister's speech 26 October"
```

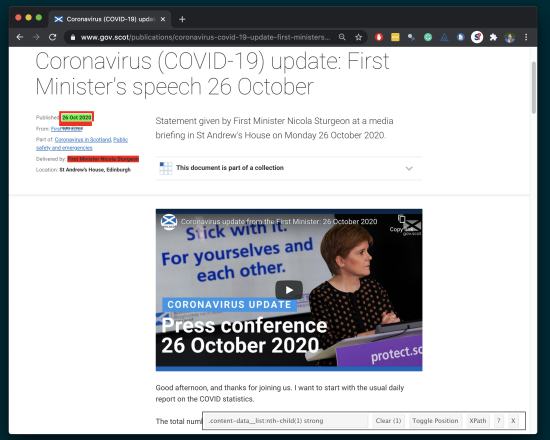
Extract date

```
library(lubridate)

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

```
## [1] "26 Oct 2020"
```

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

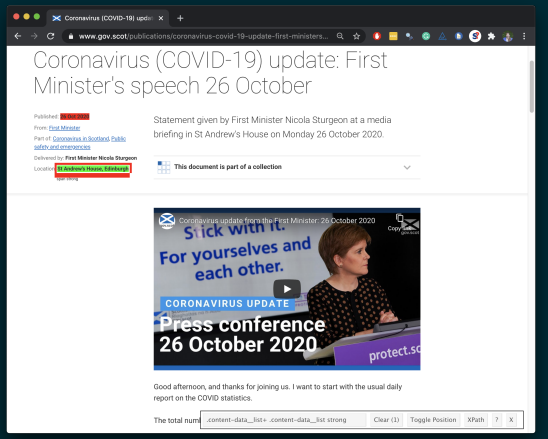
```
## [1] "2020-10-26"
```



Extract location

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

location
```

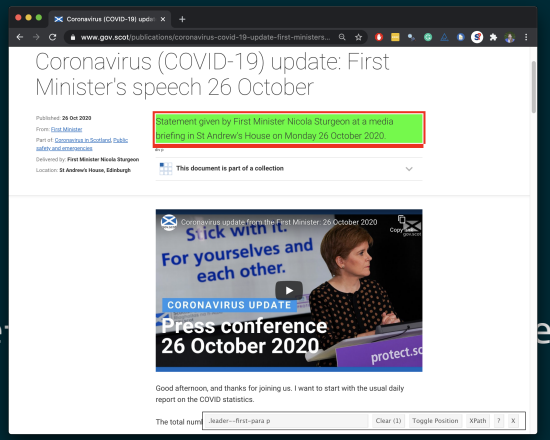
```
## [1] "St Andrew's House, Edinburgh"
```



Extract abstract

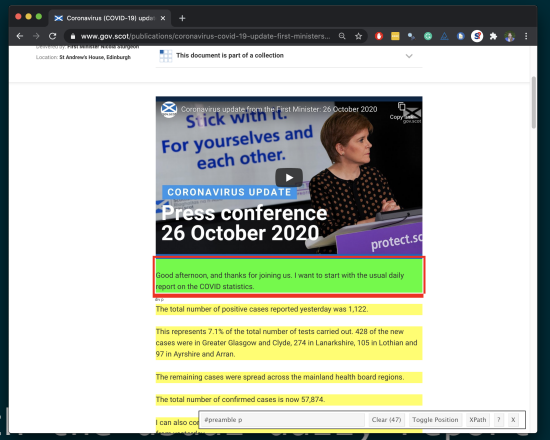
```
abstract <- speech_page %>%
    html_node(".leader--first-para p") %>%
    html_text()

abstract
```

```
## [1] "Statement given by First Minister Nicola Sturgeon at a media brie                e on Mon
```



Extract text

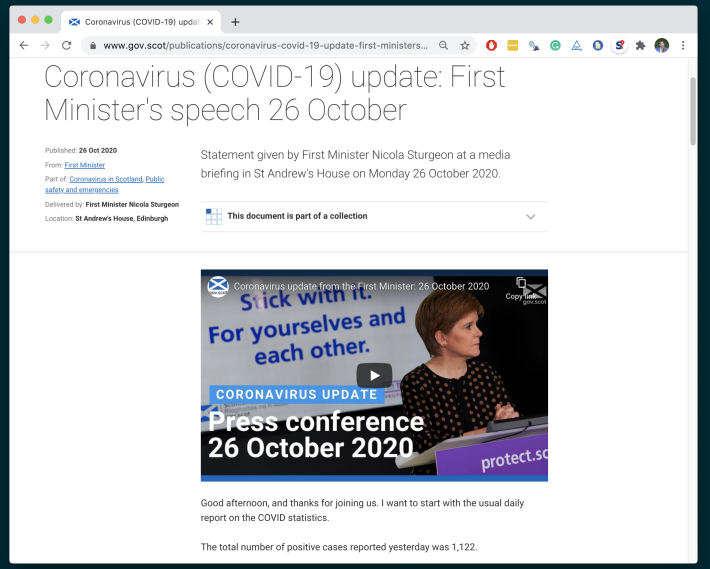
```
text <- speech_page %>%
    html_nodes("#preamble p") %>%
    html_text() %>%
    list()

