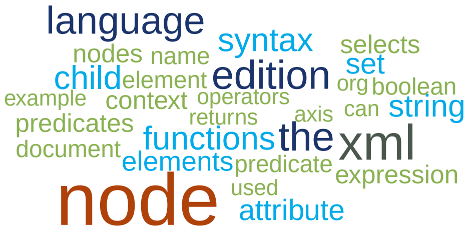
Introduction to Web Scraping with R

Using SelectorGadget



Simon Munzert | IPSDS

SelectorGadget

SelectorGadget

Scraping webpages

- the most cumbersome part of web scraping (data tidying aside) is the construction of XPath expressions that match the components of a page you want to extract
- it will take a couple of scraping projects until you'll truly have mastered XPath

SelectorGadget

Scraping webpages

- the most cumbersome part of web scraping (data tidying aside) is the construction of XPath expressions that match the components of a page you want to extract
- it will take a couple of scraping projects until you'll truly have mastered XPath

A much-appreciated helper

- **SelectorGadget** is a JavaScript browser plugin that constructs XPath statements (or CSS selectors) via a point-and-click approach
- it is available here: http://selectorgadget.com/ (there s also a Chrome extension)
- the tool is magic and you will love it

What does SelectorGadget do?

What does SelectorGadget do?

On the surface

1. you activate the tool on any webpage you want to scrape
2. based on your selection of components, the tool learns about your desired components and generates an XPath expression (or CSS selector) for you

What does SelectorGadget do?

On the surface

1. you activate the tool on any webpage you want to scrape
2. based on your selection of components, the tool learns about your desired components and generates an XPath expression (or CSS selector) for you

Under the hood

- based on your selection(s), the tool looks for similar elements on the page
- the underlying algorithm, which draws on Google's diff-match-patch libraries, focuses on CSS characteristics, such as tag names and `<div>` and `<span>` attributes

Installation and use

Installation

Go to http://selectorgadget.com/ and drag the link provided at the end of the page ("Or drag this link to your bookmark bar") to your bookmark bar.

Installation and use

Installation

Go to http://selectorgadget.com/ and drag the link provided at the end of the page ("Or drag this link to your bookmark bar") to your bookmark bar.

Use

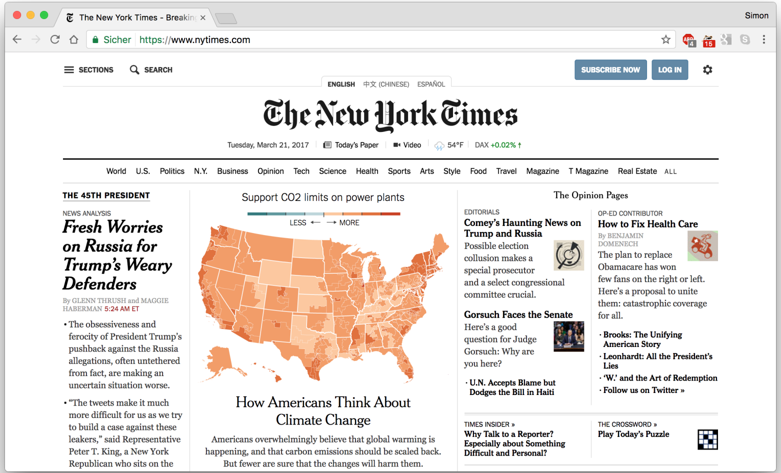
1. open the webpage you want to scrape
2. click on the SelectorGadget bookmarklet
3. click on the elements you want to extract. Manually selected elements will turn **green**. Elements that match the selector will be highlighted **yellow**
4. de-select the elements you do not want to extract - they will turn **red**.
5. repeat steps 3–4 until the marked set of elements matches the set you want to extract
6. click on the "XPath" button in the console
7. copy the generated XPath expression and process it with `rvest`

Example

Example

Example

- source: https://nytimes.com



The New York Times - Breakin ×

Simon

Sicher https://www.nytimes.com

SECTIONS SEARCH SUBSCRIBE NOW LOG IN

ENGLISH 中文 (CHINESE) ESPAÑOL

The New York Times

Tuesday, March 21, 2017 | Today's Paper | Video | 54°F | DAX +0.02% ↑

World U.S. Politics N.Y. Business Opinion Tech Science Health Sports Arts Style Food Travel Magazine T Magazine Real Estate ALL

THE 45TH PRESIDENT

NEWS ANALYSIS

Fresh Worries on Russia for Trump's Weary Defenders

By GLENN THRUSH and MAGGIE HABERMAN 5:24 AM ET

- The obsessiveness and ferocity of President Trump's pushback against the Russia allegations, often untethered from fact, are making an uncertain situation worse.
- "The tweets make it much more difficult for us as we try to build a case against these leakers," said Representative Peter T. King, a New York Republican who sits on the

Support CO2 limits on power plants

LESS ← → MORE

How Americans Think About Climate Change

Americans overwhelmingly believe that global warming is happening, and that carbon emissions should be scaled back. But fewer are sure that the changes will harm them.

