Web Scraping

A brief overview by Adam Stapleton

What is Web Scraping?

“Web scraping, web harvesting, or web data extraction is data scraping used for extracting data from websites.”

- Wikipedia

Is web scraping ILLEGAL?

https://en.wikipedia.org/robots.txt
https://www.google.com/robots.txt
https://www.facebook.com/robots.txt
http://www.daft.ie/robots.txt

Tip

Ask Google.

They've pretty much perfected it.

Add "/robots.txt" to the end of any homepage url and you'll see

So, is it illegal?

Well, is it illegal to rearrange the furniture in a restaurant that you're eating at?

Packages

Necessary

- requests
- bs4 (BeautifulSoup)
- Selenium
  - Geckodriver (Firefox)

Useful

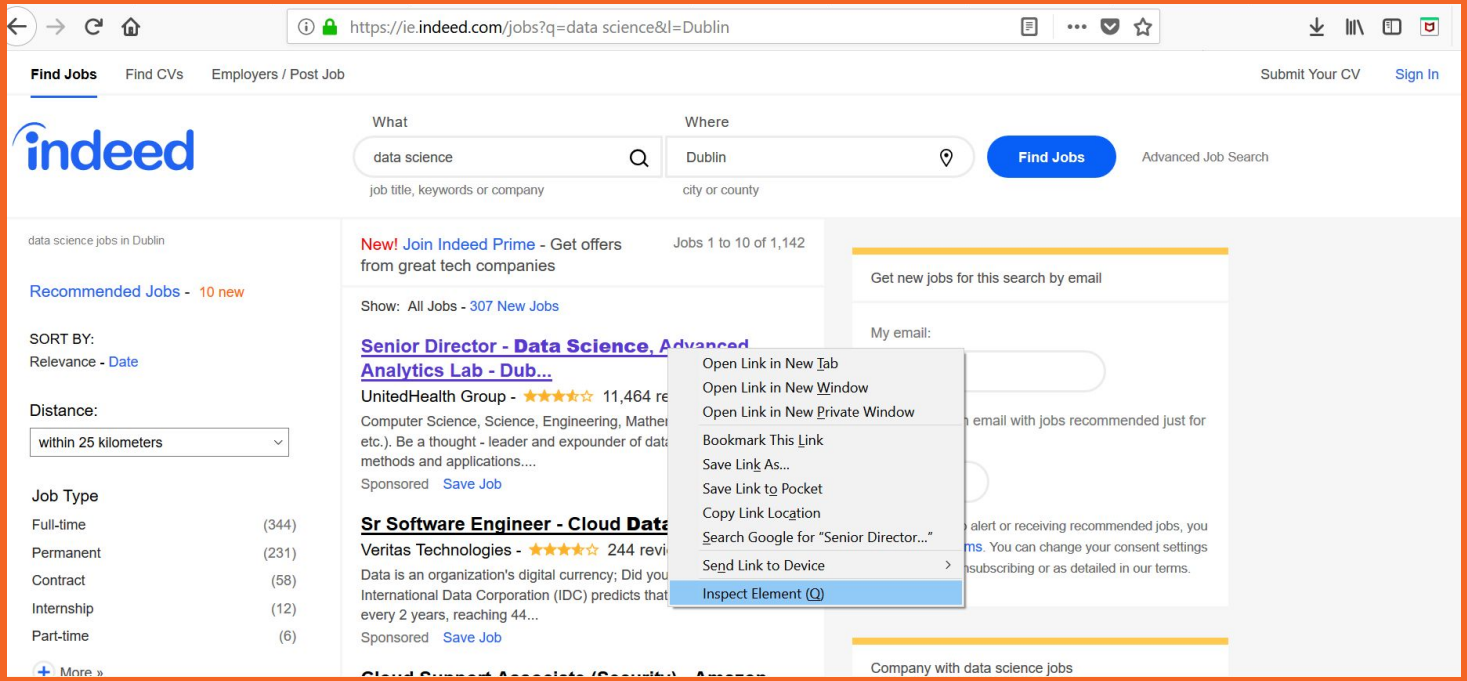
- Pandas
- random

Approach

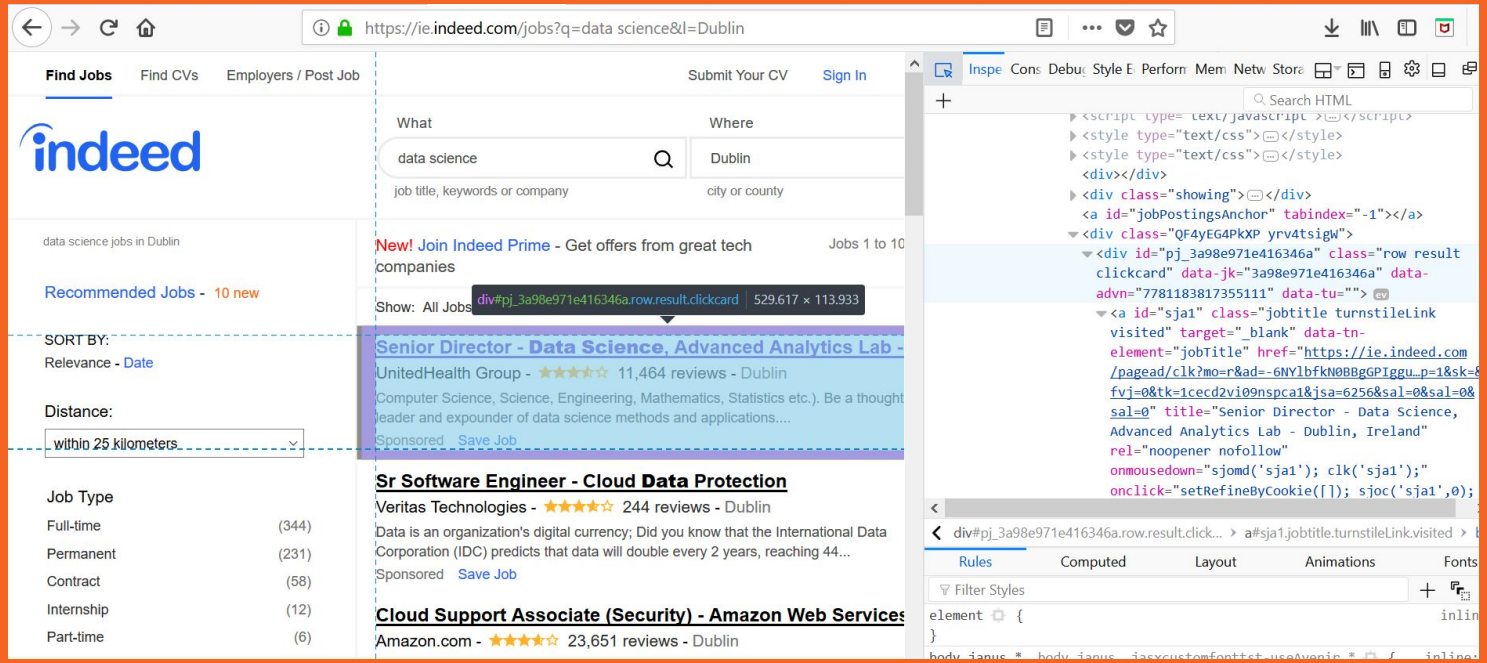
In general, this is how to go about extracting data from a website.

- **Inspect html**
  Find what container(s) your information is in
- **requests/selenium**
  Extract website body (which will contain your information)
- **BeautifulSoup**
  Parse the body to find the containers you want, and extract the information you need