text
```




```
## [[1]]
##  [1] "\nGood afternoon, and thanks for joining us. I want to start wit                    on the
##  [2] "The total number of positive cases reported yesterday was 1,122."
##  [3] "This represents 7.1% of the total number of tests carried out. 428 of the new cases were in Grea
##  [4] "The remaining cases were spread across the mainland health board regions. "
##  [5] "The total number of confirmed cases is now 57,874."
##  [6] "I can also confirm that 1,152 people are in hospital – that is an increase of 36 from yesterday
##  [7] "90 people are in intensive care, which is four more than yesterday."
##  [8] "And I regret to say that in the last 24 hours, one further death has been registered of a patien
##  [9] "We also reported 11 deaths on Saturday, and one yesterday.  So since the last briefing on F
## [10] "That reminds us again of how dangerous this virus can be and I want to send my condolences to ev
...
```

Put it all in a data frame

```
oct_26_speech <- tibble(
  title    = title,
  date     = date,
  location = location,
  abstract = abstract,
  text     = text,
  url      = url
)

oct_26_speech
```




```
## # A tibble: 1 x 6
##   title       date       location   abstract     text  url
##   <chr>       <date>     <chr>      <chr>        <lis> <chr>
## 1 Coronaviru~ 2020-10-26 St Andrew~ Statement g~ <chr~ https://w~
```

Read page for 23 Oct speech

```
url <- "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-speech-23-oc
speech_page <- read_html(url)
```

```
speech_page
```

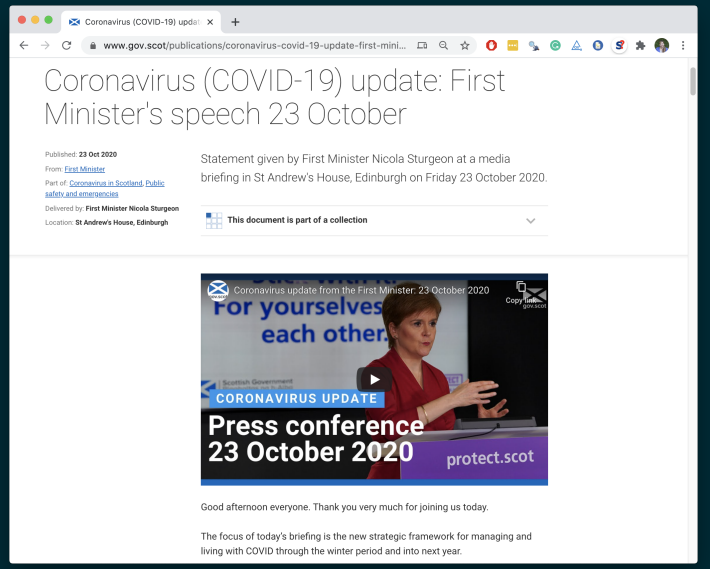
```
## {html_document}
## <html dir="ltr" lang="en">
## [1] <head>\n<meta http-equiv="Content-Type" content="text/html ...
## [2] <body class="fontawesome site-header__container">\n\n\n\n\ ...
```

Extract components of 23 Oct speech

```
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()
```

Put it all in a data frame

```
oct_23_speech <- tibble(
  title    = title,
  date     = date,
  location = location,
  abstract = abstract,
  text     = text,
  url      = url
)