The Opinion Pages

EDITORIALS

Comey's Haunting News on Trump and Russia

Possible election collusion makes a special prosecutor and a select congressional committee crucial.

Gorsuch Faces the Senate

Here's a good question for Judge Gorsuch: Why are you here?

- **U.N. Accepts Blame but Dodges the Bill in Haiti**

OP-ED CONTRIBUTOR

How to Fix Health Care

By BENJAMIN DOMENECH

The plan to replace Obamacare has won few fans on the right or left. Here's a proposal to unite them: catastrophic coverage for all.

- **Brooks: The Unifying American Story**
- **Leonhardt: All the President's Lies**
- **'W.' and the Art of Redemption**
- **Follow us on Twitter »**

TIMES INSIDER »

Why Talk to a Reporter? Especially about Something Difficult and Personal?

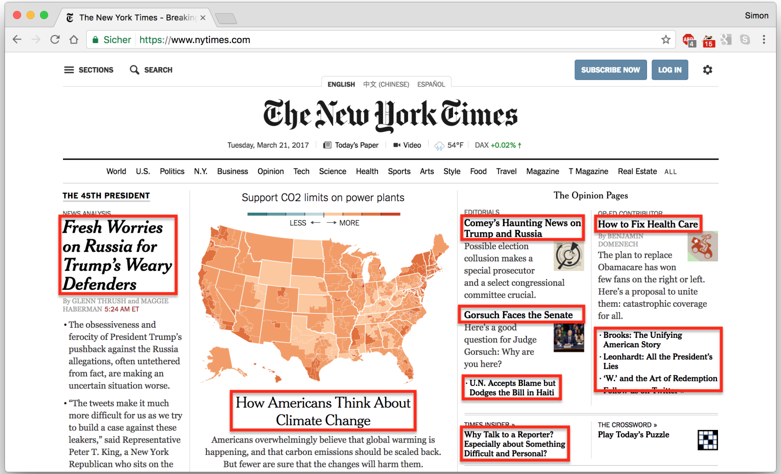
THE CROSSWORD »

Play Today's Puzzle

Example

Example

- source: https://nytimes.com
- goal: scrape all headlines



The New York Times

Tuesday, March 21, 2017 | Today's Paper | Video | 54°F | DAX +0.02%

World U.S. Politics N.Y. Business Opinion Tech Science Health Sports Arts Style Food Travel Magazine T Magazine Real Estate ALL

THE 45TH PRESIDENT

NEWS ANALYSIS

Fresh Worries on Russia for Trump's Weary Defenders

By GLENN THRUSH and MAGGIE HABERMAN 5:24 AM ET

- The obsessiveness and ferocity of President Trump's pushback against the Russia allegations, often untethered from fact, are making an uncertain situation worse.
- "The tweets make it much more difficult for us as we try to build a case against these leakers," said Representative Peter T. King, a New York Republican who sits on the

Support CO2 limits on power plants

LESS ← → MORE

How Americans Think About Climate Change

Americans overwhelmingly believe that global warming is happening, and that carbon emissions should be scaled back. But fewer are sure that the changes will harm them.

The Opinion Pages

EDITORIALS

Comey's Haunting News on Trump and Russia

Possible election collusion makes a special prosecutor and a select congressional committee crucial.

Gorsuch Faces the Senate

Here's a good question for Judge Gorsuch: Why are you here?

- **U.N. Accepts Blame but Dodges the Bill in Haiti**

OP-ED CONTRIBUTOR

How to Fix Health Care

By BENJAMIN DOMENECH

The plan to replace Obamacare has won few fans on the right or left. Here's a proposal to unite them: catastrophic coverage for all.

