Wiki Web Scrape

Mark Gallo

- 1. Get HTML
- 2. Get the info box
- 3. Make a data frame
- 4. Make a function

```
library(rvest)
```

```
## Loading required package: xml2
```

```
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(purrr)
```

```
##
## Attaching package: 'purrr'
```

```
## The following object is masked from 'package:rvest':
##
##     pluck
```

```
library(XML)
```

```
##
## Attaching package: 'XML'
```

```
## The following object is masked from 'package:rvest':
##
##     xml
```

1. Get HTML

```
mitch_url <- "https://en.wikipedia.org/wiki/Mitch_McConnell"

mitchWiki <- read_html(mitch_url)
```

2. Get the info box

```
table_node <- html_node(mitchWiki, css = "table")
table_node
```

```
## {html_node}
## <table class="infobox vcard" style="width:22em">
## [1] <tbody>\n<tr><th colspan="2" style="text-align:center;font-size:125% ...
```

3. Make a data frame

```
mitchData <- html_table(table_node, header = FALSE)


names(mitchData)[1]<-"Key"
names(mitchData)[2]<-"Value"
```

```
mitchData[1,1] = c("Full Name")
mitchData <- mitchData[c(1,47,49),]
data.frame(mitchData)
```

```
##                Key           Value
## 1        Full Name Mitch McConnell
## 47 Political party      Republican
## 49        Children               3
```

4. Make a function

a.

```
get_wiki_info <- function(u = "https://en.wikipedia.org/wiki/Mitch_McConnell"){
  wiki <- read_html(u)
  table_node <- html_node(wiki, css = "table")
  Data <- html_table(table_node, header = FALSE)


names(Data)[1]<-"Key"
names(Data)[2]<-"Value"
Data[1,1] = "Full Name"


polData <- filter(Data, Key %in% c("Full Name", "Political party","Children"))

if(polData[1,2] == "Susan Collins"){
  polData[3,1] = "Children"
  polData[3,2] = NA


}
 print(polData)
}
```

b.

```
get_wiki_info('https://en.wikipedia.org/wiki/Tammy_Duckworth')
```

```
##               Key           Value
## 1       Full Name Tammy Duckworth
## 2 Political party      Democratic
## 3        Children               2
```

```
get_wiki_info("https://en.wikipedia.org/wiki/Susan_Collins")
```

```
##               Key         Value
## 1       Full Name Susan Collins
## 2 Political party    Republican
## 3        Children          <NA>
```