Class 7: Web Scraping in R

MAST5953: Web Scraping and Text Mining

Dr. Miriam Sorace

www.miriamsorace.net

25 January 2023

Outline of Today's Class

Web Scraping Recap

Web Scraping Methods

Scraping in R: Core Functions w. Examples

Web Scraping Recap

Web Scraping

- Scrapers: programmes that grab *specific* content from webpages
- Spiders (or Crawlers): programmes that index entire pages and follow every link.
- Web scraping involves some spidering when - for example - parsing entire web-pages. But scrapers are *selective/purposive*

Web Scraping: Managing Expectations

- It is rarely a one-shot affair: websites change constantly
- Difficult to generalise: each website is different
- Writing a scraping code that works well takes a long time

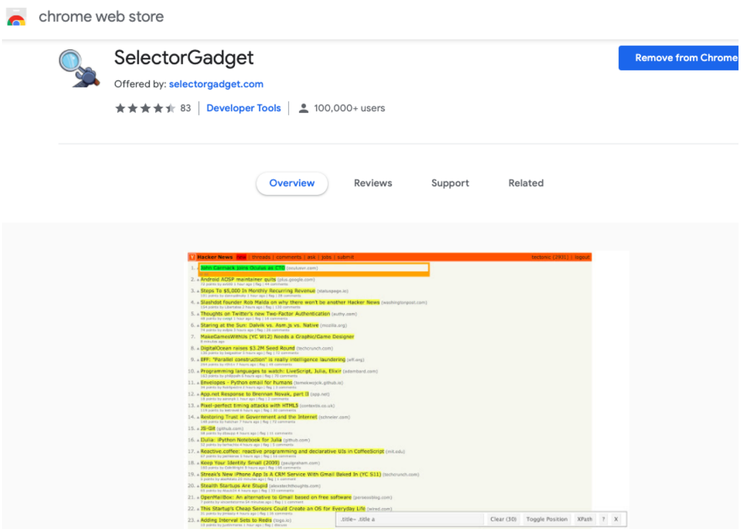
Web Scraping: Netiquette!

- Search for & respect terms of service
  - add "/robot.txt" at the end of URL
- Careful with personal/private data
- Check if there are APIs or ready-to-download files already
- Stay identifiable (do not mask your IP)
- Reduce traffic as much as possible
  - light/efficient scrapers
  - use pauses (`Sys.sleep()`)
  - avoid over-scraping

Web Scraping Steps

1. Analyse the structure of the webpage (e.g. HTML/XML trees)
    - Identify relevant nodes
    - www.selectorgadget.com
2. Test: write scraper and debug
3. Execute the download

First Task: Install SelectorGadget!



chrome web store

SelectorGadget

Offered by: selectorgadget.com

★★★★½ 83 | Developer Tools | 100,000+ users

Remove from Chrome

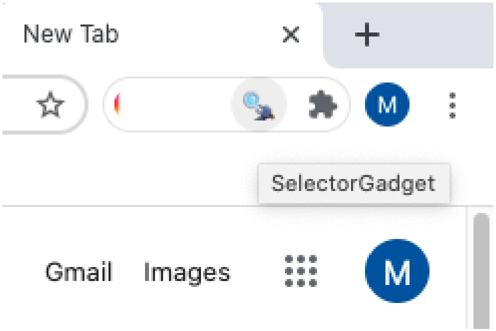
Overview Reviews Support Related

Hacker News new | threads | comments | ask | jobs | submit tectonic (2931) | logout

1. John Carmack joins Oculus as CTO (oculusvr.com)
2. Android AOSP maintainer quits (plus.google.com)
3. Steps To $5,000 In Monthly Recurring Revenue (statuspage.io)
4. Slashdot founder Rob Malda on why there won't be another Hacker News (washingtonpost.com)
5. Thoughts on Twitter's new Two-Factor Authentication (authy.com)
6. Staring at the Sun: Dalvik vs. Asm.js vs. Native (mozilla.org)
7. MakeGamesWithUs (YC W12) Needs a Graphic/Game Designer
8. DigitalOcean raises $3.2M Seed Round (techcrunch.com)
9. EFF: "Parallel construction" is really intelligence laundering (eff.org)
10. Programming languages to watch: LiveScript, Julia, Elixir (adambard.com)
11. Envelopes - Python email for humans (tomekwojcik.github.io)
12. App.net Response to Brennan Novak, part II (app.net)
13. Pixel-perfect timing attacks with HTML5 (contextis.co.uk)
14. Restoring Trust in Government and the Internet (schneier.com)
15. JS-Git (github.com)
16. IJulia: iPython Notebook for Julia (github.com)
17. Reactive.coffee: reactive programming and declarative UIs in CoffeeScript (mit.edu)
18. Keep Your Identity Small (2009) (paulgraham.com)
19. Streak's New iPhone App Is A CRM Service With Gmail Baked In (YC S11) (techcrunch.com)
20. Stealth Startups Are Stupid (alexstechthoughts.com)
21. OpenMailBox: An alternative to Gmail based on free software (persoosblog.com)
22. This Startup's Cheap Sensors Could Create an OS for Everyday Life (wired.com)
23. Adding Interval Sets to Redis (togo.io)

.title~ .title a

First Task: Install SelectorGadget!



Web Scraping Methods

Extraction Methods

1. Regular Expressions
2. Node queries
3. APIs (*Application Programming Interfaces*)
   - We'll explore them in the next class on social media data

Regular Expressions

- Scraper works by extracting from meaningful alphanumeric patterns in the webpage text
  - E.g.: grab all text in-between quotes (“”).

Regular Expressions

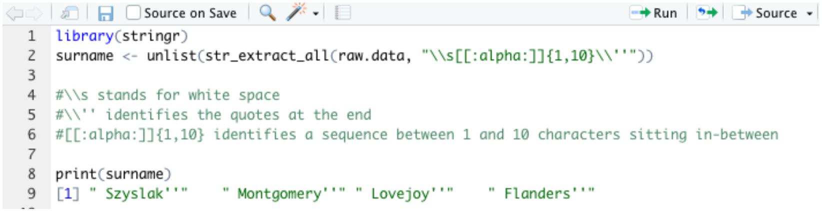
```
[1] "Moe Szyslak"          "Burns, C. Montgomery" "Rev. Timothy Lovejoy"
[4] "Ned Flanders"
```

- Suppose from the web content above, you wanted to extract the surnames
- You can exploit the regular pattern of surnames sitting between a space and closing with ”

