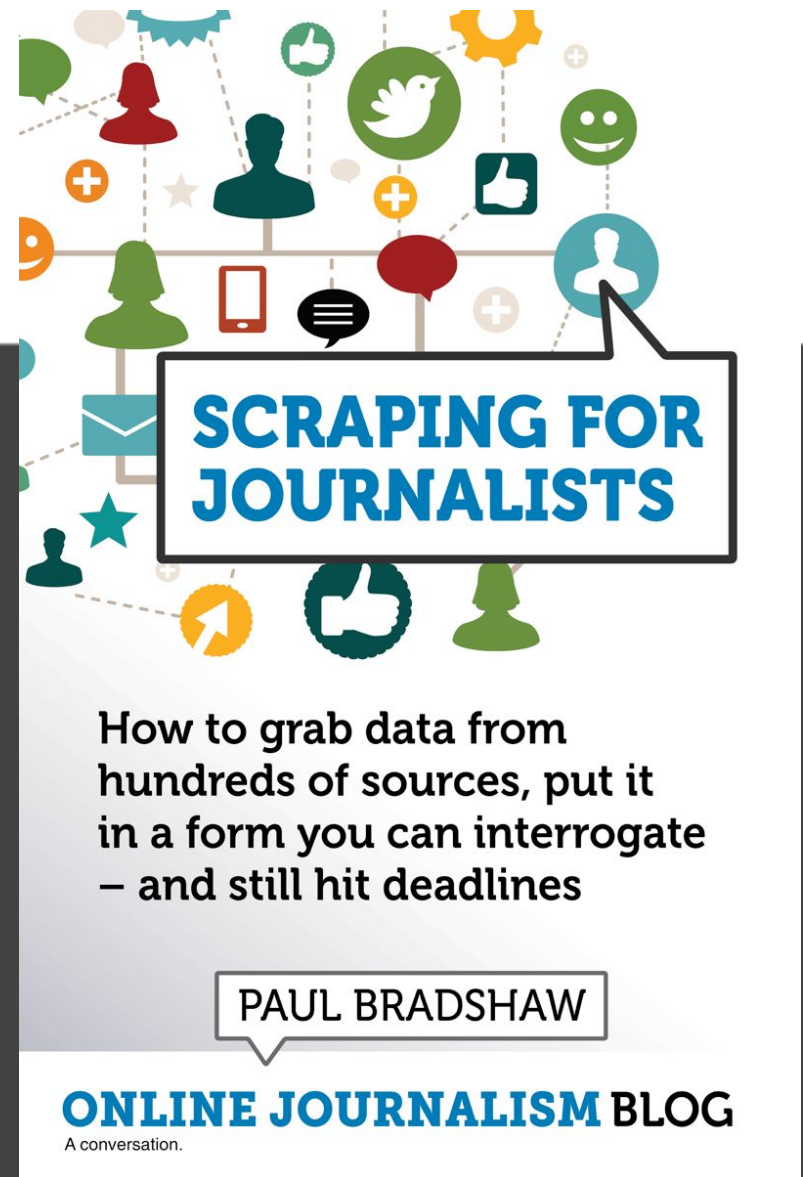


# Using CSS selectors in a scraper



Paul Bradshaw  
[Leanpub.com/scrapingforjournalists](http://leanpub.com/scrapingforjournalists)

# What we'll cover

- What are **CSS selectors** - and why they are useful in scraping
- How to **use** CSS selectors in a Colab Python notebook, with **cssselect**

# CSS selectors

- Created so web designers could style elements, e.g. 'make links red'
- They 'select' elements within HTML tags (e.g. links, headings, images) so they can be styled
- E.g. to select anything inside a `<img>` tag, the selector would be simply `img`

# Detour: HTML

- HTML webpages are created using HTML tags
- Most tags are like buttons, with an 'opening' tag turning something on (e.g. bold), and a 'closing' tag turning it off, e.g. `<p>` `</p>`
- Tags can have attributes and values, e.g. `<p class="firstpar">`
- 'class' and 'id' are common attributes. The value comes after, normally in quotes
- Tags are nested within each other, e.g. a tag to make a word bold will be nested within a paragraph, nested within an article and so on