oct_23_speech
```



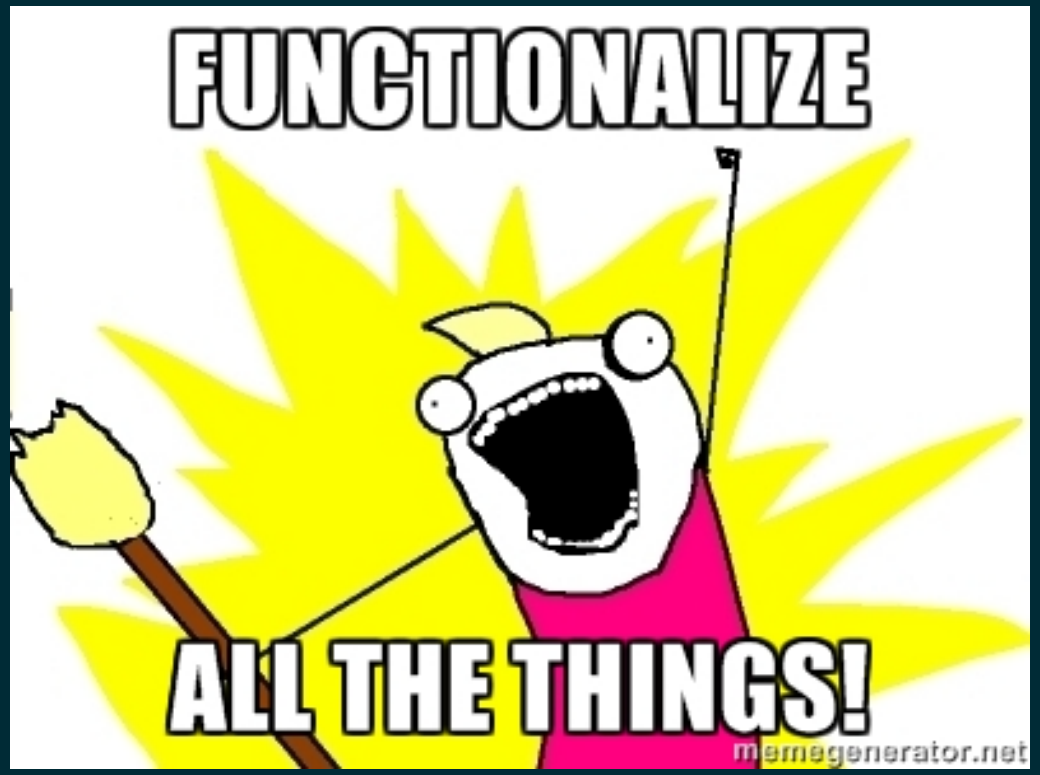

```
## # A tibble: 1 x 6
##   title       date       location   abstract     text  url
##   <chr>       <date>     <chr>      <chr>        <lis> <chr>
## 1 Coronaviru~ 2020-10-23 St Andrew~ Statement g~ <chr~ https://w~
```

this is getting tiring...

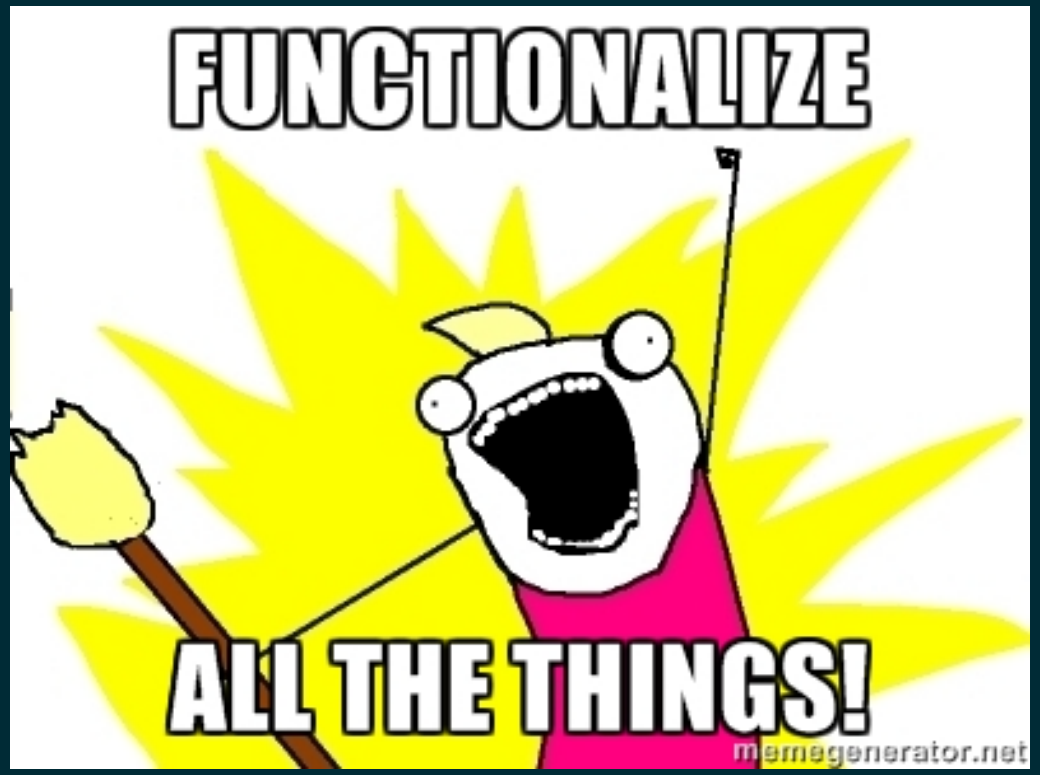
Functions

When should you write a function?

When should you write a function?

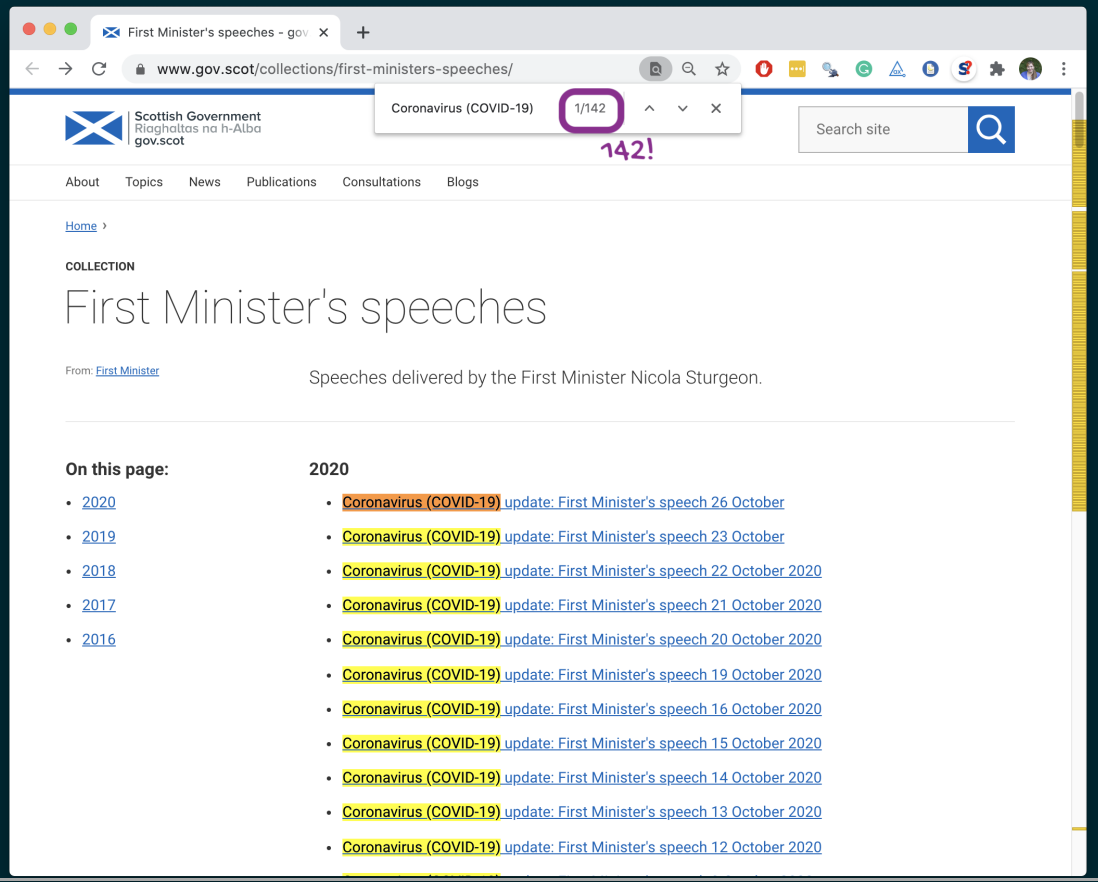


When should you write a function?



When you've copied and pasted a block of code more than twice.

How many times will we need to copy and paste the code we developed to scrape data on all of First Minister's COVID-19 speeches?



Why functions?

- Automate common tasks in a more powerful and general way than copy-and-pasting:
  - Give your function an evocative name that makes your code easier to understand
  - As requirements change, only need to update code in one place, instead of many
  - Eliminate chance of making incidental mistakes when you copy and paste (i.e. updating a variable name in one place, but not in another)

Why functions?

- Automate common tasks in a more powerful and general way than copy-and-pasting:
    - Give your function an evocative name that makes your code easier to understand
    - As requirements change, only need to update code in one place, instead of many
    - Eliminate chance of making incidental mistakes when you copy and paste (i.e. updating a variable name in one place, but not in another)
- Down the line: Improve your reach as a data scientist by writing functions (and packages!) that others use

Assuming that the page structure is the same for each speech page, how many "things" do you need to know for each speech page to scrape the data we want from it?