Regular Expressions

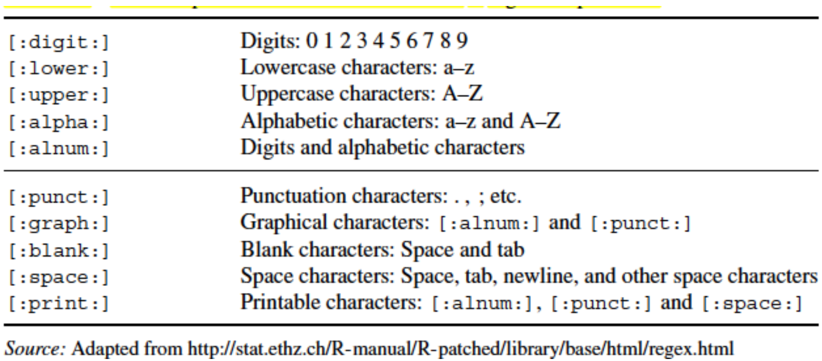
```
[1] "Moe Szyslak"       "Burns, C. Montgomery" "Rev. Timothy Lovejoy"
[4] "Ned Flanders"
```

- You would write a regular expression scraper as the below



```
library(stringr)
surname <- unlist(str_extract_all(raw.data, "\\s[[:alpha:]]{1,10}\\''"))

#\\s stands for white space
#\\'' identifies the quotes at the end
#[[:alpha:]]{1,10} identifies a sequence between 1 and 10 characters sitting in-between

print(surname)
[1] " Szyslak''"    " Montgomery''" " Lovejoy''"    " Flanders''"
```

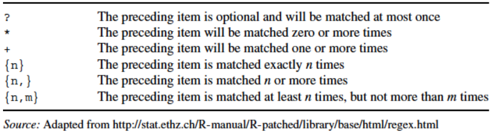
Regular Expressions



`[:digit:]`	Digits: 0 1 2 3 4 5 6 7 8 9
`[:lower:]`	Lowercase characters: a–z
`[:upper:]`	Uppercase characters: A–Z
`[:alpha:]`	Alphabetic characters: a–z and A–Z
`[:alnum:]`	Digits and alphabetic characters
`[:punct:]`	Punctuation characters: . , ; etc.
`[:graph:]`	Graphical characters: `[:alnum:]` and `[:punct:]`
`[:blank:]`	Blank characters: Space and tab
`[:space:]`	Space characters: Space, tab, newline, and other space characters
`[:print:]`	Printable characters: `[:alnum:]`, `[:punct:]` and `[:space:]`

Source: Adapted from http://stat.ethz.ch/R-manual/R-patched/library/base/html/regex.html

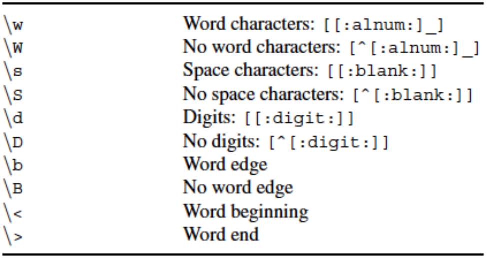
Regular Expressions



`?`	The preceding item is optional and will be matched at most once
`*`	The preceding item will be matched zero or more times
`+`	The preceding item will be matched one or more times
`{n}`	The preceding item is matched exactly *n* times
`{n,}`	The preceding item is matched *n* or more times
`{n,m}`	The preceding item is matched at least *n* times, but not more than *m* times

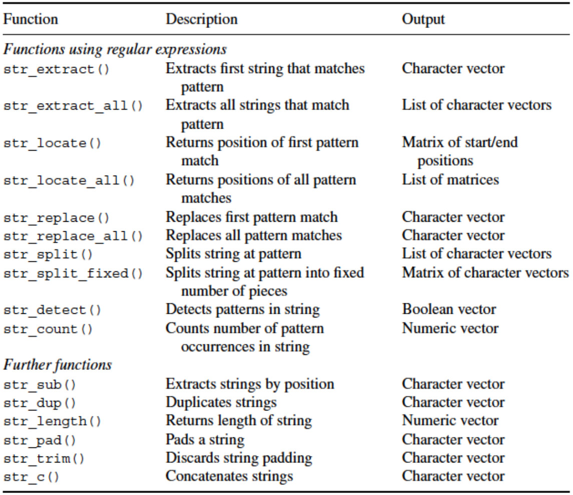
Source: Adapted from http://stat.ethz.ch/R-manual/R-patched/library/base/html/regex.html

Table 8.3 Selected symbols with special meaning



`\w`	Word characters: `[[:alnum:]_]`
`\W`	No word characters: `[^[:alnum:]_]`
`\s`	Space characters: `[[:blank:]]`
`\S`	No space characters: `[^[:blank:]]`
`\d`	Digits: `[[:digit:]]`
`\D`	No digits: `[^[:digit:]]`
`\b`	Word edge
`\B`	No word edge
`\<`	Word beginning
`\>`	Word end

Regular Expressions



Function	Description	Output
Functions using regular expressions		
`str_extract()`	Extracts first string that matches pattern	Character vector
`str_extract_all()`	Extracts all strings that match pattern	List of character vectors
`str_locate()`	Returns position of first pattern match	Matrix of start/end positions
`str_locate_all()`	Returns positions of all pattern matches	List of matrices
`str_replace()`	Replaces first pattern match	Character vector
`str_replace_all()`	Replaces all pattern matches	Character vector
`str_split()`	Splits string at pattern	List of character vectors
`str_split_fixed()`	Splits string at pattern into fixed number of pieces	Matrix of character vectors
`str_detect()`	Detects patterns in string	Boolean vector
`str_count()`	Counts number of pattern occurrences in string	Numeric vector
Further functions		
`str_sub()`	Extracts strings by position	Character vector
`str_dup()`	Duplicates strings	Character vector
`str_length()`	Returns length of string	Numeric vector
`str_pad()`	Pads a string	Character vector
`str_trim()`	Discards string padding	Character vector
`str_c()`	Concatenates strings	Character vector

Regular Expressions
Advantages

- Only useful when HTML/XML are malformed **and** clear patterns are retrievable in the page
- Mostly used in creating URL lists to scrape multiple pages
- Mostly used in data cleaning/variable creation

Regular Expressions

Disadvantages

- Building and interpreting a regular expression scraper is challenging: hard to test and debug
- Limited applicability: not all web pages have consistent repeated patterns that *uniquely* identify a specific content
- Ignores the useful and meaningful hierarchical structure of web pages
- Strongly discouraged for webscraping, best for data cleaning

Node Queries

- This is the most common procedure
- The scraper is built by exploiting the tree structure of the web page
  - Inspected via web developer tools & SelectorGadget
- Utilises query functions from R library such as `rvest` to locate nodes and extract information

Node Queries
Advantages

- Intuitive to write and read
- Uniquely identifies the relevant content

Node Queries

Disadvantages

- Costly when the HTML/XML structure is malformed
- Costly when web pages get constantly updated

Application Programming Interfaces

- Scraper is built by using APIs ‘wrappers’, exploiting the webpage’s API, which provides the website’s data in pure/cleaned form
- We’ll see concrete examples of this in the next class where we’ll look at social media data