# CSS selectors

- You can specify combinations of HTML tags, e.g. to select any bold text within a paragraph within a div tag:

```
div p strong
```

- You can also specify attributes of those tags, such as their class or id

```
div[@class="article"]
```

- Or a combination of those

```
div[@class="article"] p strong
```

```
pars = root.cssselect('p')
```

# cssselect

- The cssselect library has functions that allow you to use CSS selectors to extract information from a webpage that's been converted using lxml.html (often stored in a variable called 'root') - e.g.  


```
pars = root.cssselect('p')
```
- The cssselect function is attached to `root` with a period, and the selector put in parentheses
- The result is always a **list** - even if it's a list of one, or zero, results

**Let's apply this to a  
webpage...**



## Find services

You can search all of our service directories from here. Try searching by service name, service type, condition or surgical procedure.

<b>Find</b>	<input type="text" value="Eating disorders"/>	<b>Location</b>	<input type="text" value="nottingham"/>		<b>Search</b>
-------------	---	-----------------	---	---	---------------

or browse the Services A-Z

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

## Select your location for Eating disorders in nottingham

Looks like there is more than one "nottingham" to choose from. Please select your location from th

**Nottingham, Nottingham, NG1**

**Nottinghamshire, Nottinghamshire, NG22**

<https://www.nhs.uk/service-search/other-services>

# Select your service for Eating disorders **near** Nottingham

Looks like there is more than one service available that could help with your condition. Please select c

## Services

Care services for people with eating disorders

Child and adolescent mental health services (CAMHS)

Eating disorder support

Eating disorders - inpatient

Eating disorders - outpatient

## Treatments

Eating disorders (adults)

Eating disorders (children and adolescents)

# Results for **Eating disorder support** in **Nottingham**

**Store Nottingham as your main location for future visits?**

**Narrow search** or **start new search**

Showing 1-10 of 28 results | Results per page



Please check travel times before starting your journey. Distances are given in a straight line and may n

## Address & contact details

## Information supplied by

### **Addictive Eaters Anonymous - Nottingham**

**Tel: 03301333615**

Station Street  
Nottingham  
NG2 3NG  
0.7 miles away

Beat is the UK's leading charity for people with eating disorders and their families. Eating disorders are a serious mental illness affecting over 725, 000 people in the UK. The charity provides helplines for adults and young people, message boards, online chat groups, and emotional overeating su...

**Continue reading overview**

### **Nottinghamshire Adult Eating Disorder Team**

**Tel: 0115 876 0162**

Mandala Centre  
Gregory Boulevard

Beat is the UK's leading charity for people with eating disorders and their families. Eating disorders are a serious mental illness affecting over 725, 000 people in the UK. The charity provides helplines for adults

<https://www.nhs.uk/service-search/other-services/Eating-disorders/Nottingham/Results/102/-1.158/52.955/1797/15942?distance=25>

# Results for **Eating disorder support** in **Nottingham**

Store Nottingham as your main location for future visits?

**Narrow search** or **start new search**

Showing 1-10 of 805 results | Results per page



Please check travel times before starting your journey. Distances are given in a straight line and may

Address & contact details

Information supplied by

## **Addictive Eaters Anonymous - Nottingham**

**Tel: 03301333615**

Station Street

Nottingham

NG2 3NG

**0.7 miles away**

Beat is the UK's leading charity for people with eating disorders and their families. Eating disorders are a serious mental illness affecting over 725, 000 people in the UK. The charity provides helplines for adults and young people, message boards, online chat groups, and emotional overeating su...

**Continue reading overview**

## **Nottinghamshire Adult Eating Disorder Team**

<https://www.nhs.uk/service-search/other-services/Eating-disorders/Nottingham/Results/102/-1.158/52.955/1797/15942?distance=500>



# Results for Eating disorder support in Nottingham

Store Nottingham as your main location for future visits?