```
url_23_oct <- "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-speech-23-october/"
speech_page <- read_html(url_23_oct)

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
)
```

Turn your code into a function

- Pick a short but informative **name**, preferably a verb.

```
scrape_speech <-
```

Turn your code into a function

- Pick a short but evocative **name**, preferably a verb.
- List inputs, or **arguments**, to the function inside `function`. If we had more the call would look like `function(x, y, z)`.

```
scrape_speech <- function(x){



}
```

Turn your code into a function

- Pick a short but informative **name**, preferably a verb.
- List inputs, or **arguments**, to the function inside `function`. If we had more the call would look like `function(x, y, z)`.
- Place the **code** you have developed in body of the function, a `{` block that immediately follows `function(...)`.

```
scrape_speech <- function(url){

  # code we developed earlier to scrape info
  # on single art piece goes here

}
```

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
  )
}
```

Function in action

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

```
## Rows: 1
## Columns: 6
## $ title    <chr> "Coronavirus (COVID-19) update: First Minister~
## $ date     <date> 2020-10-26
## $ location <chr> "St Andrew's House, Edinburgh"
## $ abstract <chr> NA
## $ text     <list> <"\nGood afternoon, and thanks for joining us~
## $ url      <chr> "https://www.gov.scot/publications/coronavirus~
```

Function in action

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

```
## Rows: 1
## Columns: 6
## $ title    <chr> "Coronavirus (COVID-19) update: First Minister~
## $ date     <date> 2020-10-23
## $ location <chr> "St Andrew's House, Edinburgh"
## $ abstract <chr> NA
## $ text     <list> <"\nGood afternoon everyone. Thank you very m~
## $ url      <chr> "https://www.gov.scot/publications/coronavirus~
```

Function in action

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

```
## Rows: 1
## Columns: 6
## $ title    <chr> "Coronavirus (COVID-19) update: First Minister~
## $ date     <date> 2020-10-22
## $ location <chr> "St Andrew's House, Edinburgh"
## $ abstract <chr> NA
## $ text     <list> <"\nGood afternoon, let me start as usual wit~
## $ url      <chr> "https://www.gov.scot/publications/coronavirus~
```

Writing functions

What goes in / what comes out?

- They take input(s) defined in the function definition

```
function([inputs separated by commas]){
  # what to do with those inputs
}
```

- By default they return the last value computed in the function

```
scrape_page <- function(x){
  # do bunch of stuff with the input...

  # return a tibble
  tibble(...)
}
```

- You can define more outputs to be returned in a list as well as nice print methods (but we won't go there for now...)

What is going on here?

```
add_2 <- function(x){
  x + 2
  1000
}
```

```
add_2(3)
```

```
## [1] 1000
```

```
add_2(10)
```

```
## [1] 1000
```

Naming functions

> "There are only two hard things in Computer Science: cache invalidation and naming things." - Phil Karlton

Naming functions

- Names should be short but clearly evoke what the function does

Naming functions

- Names should be short but clearly evoke what the function does
- Names should be verbs, not nouns

Naming functions

- Names should be short but clearly evoke what the function does
- Names should be verbs, not nouns
- Multi-word names should be separated by underscores (`snake_case` as opposed to `camelCase`)

Naming functions

- Names should be short but clearly evoke what the function does
- Names should be verbs, not nouns
- Multi-word names should be separated by underscores (`snake_case` as opposed to `camelCase`)
- A family of functions should be named similarly (`scrape_page()`, `scrape_speech()` OR `str_remove()`, `str_replace()` etc.)