- **Brooks: The Unifying American Story**
- **Leonhardt: All the President's Lies**
- **'W.' and the Art of Redemption**

Follow us on Twitter »

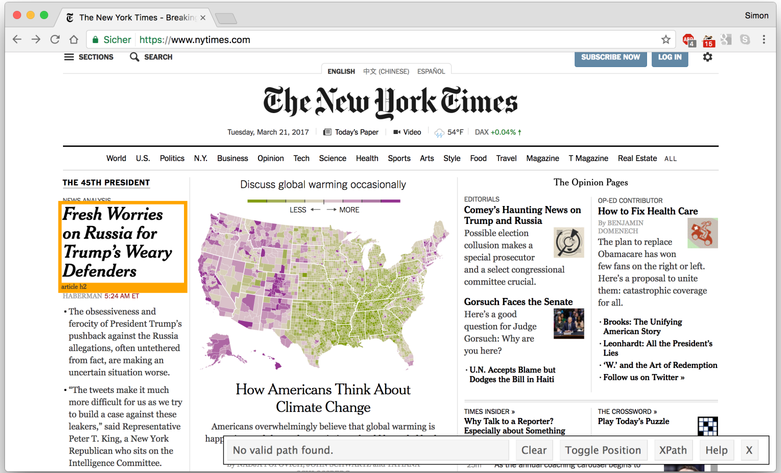
TIMES INSIDER »

Why Talk to a Reporter? Especially about Something Difficult and Personal?

THE CROSSWORD »

Play Today's Puzzle

Example



The New York Times - Breakin ×

Simon

Sicher https://www.nytimes.com

SECTIONS
SEARCH
SUBSCRIBE NOW
LOG IN
ENGLISH
中文 (CHINESE)
ESPAÑOL

The New York Times

Tuesday, March 21, 2017 | Today's Paper | Video | 54°F | DAX +0.04% ↑

World U.S. Politics N.Y. Business Opinion Tech Science Health Sports Arts Style Food Travel Magazine T Magazine Real Estate ALL

THE 45TH PRESIDENT

NEWS ANALYSIS

Fresh Worries on Russia for Trump's Weary Defenders

article h2

HABERMAN 5:24 AM ET

- The obsessiveness and ferocity of President Trump's pushback against the Russia allegations, often untethered from fact, are making an uncertain situation worse.
- "The tweets make it much more difficult for us as we try to build a case against these leakers," said Representative Peter T. King, a New York Republican who sits on the Intelligence Committee.

Discuss global warming occasionally

LESS ⟵ ⟶ MORE

How Americans Think About Climate Change

Americans overwhelmingly believe that global warming is happ

The Opinion Pages

EDITORIALS

Comey's Haunting News on Trump and Russia

Possible election collusion makes a special prosecutor and a select congressional committee crucial.

Gorsuch Faces the Senate

Here's a good question for Judge Gorsuch: Why are you here?

- **U.N. Accepts Blame but Dodges the Bill in Haiti**

OP-ED CONTRIBUTOR

How to Fix Health Care

By BENJAMIN DOMENECH

The plan to replace Obamacare has won few fans on the right or left. Here's a proposal to unite them: catastrophic coverage for all.

- **Brooks: The Unifying American Story**
- **Leonhardt: All the President's Lies**
- **'W.' and the Art of Redemption**
- **Follow us on Twitter »**

TIMES INSIDER »

Why Talk to a Reporter? Especially about Something

THE CROSSWORD »

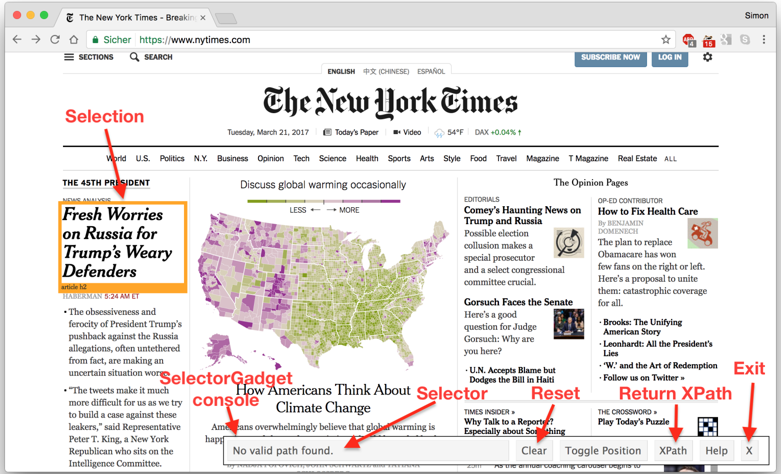
Play Today's Puzzle

No valid path found.