Application Programming Interfaces

Advantages

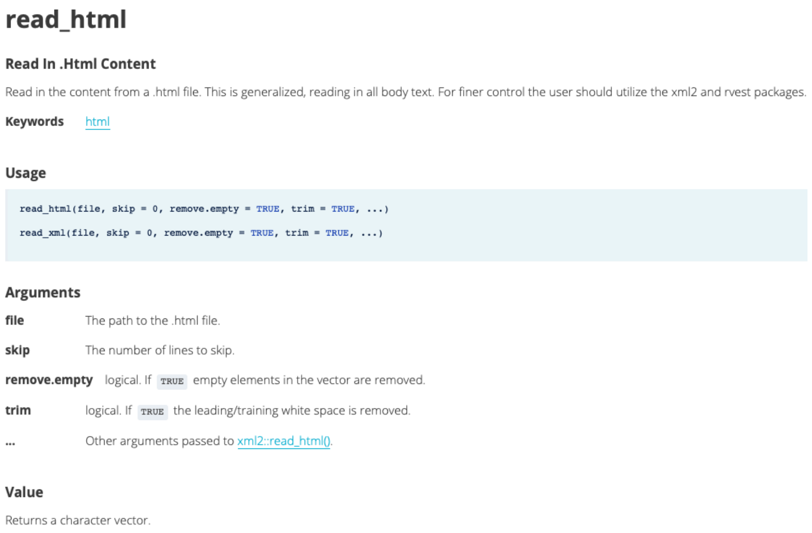
- This is the scraping gold standard.
- Very easy: clean, standardised data
- APIs codes are robust and regularly updated

Application Programming Interfaces
Disadvantages

- Not many websites provide their APIs
- Websites that do, often have licences/fees/limits attached to API use
- API owners often limit their functionality from one day to the next, rendering the scraper useless.

Scraping in R: Core Functions w. Examples

Parsing the Page



read_html

Read In .Html Content

Read in the content from a .html file. This is generalized, reading in all body text. For finer control the user should utilize the xml2 and rvest packages.

Keywords html

Usage

```
read_html(file, skip = 0, remove.empty = TRUE, trim = TRUE, ...)

read_xml(file, skip = 0, remove.empty = TRUE, trim = TRUE, ...)
```

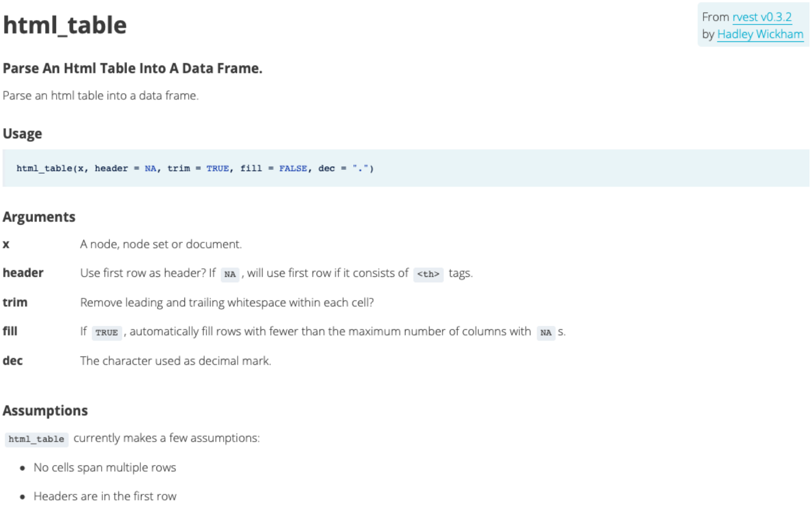
Arguments

file	The path to the .html file.
skip	The number of lines to skip.
remove.empty	logical. If `TRUE` empty elements in the vector are removed.
trim	logical. If `TRUE` the leading/training white space is removed.
...	Other arguments passed to xml2::read_html().

Value

Returns a character vector.

Extracting Tables



html_table

From rvest v0.3.2
by Hadley Wickham

Parse An Html Table Into A Data Frame.

Parse an html table into a data frame.

Usage

```
html_table(x, header = NA, trim = TRUE, fill = FALSE, dec = ".")
```

Arguments

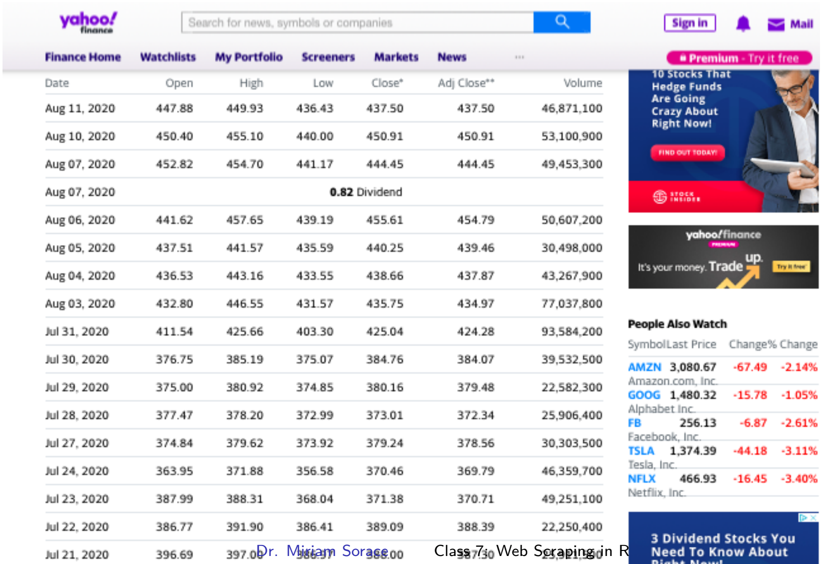
x	A node, node set or document.
header	Use first row as header? If `NA`, will use first row if it consists of `<th>` tags.
trim	Remove leading and trailing whitespace within each cell?
fill	If `TRUE`, automatically fill rows with fewer than the maximum number of columns with `NA` s.
dec	The character used as decimal mark.

Assumptions

`html_table` currently makes a few assumptions:

- No cells span multiple rows
- Headers are in the first row

Extracting Tables
Example


yahoo! finance
Search for news, symbols or companies
Sign in
Mail
Finance Home
Watchlists
My Portfolio
Screeners
Markets
News
Premium - Try it free
10 Stocks That Hedge Funds Are Going Crazy About Right Now!
FIND OUT TODAY!
STOCK INSIDER
yahoo/finance PREMIUM
It's your money. Trade up.
Try it free
3 Dividend Stocks You Need To Know About Right Now!