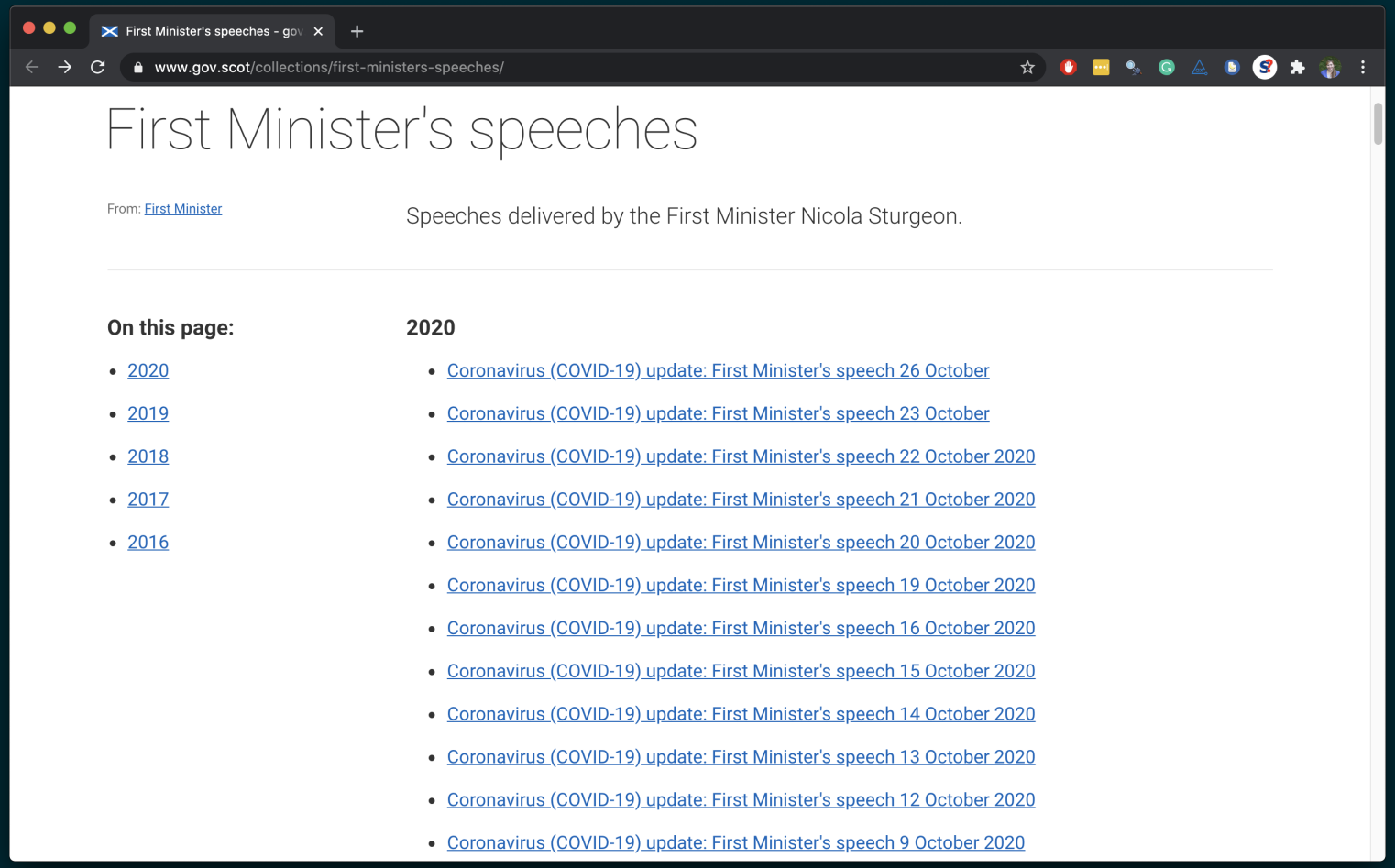
Naming functions

- Names should be short but clearly evoke what the function does
- Names should be verbs, not nouns
- Multi-word names should be separated by underscores (`snake_case` as opposed to `camelCase`)
- A family of functions should be named similarly (`scrape_page()`, `scrape_speech()` OR `str_remove()`, `str_replace()` etc.)
- Avoid overwriting existing (especially widely used) functions

```
# JUST DON'T
mean <- function(x){
  x * 3
  }
```