Example

1. click on SelectorGadget bookmarklet (not shown)
2. click on first headline

Example

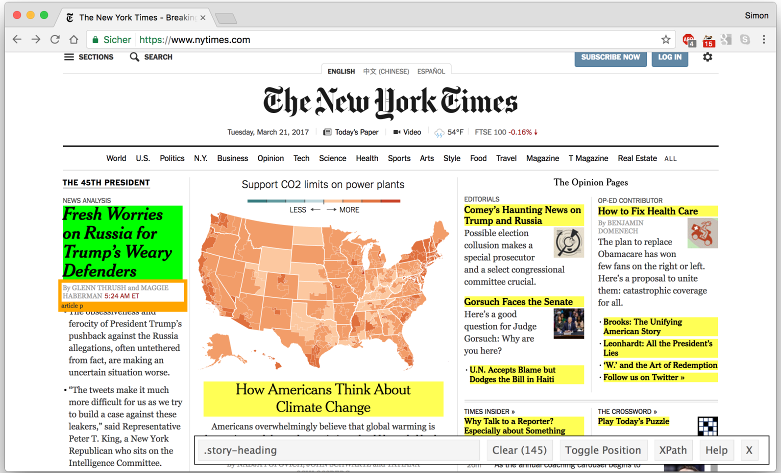


The New York Times - Breaking
Simon
Sicher https://www.nytimes.com
SECTIONS SEARCH
SUBSCRIBE NOW LOG IN
ENGLISH 中文 (CHINESE) ESPAÑOL
The New York Times
Tuesday, March 21, 2017 Today's Paper Video 54°F DAX +0.04%
World U.S. Politics N.Y. Business Opinion Tech Science Health Sports Arts Style Food Travel Magazine T Magazine Real Estate ALL
Selection
THE 45TH PRESIDENT
Fresh Worries on Russia for Trump's Weary Defenders
article h2
Discuss global warming occasionally
LESS MORE
How Americans Think About Climate Change
The Opinion Pages
SelectorGadget console
Selector
Reset
Return XPath
Exit
No valid path found.
Clear
Toggle Position
XPath
Help
X

Example

1. click on SelectorGadget bookmarklet (not shown)
2. click on first headline

Example



The New York Times - Breakin ×

Simon

Sicher https://www.nytimes.com

SECTIONS SEARCH SUBSCRIBE NOW LOG IN

ENGLISH 中文 (CHINESE) ESPAÑOL

The New York Times

Tuesday, March 21, 2017 | Today's Paper | Video | 54°F | FTSE 100 -0.16% ↓

World U.S. Politics N.Y. Business Opinion Tech Science Health Sports Arts Style Food Travel Magazine T Magazine Real Estate ALL

THE 45TH PRESIDENT

NEWS ANALYSIS

Fresh Worries on Russia for Trump's Weary Defenders

By GLENN THRUSH and MAGGIE HABERMAN 5:24 AM ET

article p

- The obsessiveness and ferocity of President Trump's pushback against the Russia allegations, often untethered from fact, are making an uncertain situation worse.
- "The tweets make it much more difficult for us as we try to build a case against these leakers," said Representative Peter T. King, a New York Republican who sits on the Intelligence Committee.

Support CO2 limits on power plants

LESS ⟵ ⟶ MORE

How Americans Think About Climate Change

Americans overwhelmingly believe that global warming is

The Opinion Pages

EDITORIALS

Comey's Haunting News on Trump and Russia

Possible election collusion makes a special prosecutor and a select congressional committee crucial.

Gorsuch Faces the Senate

Here's a good question for Judge Gorsuch: Why are you here?

- **U.N. Accepts Blame but Dodges the Bill in Haiti**

OP-ED CONTRIBUTOR

How to Fix Health Care

By BENJAMIN DOMENECH

The plan to replace Obamacare has won few fans on the right or left. Here's a proposal to unite them: catastrophic coverage for all.

- **Brooks: The Unifying American Story**
- **Leonhardt: All the President's Lies**
- **'W.' and the Art of Redemption**
- **Follow us on Twitter »**

TIMES INSIDER »

Why Talk to a Reporter? Especially about Something

THE CROSSWORD »

Play Today's Puzzle

.story-heading | Clear (145) | Toggle Position | XPath | Help | X

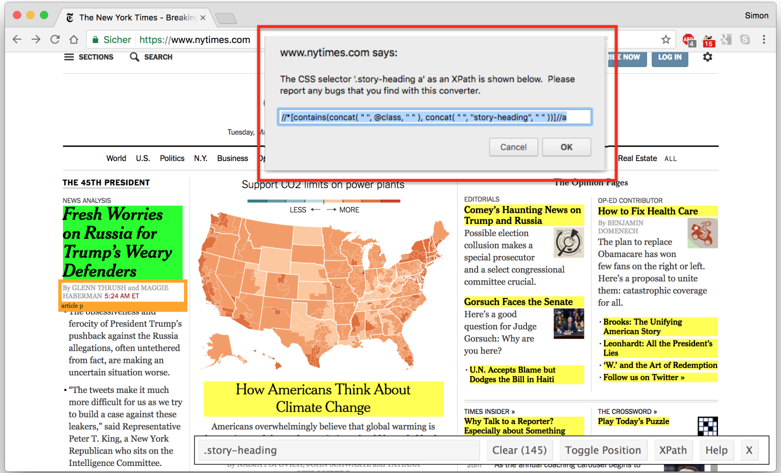
Example

1. click on SelectorGadget bookmarklet (not shown)
2. click on first headline
3. all relevant headlines selected