Date	Open	High	Low	Close*	Adj Close**	Volume
Aug 11, 2020	447.88	449.93	436.43	437.50	437.50	46,871,100
Aug 10, 2020	450.40	455.10	440.00	450.91	450.91	53,100,900
Aug 07, 2020	452.82	454.70	441.17	444.45	444.45	49,453,300
Aug 07, 2020				0.82 Dividend		
Aug 06, 2020	441.62	457.65	439.19	455.61	454.79	50,607,200
Aug 05, 2020	437.51	441.57	435.59	440.25	439.46	30,498,000
Aug 04, 2020	436.53	443.16	433.55	438.66	437.87	43,267,900
Aug 03, 2020	432.80	446.55	431.57	435.75	434.97	77,037,800
Jul 31, 2020	411.54	425.66	403.30	425.04	424.28	93,584,200
Jul 30, 2020	376.75	385.19	375.07	384.76	384.07	39,532,500
Jul 29, 2020	375.00	380.92	374.85	380.16	379.48	22,582,300
Jul 28, 2020	377.47	378.20	372.99	373.01	372.34	25,906,400
Jul 27, 2020	374.84	379.62	373.92	379.24	378.56	30,303,500
Jul 24, 2020	363.95	371.88	356.58	370.46	369.79	46,359,700
Jul 23, 2020	387.99	388.31	368.04	371.38	370.71	49,251,100
Jul 22, 2020	386.77	391.90	386.41	389.09	388.39	22,250,400
Jul 21, 2020	396.69	[illegible]	[illegible]	388.00	387.30	[illegible]

People Also Watch

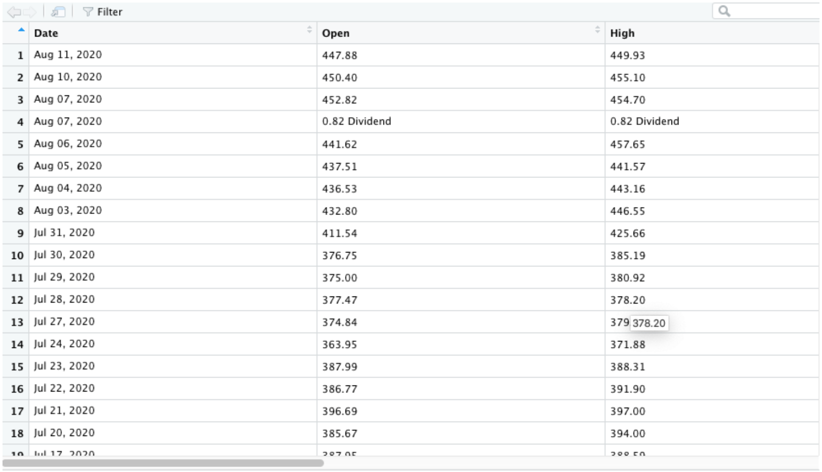
Symbol	Last Price	Change	% Change
AMZN Amazon.com, Inc.	3,080.67	-67.49	-2.14%
GOOG Alphabet Inc.	1,480.32	-15.78	-1.05%
FB Facebook, Inc.	256.13	-6.87	-2.61%
TSLA Tesla, Inc.	1,374.39	-44.18	-3.11%
NFLX Netflix, Inc.	466.93	-16.45	-3.40%

Extracting Tables
Example

- `url<-"https://finance.yahoo.com/quote/AAPL/history?p=AAPL"`
- `apple_finance<- read_html(url)`
- `tableaslist<-html_table(apple_finance)`
- `table<-as.data.frame(tableaslist)`

Extracting Tables

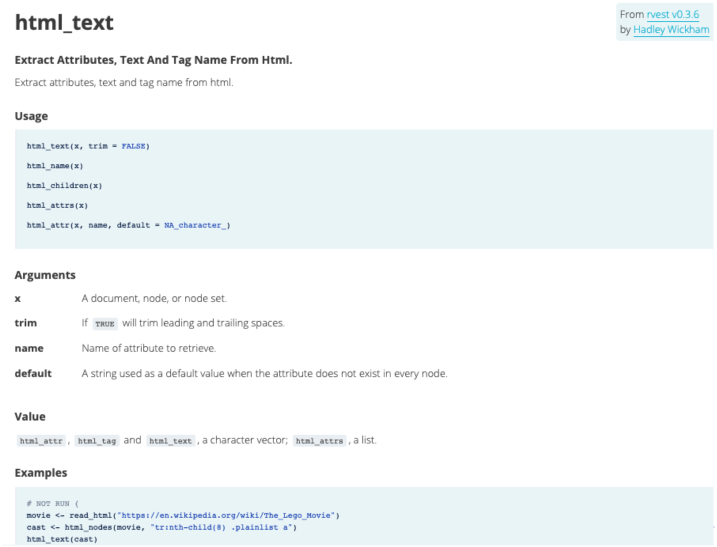
Example



Filter

	Date	Open	High
1	Aug 11, 2020	447.88	449.93
2	Aug 10, 2020	450.40	455.10
3	Aug 07, 2020	452.82	454.70
4	Aug 07, 2020	0.82 Dividend	0.82 Dividend
5	Aug 06, 2020	441.62	457.65
6	Aug 05, 2020	437.51	441.57
7	Aug 04, 2020	436.53	443.16
8	Aug 03, 2020	432.80	446.55
9	Jul 31, 2020	411.54	425.66
10	Jul 30, 2020	376.75	385.19
11	Jul 29, 2020	375.00	380.92
12	Jul 28, 2020	377.47	378.20
13	Jul 27, 2020	374.84	379 378.20
14	Jul 24, 2020	363.95	371.88
15	Jul 23, 2020	387.99	388.31
16	Jul 22, 2020	386.77	391.90
17	Jul 21, 2020	396.69	397.00
18	Jul 20, 2020	385.67	394.00
19	Jul 17, 2020	387.95	388.50

Extracting Text



html_text

From rvest v0.3.6
by Hadley Wickham

Extract Attributes, Text And Tag Name From Html.

Extract attributes, text and tag name from html.

Usage

```
html_text(x, trim = FALSE)

html_name(x)

html_children(x)

html_attrs(x)

html_attr(x, name, default = NA_character_)
```

Arguments

x	A document, node, or node set.
trim	If `TRUE` will trim leading and trailing spaces.
name	Name of attribute to retrieve.
default	A string used as a default value when the attribute does not exist in every node.