Example

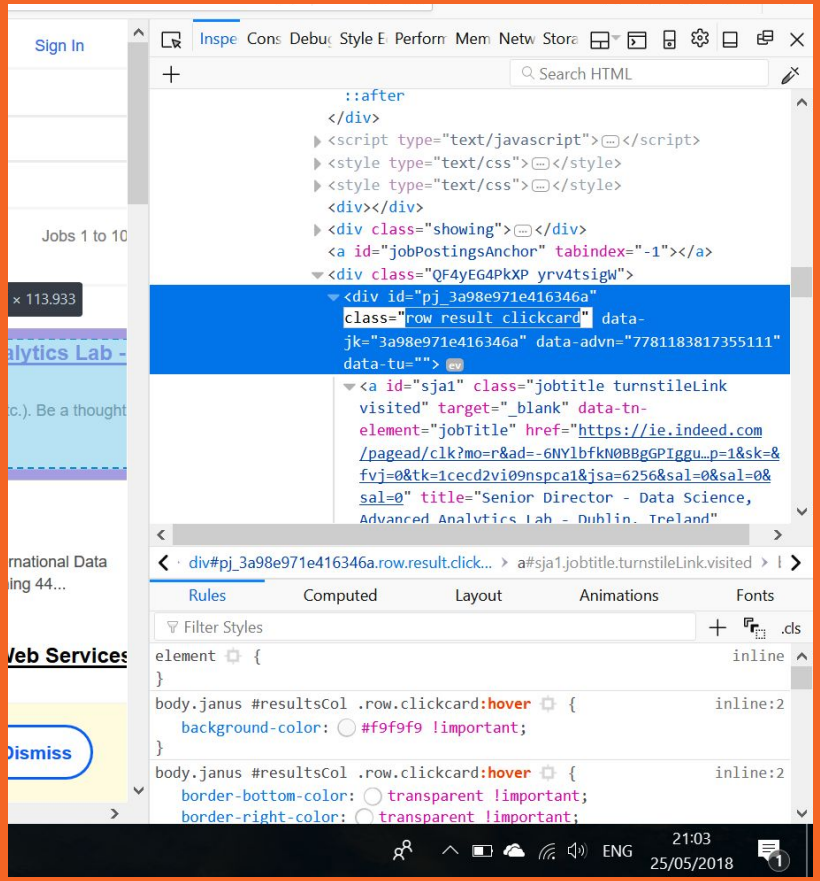

https://ie.indeed.com/jobs?q=data science&l=Dublin
Find Jobs
Find CVs
Employers / Post Job
Submit Your CV
Sign In
indeed
What
Where
data science
Dublin
Find Jobs
Advanced Job Search
job title, keywords or company
city or county
data science jobs in Dublin
Recommended Jobs - 10 new
SORT BY:
Relevance - Date
Distance:
within 25 kilometers
Job Type
Full-time (344)
Permanent (231)
Contract (58)
Internship (12)
Part-time (6)
More »
New! Join Indeed Prime - Get offers from great tech companies
Jobs 1 to 10 of 1,142
Show: All Jobs - 307 New Jobs
Senior Director - Data Science, Advanced Analytics Lab - Dub...
UnitedHealth Group - 11,464 re
Computer Science, Science, Engineering, Mathe
etc.). Be a thought - leader and expounder of dat
methods and applications....
Sponsored Save Job
Sr Software Engineer - Cloud Data
Veritas Technologies - 244 revi
Data is an organization's digital currency; Did you
International Data Corporation (IDC) predicts that
every 2 years, reaching 44...
Sponsored Save Job
Open Link in New Tab
Open Link in New Window
Open Link in New Private Window
Bookmark This Link
Save Link As...
Save Link to Pocket
Copy Link Location
Search Google for "Senior Director..."
Send Link to Device
Inspect Element (Q)
Get new jobs for this search by email
My email:
email with jobs recommended just for
alert or receiving recommended jobs, you
ms. You can change your consent settings
nsubscribing or as detailed in our terms.
Company with data science jobs

Example


https://ie.indeed.com/jobs?q=data science&l=Dublin
Find Jobs
Find CVs
Employers / Post Job
Submit Your CV
Sign In
indeed
What
Where
data science
Dublin
job title, keywords or company
city or county
data science jobs in Dublin
New! Join Indeed Prime - Get offers from great tech companies
Jobs 1 to 10
Recommended Jobs - 10 new
Show: All Jobs
div#pj_3a98e971e416346a.row.result.clickcard 529.617 × 113.933
SORT BY:
Relevance - Date
Distance:
within 25 kilometers
Senior Director - Data Science, Advanced Analytics Lab -
UnitedHealth Group - 11,464 reviews - Dublin
Computer Science, Science, Engineering, Mathematics, Statistics etc.). Be a thought leader and expounder of data science methods and applications....
Sponsored Save Job
Job Type