Narrow search or start new search

Showing 1-100 of 805 results | Results per page

 Please check travel times before starting your journey. Distances are given in a straight line and may n

Address & contact details	Information supplied by
<b>Addictive Eaters Anonymous - Nottingham</b>	
<b>Tel: 03301333615</b> Station Street Nottingham NG2 3NG 0.7 miles away	Beat is the UK's leading charity for people with eating disorders and their families. Eating disorders are a serious mental illness affecting over 725, 000 people in the UK. The charity provides helplines for adults and young people, message boards, online chat groups, and emotional overeating su... <b>Continue reading overview</b>
<b>Nottinghamshire Adult Eating Disorder Team</b>	

<b>Tel: 0115 876 0162</b>	Beat is the UK's leading charity for people with eating disorders and their families. Eating disorders are a
---------------------------	--

https://www.nhs.uk/service-search/other-services/Eating-disorders/Nottingham/Results/102/-1.158/52.955/1797/15942?distance=500&ResultsOnPageValue=100



## Table of contents



+ Code + Text

Connect ▾

Editing



## An example scraper showing how to use cssselect

- <> Adding a user agent
- Drilling down into the HTML
- Capturing both 'columns' of data
- Improving the scraper
- Exporting the results
- Improvement 1: Cleaning/splitting the data
- Improvement 2: Grabbing the links to detail pages
- Improvement 3: Scraping multiple pages

## + Section

## An example scraper showing how to use cssselect

This notebook explains how to scrape an example webpage as a way of demonstrating how to apply the `cssselect` library.

First, we import the libraries we will need.

```
#install the libraries
#scraperwiki is a library for scraping webpages
!pip install scraperwiki
import scraperwiki
#lxml.html is used to convert it into xml (more structured)
import lxml.html
#cssselect is used to drill down into that and find data in tags
!pip install cssselect
import cssselect
#the pandas library which is used to work with data
import pandas
```

And the first lines of our scraper.

```
[ ] #store the url we want to scrape
theurl = "https://www.nhs.uk/service-search/other-services/Eating-disorders/Nottingham/Results/102/-1.158/52.955/"
#scrape the webpage at that url and store in 'html'
#without a user agent we get a 403 error on this webpage
```

```
#store the url we want to scrape
```

```
theurl =
```

```
"https://www.nhs.uk/service-search/other-services/Eating-disorders/Nottingham/Results/102/-1.158/52.955/1797/15942?distance=500&ResultsOnPageValue=100"
```

```
#scrape the webpage at that url and store in 'html'
```

```
html = scraperwiki.scrape(theurl, user_agent="Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/68.0.3440.106 Safari/537.36")
```

```
#convert 'html' into an lxml object so we can drill into it
```

```
root = lxml.html.fromstring(html)
```

```
#store the url we want to scrape

theurl =
"https://www.nhs.uk/service-search/other-services/Eating-disorders/Nottingham/Results/102/-1.158/52.955/1797/15942?distance=500&ResultsOnPageValue=100"

#scrape the webpage at that url and store in 'html'

html = scraperwiki.scrape(theurl, user_agent="Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/68.0.3440.106 Safari/537.36")

#convert 'html' into an lxml object so we can drill into it

root = lxml.html.fromstring(html)

#grab the contents of every <th> tag

servicenames = root.cssselect('th')

#check how many results - there should be 100

len(servicenames)
```



```
#grab the contents of every <th> tag
servicenames = root.cssselect('th')

#check how many results - there should be 100
len(servicenames)

#Loop through the results
for i in servicenames:

    #print the text inside the tag
    print(i.text_content())
```

```
#Loop through the results
for i in servicenames[-100:]:
    #print the text inside the tag
    print(i.text_content())

#grab the contents of each <div class="fcdetailsleft"> tag
tels = root.cssselect('div.fcdetailsleft')

#count how many matches are in that list
len(tels)

firsttel = tels[0].text_content()
print(firsttel)
```

# Recap

- Use `cssselect` to drill down into the `lxml.html` object 'root'

```
tels = root.cssselect('div.fcdetailsleft')
```

- Always generates a list - access items using a for loop or an index/indices
- Add `.text_content()` to extract text

```
tels[0].text_content()  
for i in servicenames:  
    print(i.text_content)
```

# Try it now:

- In your notebook scrape the page and extract the contents of:
  - `<th>` tags
  - `<div class="fcdetailsleft">` tags
- Loop through those tags and print the `.text_content()`
- Access the first match and print the `.text_content()`