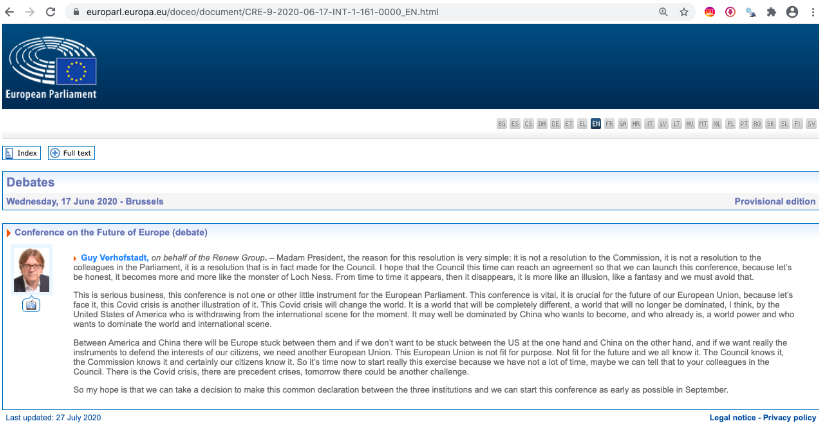
Value

`html_attr`, `html_tag` and `html_text`, a character vector; `html_attrs`, a list.

Examples

```
# NOT RUN {
movie <- read_html("https://en.wikipedia.org/wiki/The_Lego_Movie")
cast <- html_nodes(movie, "tr:nth-child(8) .plainlist a")
html_text(cast)
```

Extracting Text
Example


europarl.europa.eu/doceo/document/CRE-9-2020-06-17-INT-1-161-0000_EN.html

European Parliament

BG ES CS DA DE ET EL EN FR GA HR IT LV LT HU MT NL PL PT RO SK SL FI SV

Index | Full text

Debates

Wednesday, 17 June 2020 - Brussels **Provisional edition**

Conference on the Future of Europe (debate)

Guy Verhofstadt, *on behalf of the Renew Group*. – Madam President, the reason for this resolution is very simple: it is not a resolution to the Commission, it is not a resolution to the colleagues in the Parliament, it is a resolution that is in fact made for the Council. I hope that the Council this time can reach an agreement so that we can launch this conference, because let's be honest, it becomes more and more like the monster of Loch Ness. From time to time it appears, then it disappears, it is more like an illusion, like a fantasy and we must avoid that.

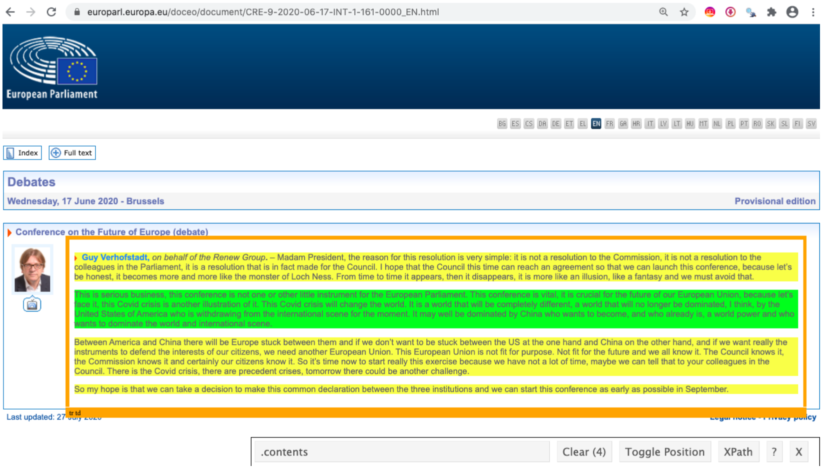
This is serious business, this conference is not one or other little instrument for the European Parliament. This conference is vital, it is crucial for the future of our European Union, because let's face it, this Covid crisis is another illustration of it. This Covid crisis will change the world. It is a world that will be completely different, a world that will no longer be dominated, I think, by the United States of America who is withdrawing from the international scene for the moment. It may well be dominated by China who wants to become, and who already is, a world power and who wants to dominate the world and international scene.

Between America and China there will be Europe stuck between them and if we don't want to be stuck between the US at the one hand and China on the other hand, and if we want really the instruments to defend the interests of our citizens, we need another European Union. This European Union is not fit for purpose. Not fit for the future and we all know it. The Council knows it, the Commission knows it and certainly our citizens know it. So it's time now to start really this exercise because we have not a lot of time, maybe we can tell that to your colleagues in the Council. There is the Covid crisis, there are precedent crises, tomorrow there could be another challenge.

So my hope is that we can take a decision to make this common declaration between the three institutions and we can start this conference as early as possible in September.

Last updated: 27 July 2020 **Legal notice - Privacy policy**

Extracting Text
Example


europarl.europa.eu/doceo/document/CRE-9-2020-06-17-INT-1-161-0000_EN.html
European Parliament
BG ES CS DA DE ET EL EN FR GA HR IT LV LT HU MT NL PL PT RO SK SL FI SV
Index
Full text
tr td
Last updated: 27 July 2020
Legal notice - Privacy policy
.contents
Clear (4)
Toggle Position
XPath
?
X

Debates

Wednesday, 17 June 2020 - Brussels **Provisional edition**

Conference on the Future of Europe (debate)

Guy Verhofstadt, *on behalf of the Renew Group.* – Madam President, the reason for this resolution is very simple: it is not a resolution to the Commission, it is not a resolution to the colleagues in the Parliament, it is a resolution that is in fact made for the Council. I hope that the Council this time can reach an agreement so that we can launch this conference, because let's be honest, it becomes more and more like the monster of Loch Ness. From time to time it appears, then it disappears, it is more like an illusion, like a fantasy and we must avoid that.

This is serious business, this conference is not one or other little instrument for the European Parliament. This conference is vital, it is crucial for the future of our European Union, because let's face it, this Covid crisis is another illustration of it. This Covid crisis will change the world. It is a world that will be completely different, a world that will no longer be dominated, I think, by the United States of America who is withdrawing from the international scene for the moment. It may well be dominated by China who wants to become, and who already is, a world power and who wants to dominate the world and international scene.

Between America and China there will be Europe stuck between them and if we don't want to be stuck between the US at the one hand and China on the other hand, and if we want really the instruments to defend the interests of our citizens, we need another European Union. This European Union is not fit for purpose. Not fit for the future and we all know it. The Council knows it, the Commission knows it and certainly our citizens know it. So it's time now to start really this exercise because we have not a lot of time, maybe we can tell that to your colleagues in the Council. There is the Covid crisis, there are precedent crises, tomorrow there could be another challenge.

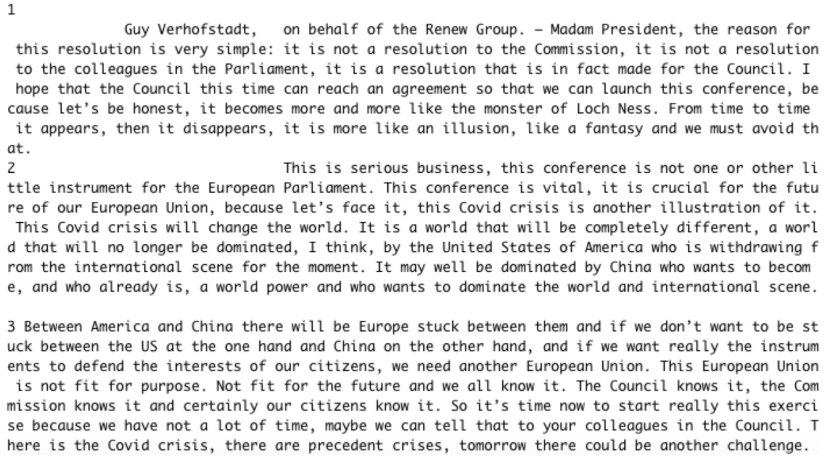
So my hope is that we can take a decision to make this common declaration between the three institutions and we can start this conference as early as possible in September.