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 x 6
##    title     date       location   abstract     text     url
##    <chr>     <date>     <chr>      <chr>        <chr>    <chr>
##  1 Coronavi~ 2021-04-20 St Andrew~ Statement g~ "Good a~ https:/~
##  2 Coronavi~ 2021-04-13 St Andrew~ Statement g~ "Thanks~ https:/~
##  3 Coronavi~ 2021-04-06 St Andrew~ Statement g~ "Good a~ https:/~
##  4 Coronavi~ 2021-03-30 St Andrew~ Statement g~ "Thanks~ https:/~
##  5 Coronavi~ 2021-03-24 Scottish ~ Statement g~ "Thank ~ https:/~
##  6 Coronavi~ 2021-03-23 The Scott~ Statement g~ "Presid~ https:/~
##  7 Coronavi~ 2021-03-18 Scottish ~ Statement g~ "Thank ~ https:/~
##  8 Coronavi~ 2021-03-17 St Andrew~ Statement g~ "\nGood~ https:/~
##  9 Coronavi~ 2021-03-16 Scottish ~ Statement g~ "Presid~ https:/~
## 10 Coronavi~ 2021-03-15 St Andrew~ Statement g~ "\nGood~ https:/~
## 11 Coronavi~ 2021-03-11 Scottish ~ Statement g~ "I can ~ https:/~
## 12 Coronavi~ 2021-03-09 Scottish ~ Statement g~ "Presid~ https:/~
## 13 Coronavi~ 2021-03-05 Scottish ~ Parliamenta~ "Hello.~ https:/~
## 14 Coronavi~ 2021-03-04 Scottish ~ Parliamenta~ "I will~ https:/~
## 15 Coronavi~ 2021-03-02 Scottish ~ Statement g~ "Presid~ https:/~
## # ... 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-ministers-speec
scrape_speech(url = url_26_oct)
```

```
## # A tibble: 1 x 6
##   title        date       location    abstract text   url
##   <chr>        <date>     <chr>       <chr>    <list> <chr>
## 1 Coronavirus~ 2020-10-26 St Andrew'~ <NA>     <chr ~ https://ww~
```

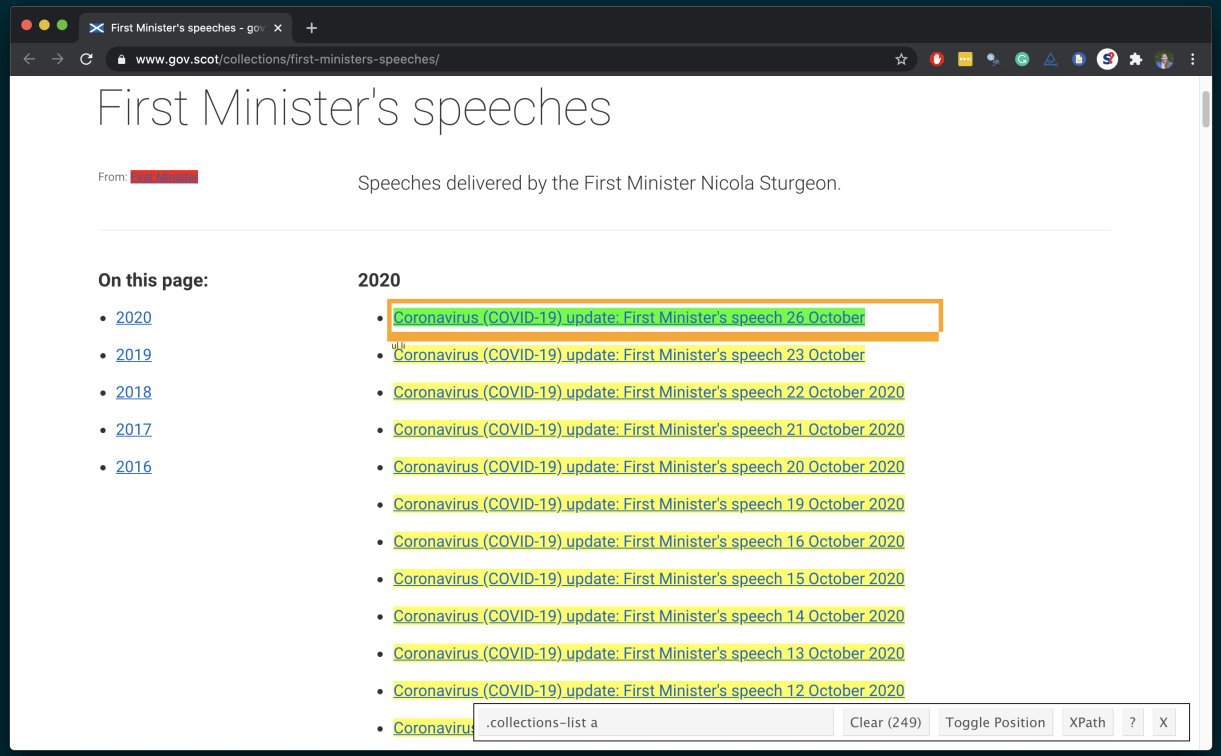
```
url_23_oct <- "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-speec
scrape_speech(url = url_23_oct)
```