Example

Example

1. click on SelectorGadget bookmarklet (not shown)
2. click on first headline
3. all relevant headlines selected
4. click on "XPath"
5. copy proposed expression



www.nytimes.com says:

The CSS selector '.story-heading a' as an XPath is shown below. Please report any bugs that you find with this converter.

```
//*[contains(concat( " ", @class, " " ), concat( " ", "story-heading", " " ))]//a
```

THE 45TH PRESIDENT

NEWS ANALYSIS

Fresh Worries on Russia for Trump's Weary Defenders

By GLENN THRUSH and MAGGIE HABERMAN 5:24 AM ET

ferocity of President Trump's pushback against the Russia allegations, often untethered from fact, are making an uncertain situation worse.

• "The tweets make it much more difficult for us as we try to build a case against these leakers," said Representative Peter T. King, a New York Republican who sits on the Intelligence Committee.

Support CO2 limits on power plants

LESS ⟵ ⟶ MORE

How Americans Think About Climate Change

Americans overwhelmingly believe that global warming is

The Opinion Pages

EDITORIALS

Comey's Haunting News on Trump and Russia

Possible election collusion makes a special prosecutor and a select congressional committee crucial.

Gorsuch Faces the Senate

Here's a good question for Judge Gorsuch: Why are you here?

• U.N. Accepts Blame but Dodges the Bill in Haiti

OP-ED CONTRIBUTOR

How to Fix Health Care

By BENJAMIN DOMENECH

The plan to replace Obamacare has won few fans on the right or left. Here's a proposal to unite them: catastrophic coverage for all.

• Brooks: The Unifying American Story

• Leonhardt: All the President's Lies

• 'W.' and the Art of Redemption

• Follow us on Twitter »

TIMES INSIDER »

Why Talk to a Reporter? Especially about Something

THE CROSSWORD »

Play Today's Puzzle

Example

Now, we switch to R

R code

```
library(rvest)
url_p <- read_html("https://www.nytimes.com")
headlines <- html_nodes(url_p, xpath = '//*[contains(concat( " ", @class, " " ), concat( " ",
"story-heading", " " ))]//a') # we use single quotation marks here to wrap around the
expression!
headlines_raw <- html_text(headlines)
head(headlines_raw)
 [1] "Fresh Worries on Russia for Trump's Weary Defenders"
 [2] "F.B.I. Confirms Inquiry on Trump Team's Russia Ties"
 [3] "Takeaways From the Hearing "
 [4] "\n      Trump and the Russians: Links? No Links?"
 [5] "G.O.P. Responds by Changing Subject"
 [6] "U.S. Limits Devices on Foreign Airlines From 8 Countries"
```

end

Example

Let’s clean the data

R code

```
headlines_clean <- headlines_raw %>% str_replace_all("\\n", "") %>% str_trim()
length(headlines_clean)
[1] "Fresh Worries on Russia for Trump's Weary Defenders"
[2] "F.B.I. Confirms Inquiry on Trump Team's Russia Ties"
[3] "Takeaways From the Hearing"
[4] "Trump and the Russians: Links? No Links?"
[5] "G.O.P. Responds by Changing Subject"
[6] "U.S. Limits Devices on Foreign Airlines From 8 Countries"
str_detect(headlines_clean, "Trump") %>% table()
FALSE  TRUE
  133    12
```

end

Summary

Summary

So why did I bother you with learning XPath at all?

Summary

So why did I bother you with learning XPath at all?

Caveats

- SelectorGadget is not perfect. Sometimes, the algorithm will fail in finding a useful XPath expression
- starting from a different element sometimes (but not always!) helps
- often the generated expressions are unnecessarily complex, unintuitive and therefore difficult to debug
- in my experience, SelectorGadget works 70-80% of the times when you want to scrape info from a static webpage
- you are also prepared for the remaining 20-30%!

Source: http://inspectorgadget.wikia.com/wiki/File:Inspector_Gadget_Thinking.png (DANTHEMAN123)