Extracting Text
Example

- `url<-"https://www.europarl.europa.eu/doceo/document/CRE-9-2020-06-17-INT-1-161-0000_EN.html"`
- `v_speech1<-read_html(url)`

-
  ```
  text <- v_speech1 %>%
    html_nodes(".contents") %>%
    html_text() %>%
    as.data.frame()
  ```

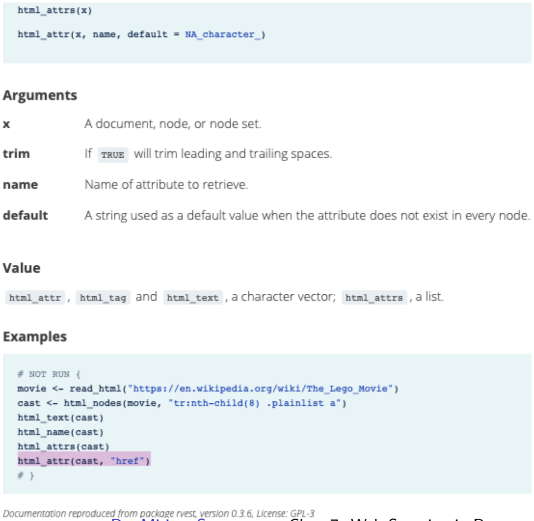
Extracting Text
Example



```
1
              Guy Verhofstadt,   on behalf of the Renew Group. – Madam President, the reason for
 this resolution is very simple: it is not a resolution to the Commission, it is not a resolution
 to the colleagues in the Parliament, it is a resolution that is in fact made for the Council. I
 hope that the Council this time can reach an agreement so that we can launch this conference, be
cause let’s be honest, it becomes more and more like the monster of Loch Ness. From time to time
 it appears, then it disappears, it is more like an illusion, like a fantasy and we must avoid th
at.
2                               This is serious business, this conference is not one or other li
ttle instrument for the European Parliament. This conference is vital, it is crucial for the futu
re of our European Union, because let’s face it, this Covid crisis is another illustration of it.
 This Covid crisis will change the world. It is a world that will be completely different, a worl
d that will no longer be dominated, I think, by the United States of America who is withdrawing f
rom the international scene for the moment. It may well be dominated by China who wants to becom
e, and who already is, a world power and who wants to dominate the world and international scene.

3 Between America and China there will be Europe stuck between them and if we don’t want to be st
uck between the US at the one hand and China on the other hand, and if we want really the instrum
ents to defend the interests of our citizens, we need another European Union. This European Union
 is not fit for purpose. Not fit for the future and we all know it. The Council knows it, the Com
mission knows it and certainly our citizens know it. So it’s time now to start really this exerci
se because we have not a lot of time, maybe we can tell that to your colleagues in the Council. T
here is the Covid crisis, there are precedent crises, tomorrow there could be another challenge.
```

Extracting Links

- Use `html_attr("href")` after identifying the relevant node



```
html_attrs(x)

html_attr(x, name, default = NA_character_)
```

Arguments

x	A document, node, or node set.
trim	If `TRUE` will trim leading and trailing spaces.
name	Name of attribute to retrieve.
default	A string used as a default value when the attribute does not exist in every node.

Value

`html_attr`, `html_tag` and `html_text`, a character vector; `html_attrs`, a list.

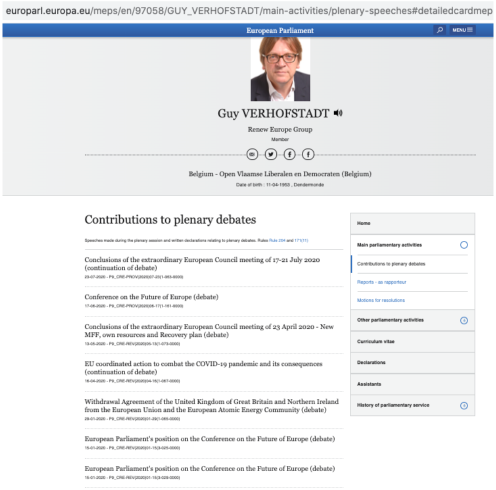
Examples

```
# NOT RUN {
movie <- read_html("https://en.wikipedia.org/wiki/The_Lego_Movie")
cast <- html_nodes(movie, "tr:nth-child(8) .plainlist a")
html_text(cast)
html_name(cast)
html_attrs(cast)
html_attr(cast, "href")
# }
```

Documentation reproduced from package rvest, version 0.3.6, License: GPL-3

Extracting Links

Example



europarl.europa.eu/meps/en/97058/GUY_VERHOFSTADT/main-activities/plenary-speeches#detailedcardmep

Guy VERHOFSTADT

Renew Europe Group

Member

Belgium - Open Vlaamse Liberalen en Democraten (Belgium)

Date of birth : 11-04-1953 , Dendermonde

Contributions to plenary debates

Speeches made during the plenary session and written declarations relating to plenary debates. Rules Rule 204 and 171(11)

Conclusions of the extraordinary European Council meeting of 17-21 July 2020 (continuation of debate)
23-07-2020 - P9_CRE-PROV(2020)07-23(1-063-0000)

Conference on the Future of Europe (debate)
17-06-2020 - P9_CRE-PROV(2020)06-17(1-161-0000)

Conclusions of the extraordinary European Council meeting of 23 April 2020 - New MFF, own resources and Recovery plan (debate)
13-05-2020 - P9_CRE-REV(2020)05-13(1-073-0000)

EU coordinated action to combat the COVID-19 pandemic and its consequences (continuation of debate)
16-04-2020 - P9_CRE-REV(2020)04-16(1-067-0000)

Withdrawal Agreement of the United Kingdom of Great Britain and Northern Ireland from the European Union and the European Atomic Energy Community (debate)
29-01-2020 - P9_CRE-REV(2020)01-29(1-065-0000)