```
## # A tibble: 1 x 6
##   title        date       location    abstract text   url
##   <chr>        <date>     <chr>       <chr>    <list> <chr>
## 1 Coronavirus~ 2020-10-23 St Andrew'~ <NA>     <chr ~ https://ww~
```

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/coronavirus-covid-19-update-first-ministers-statement-30-november-2021/"
##   [2] "/publications/coronavirus-covid-19-update-first-ministers-speech-29-november-2021/"
##   [3] "/publications/coronavirus-covid-19-update-first-ministers-statement-23-november-2021/"
##   [4] "/publications/first-ministers-statement-cop26/"
##   [5] "/publications/coronavirus-covid-19-update-first-ministers-statement-16-november-2021/"
##   [6] "/publications/first-minister-speech-43rd-t-b-macaulay-lecture/"
##   [7] "/publications/global-assembly-cop26-first-ministers-speech-1-november-2021/"
##   [8] "/publications/first-ministers-speech-before-start-cop26-1/"
##   [9] "/publications/coronavirus-covid-19-update-first-ministers-statement-26-october-2021/"
##  [10] "/publications/first-ministers-speech-scotlands-priorities-cop26/"
...
```

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-statement-30-november-2021/"
##   [2] "/publications/coronavirus-covid-19-update-first-ministers-speech-29-november-2021/"
##   [3] "/publications/coronavirus-covid-19-update-first-ministers-statement-23-november-2021/"
##   [4] "/publications/coronavirus-covid-19-update-first-ministers-statement-16-november-2021/"
##   [5] "/publications/coronavirus-covid-19-update-first-ministers-statement-26-october-2021/"
##   [6] "/publications/coronavirus-covid-19-update-first-ministers-statement-5-october-2021/"
##   [7] "/publications/coronavirus-covid-19-update-first-ministers-statement-28-september-2021/"
##   [8] "/publications/coronavirus-covid-19-update-first-ministers-statement-21-september-2021/"
##   [9] "/publications/coronavirus-covid-19-update-first-ministers-statement-14-september-2021/"
##  [10] "/publications/coronavirus-covid-19-update-first-ministers-statement-8-september-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-statement-30-nove
##   [2] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-speech-29-novembe
##   [3] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-statement-23-nove
##   [4] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-statement-16-nove
##   [5] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-statement-26-octo
##   [6] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-statement-5-octob
##   [7] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-statement-28-sept
##   [8] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-statement-21-sept
##   [9] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-statement-14-sept
##  [10] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-statement-8-septe
...
```

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-statement-30-nove
##   [2] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-speech-29-novembe
##   [3] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-statement-23-nove
##   [4] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-statement-16-nove
##   [5] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-statement-26-octo
##   [6] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-statement-5-octob
##   [7] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-statement-28-sept
##   [8] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-statement-21-sept
##   [9] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-statement-14-sept
##  [10] "https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-statement-8-septe
...
```

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`?

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.

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.

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)
)
```

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 x 6
##    title     date       location   abstract     text     url
##    <chr>     <date>     <chr>      <chr>        <chr>    <chr>
##  1 Coronavi~ 2021-04-20 St Andrew~ Statement g~ "Good a~ https:/~
##  2 Coronavi~ 2021-04-13 St Andrew~ Statement g~ "Thanks~ https:/~
##  3 Coronavi~ 2021-04-06 St Andrew~ Statement g~ "Good a~ https:/~
##  4 Coronavi~ 2021-03-30 St Andrew~ Statement g~ "Thanks~ https:/~
##  5 Coronavi~ 2021-03-24 Scottish ~ Statement g~ "Thank ~ https:/~
##  6 Coronavi~ 2021-03-23 The Scott~ Statement g~ "Presid~ https:/~
##  7 Coronavi~ 2021-03-18 Scottish ~ Statement g~ "Thank ~ https:/~
##  8 Coronavi~ 2021-03-17 St Andrew~ Statement g~ "\nGood~ https:/~
##  9 Coronavi~ 2021-03-16 Scottish ~ Statement g~ "Presid~ https:/~
## 10 Coronavi~ 2021-03-15 St Andrew~ Statement g~ "\nGood~ https:/~
## 11 Coronavi~ 2021-03-11 Scottish ~ Statement g~ "I can ~ https:/~
## 12 Coronavi~ 2021-03-09 Scottish ~ Statement g~ "Presid~ https:/~
## 13 Coronavi~ 2021-03-05 Scottish ~ Parliamenta~ "Hello.~ https:/~
## 14 Coronavi~ 2021-03-04 Scottish ~ Parliamenta~ "I will~ https:/~
## 15 Coronavi~ 2021-03-02 Scottish ~ Statement g~ "Presid~ https:/~
## # ... 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...
}
```