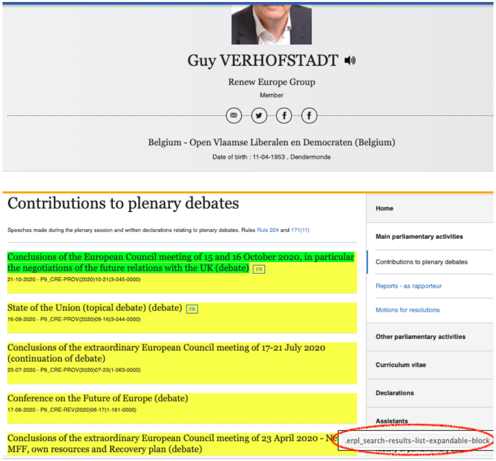
European Parliament's position on the Conference on the Future of Europe (debate)
15-01-2020 - P9_CRE-REV(2020)01-15(3-025-0000)

European Parliament's position on the Conference on the Future of Europe (debate)
15-01-2020 - P9_CRE-REV(2020)01-15(3-029-0000)

Home
Main parliamentary activities
Contributions to plenary debates
Reports - as rapporteur
Motions for resolutions
Other parliamentary activities
Curriculum vitae
Declarations
Assistants
History of parliamentary service

Extracting Links

Example



Guy VERHOFSTADT

Renew Europe Group

Member

Belgium - Open Vlaamse Liberalen en Democraten (Belgium)

Date of birth : 11-04-1953 , Dendermonde

Contributions to plenary debates

Speeches made during the plenary session and written declarations relating to plenary debates. Rules Rule 204 and 171(11)

Conclusions of the European Council meeting of 15 and 16 October 2020, in particular the negotiations of the future relations with the UK (debate) FR
21-10-2020 - P9_CRE-PROV(2020)10-21(3-045-0000)

State of the Union (topical debate) (debate) FR
16-09-2020 - P9_CRE-PROV(2020)09-16(3-044-0000)

Conclusions of the extraordinary European Council meeting of 17-21 July 2020 (continuation of debate)
23-07-2020 - P9_CRE-PROV(2020)07-23(1-063-0000)

Conference on the Future of Europe (debate)
17-06-2020 - P9_CRE-REV(2020)06-17(1-161-0000)

Conclusions of the extraordinary European Council meeting of 23 April 2020 - Ne
MFF, own resources and Recovery plan (debate)

.erpl_search-results-list-expandable-block

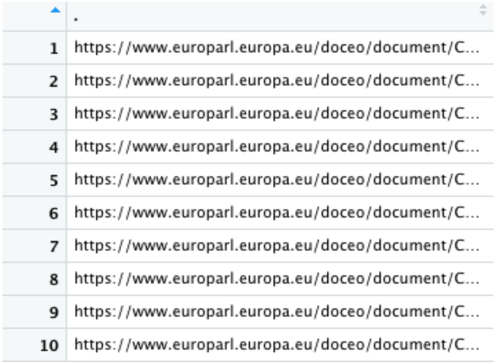
Home
Main parliamentary activities
Contributions to plenary debates
Reports - as rapporteur
Motions for resolutions
Other parliamentary activities
Curriculum vitae
Declarations
Assistants

Extracting Links
Example

- url<-"https://www.europarl.europa.eu/meps/en/97058/GUY_VERHOFSTADT/main-activities/plenary-speeches#detailedcardmep"
- verhof_speech <- read_html(url)
- speech_links <- verhof_speech %>%
  html_nodes(".erpl_search-results-list-expandable-bloc
  a") %>%
  html_attr("href") %>%
  as.data.frame()

Extracting Links

Example



	.
1	https://www.europarl.europa.eu/doceo/document/C...
2	https://www.europarl.europa.eu/doceo/document/C...
3	https://www.europarl.europa.eu/doceo/document/C...
4	https://www.europarl.europa.eu/doceo/document/C...
5	https://www.europarl.europa.eu/doceo/document/C...
6	https://www.europarl.europa.eu/doceo/document/C...
7	https://www.europarl.europa.eu/doceo/document/C...
8	https://www.europarl.europa.eu/doceo/document/C...
9	https://www.europarl.europa.eu/doceo/document/C...
10	https://www.europarl.europa.eu/doceo/document/C...

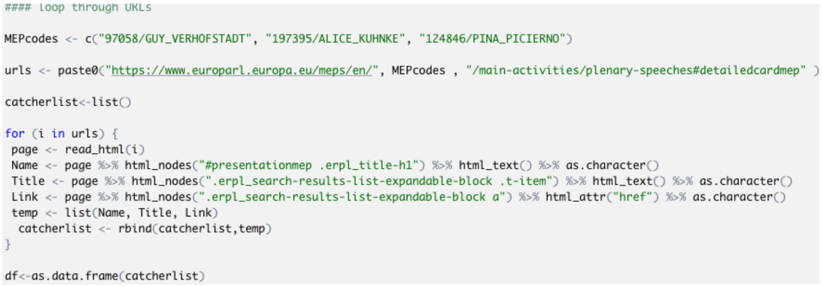
Looping over list of URLs

- This is useful in situations where urls have patterns
- E.g. In EP website, MEPs urls are identical barring from MEP number and name.
- A scraper can exploit this and loop over a series of webpages with identical content
- This ‘simply’ requires creating a list of univocal url codes, pasting those to the unchanging - generic - url and running a standard `for` loop
- The web scraper we built will be inside the `for` loop

Looping over list of URLs
Example

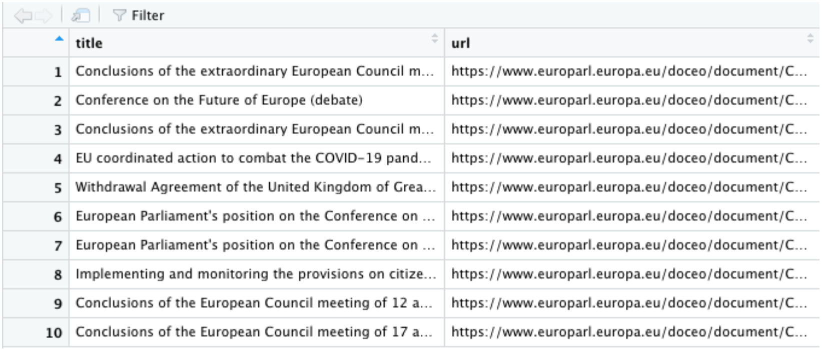
- Consider the following urls:
  1. `https://www.europarl.europa.eu/meps/en/97058/GUY_VERHOFSTADT/main-activities/plenary-speeches#detailedcardmep/`
  2. `https://www.europarl.europa.eu/meps/en/197395/ALICE_KUHNKE/main-activities/plenary-speeches#detailedcardmep`
  3. `https://www.europarl.europa.eu/meps/en/124846/PINA_PICIERNO/main-activities/plenary-speeches#detailedcardmep`

Looping over list of URLs
Example



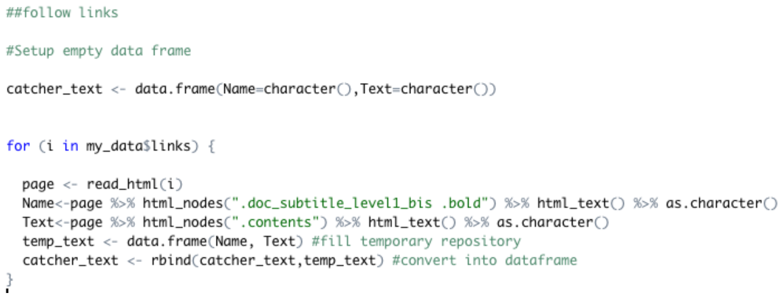
```
#### loop through URLs

MEPcodes <- c("97058/GUY_VERHOFSTADT", "197395/ALICE_KUHNKE", "124846/PINA_PICIERNO")

urls <- paste0("https://www.europarl.europa.eu/meps/en/", MEPcodes , "/main-activities/plenary-speeches#detailedcardmep" )

catcherlist<-list()

for (i in urls) {
 page <- read_html(i)
 Name <- page %>% html_nodes("#presentationmep .erpl_title-h1") %>% html_text() %>% as.character()
 Title <- page %>% html_nodes(".erpl_search-results-list-expandable-block .t-item") %>% html_text() %>% as.character()
 Link <- page %>% html_nodes(".erpl_search-results-list-expandable-block a") %>% html_attr("href") %>% as.character()
 temp <- list(Name, Title, Link)
  catcherlist <- rbind(catcherlist,temp)
}

df<-as.data.frame(catcherlist)
```

Looping over list of links
Example



	title	url
1	Conclusions of the extraordinary European Council m...	https://www.europarl.europa.eu/doceo/document/C...
2	Conference on the Future of Europe (debate)	https://www.europarl.europa.eu/doceo/document/C...
3	Conclusions of the extraordinary European Council m...	https://www.europarl.europa.eu/doceo/document/C...
4	EU coordinated action to combat the COVID-19 pand...	https://www.europarl.europa.eu/doceo/document/C...
5	Withdrawal Agreement of the United Kingdom of Grea...	https://www.europarl.europa.eu/doceo/document/C...
6	European Parliament's position on the Conference on ...	https://www.europarl.europa.eu/doceo/document/C...
7	European Parliament's position on the Conference on ...	https://www.europarl.europa.eu/doceo/document/C...
8	Implementing and monitoring the provisions on citize...	https://www.europarl.europa.eu/doceo/document/C...
9	Conclusions of the European Council meeting of 12 a...	https://www.europarl.europa.eu/doceo/document/C...
10	Conclusions of the European Council meeting of 17 a...	https://www.europarl.europa.eu/doceo/document/C...

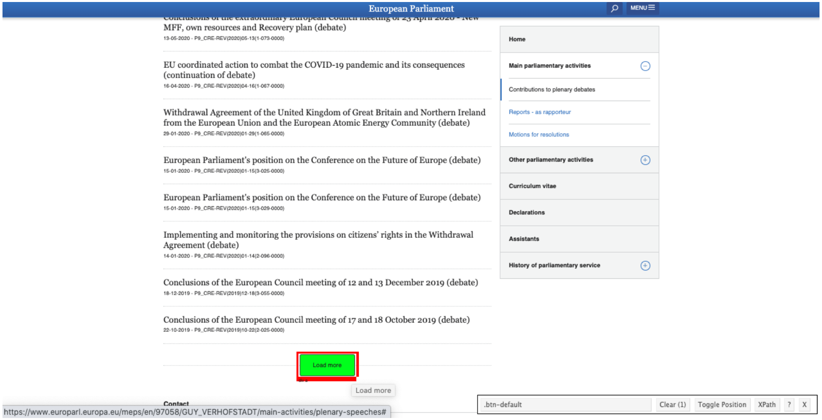
Looping over list of links
Example



```r
##follow links

#Setup empty data frame

catcher_text <- data.frame(Name=character(),Text=character())


for (i in my_data$links) {

  page <- read_html(i)
  Name<-page %>% html_nodes(".doc_subtitle_level1_bis .bold") %>% html_text() %>% as.character()
  Text<-page %>% html_nodes(".contents") %>% html_text() %>% as.character()
  temp_text <- data.frame(Name, Text) #fill temporary repository
  catcher_text <- rbind(catcher_text,temp_text) #convert into dataframe
}
```

Clicking Buttons

E.g. “Load More”



MFF, own resources and Recovery plan (debate)
13-05-2020 - P9_CRE-REV(2020)05-13(1-073-0000)

EU coordinated action to combat the COVID-19 pandemic and its consequences (continuation of debate)
16-04-2020 - P9_CRE-REV(2020)04-16(1-067-0000)

Withdrawal Agreement of the United Kingdom of Great Britain and Northern Ireland from the European Union and the European Atomic Energy Community (debate)
29-01-2020 - P9_CRE-REV(2020)01-29(1-065-0000)

European Parliament's position on the Conference on the Future of Europe (debate)
15-01-2020 - P9_CRE-REV(2020)01-15(3-025-0000)

European Parliament's position on the Conference on the Future of Europe (debate)
15-01-2020 - P9_CRE-REV(2020)01-15(3-029-0000)

Implementing and monitoring the provisions on citizens' rights in the Withdrawal Agreement (debate)
14-01-2020 - P9_CRE-REV(2020)01-14(2-096-0000)

Conclusions of the European Council meeting of 12 and 13 December 2019 (debate)
18-12-2019 - P9_CRE-REV(2019)12-18(3-055-0000)

Conclusions of the European Council meeting of 17 and 18 October 2019 (debate)
22-10-2019 - P9_CRE-REV(2019)10-22(2-025-0000)

https://www.europarl.europa.eu/meps/en/97058/GUY_VERHOFSTADT/main-activities/plenary-speeches#

`RSelenium`

- *Dynamic* interaction (filling fields, clicking buttons) with the website is not possible through `rvest`
- You need `RSelenium` for this: but **outside of the scopes of this course**
- `RSelenium` requires a bit of environment preparation:
    - It requires selenium initiation, by opening a browser session (not simply the specification of a url)
    - Only after setting up the 'Remote Server/Driver' it can navigate the specific web page ... v. time consuming
- The core functions to perform a button click are `RSelenium`'s `$findElement` and `$clickElement`
- If you want to know more: `http://joshuamccrain.com/tutorials/web_scraping_R_selenium.html`

What we have learnt today ...

- SelectorGadget can make the first step of scraping (web-page analysis) so much easier
- There are 3 core strategies in scraping
  - Regular expression (worst case)
  - Node queries (typical case)
  - APIs (gold standard but rare)
- The core functions in R to perform **node query** scraping - using the package `rvest`
- Awareness of `RSelenium` and the problems in scraping dynamic pages
- Next week: more on scraping via APIs: we're going to learn how to scrape Twitter data