Full-time (344)
Permanent (231)
Contract (58)
Internship (12)
Part-time (6)
Sr Software Engineer - Cloud Data Protection
Veritas Technologies - 244 reviews - Dublin
Data is an organization's digital currency; Did you know that the International Data Corporation (IDC) predicts that data will double every 2 years, reaching 44...
Sponsored Save Job
Cloud Support Associate (Security) - Amazon Web Services
Amazon.com - 23,651 reviews - Dublin
Search HTML
<style type="text/css"></style>

<a id="sja1" class="jobtitle turnstileLink visited" target="_blank" data-tn-element="jobTitle" href="https://ie.indeed.com/pagead/clk?mo=r&ad=-6NYlbfkN0BBgGPIggu...p=1&sk=&fvj=0&tk=1cecd2vi09nspca1&jsa=6256&sal=0&sal=0&sal=0" title="Senior Director - Data Science, Advanced Analytics Lab - Dublin, Ireland" rel="noopener nofollow" onmousedown="sjomd('sja1'); clk('sja1');" onclick="setRefineByCookie([]); sjoc('sja1',0);
div#pj_3a98e971e416346a.row.result.click... a#sja1.jobtitle.turnstileLink.visited
Rules
Computed
Layout
Animations
Fonts
Filter Styles
element {
}

Example


Sign In
Jobs 1 to 10
× 113.933
alytics Lab -
c.). Be a thought
rnational Data
ing 44...
Web Services
Dismiss
Inspe Cons Debu Style E Perform Mem Netw Stora
Search HTML
::after

<script type="text/javascript">…</script>
<style type="text/css">…</style>
<style type="text/css">…</style>

…

<div id="pj_3a98e971e416346a"
class="row result clickcard" data-
jk="3a98e971e416346a" data-advn="7781183817355111"
data-tu="">
<a id="sja1" class="jobtitle turnstileLink
visited" target="_blank" data-tn-
element="jobTitle" href="https://ie.indeed.com
/pagead/clk?mo=r&ad=-6NYlbfkN0BBgGPIggu…p=1&sk=&
fvj=0&tk=1cecd2vi09nspca1&jsa=6256&sal=0&sal=0&
sal=0" title="Senior Director - Data Science,
Advanced Analytics Lab - Dublin, Ireland"
div#pj_3a98e971e416346a.row.result.click… a#sja1.jobtitle.turnstileLink.visited
Rules Computed Layout Animations Fonts
Filter Styles .cls
element { inline
}
body.janus #resultsCol .row.clickcard:hover { inline:2
background-color: #f9f9f9 !important;
}
body.janus #resultsCol .row.clickcard:hover { inline:2
border-bottom-color: transparent !important;
border-right-color: transparent !important;
ENG
21:03
25/05/2018

Resources

- https://automatetheboringstuff.com/chapter11/
- https://www.seleniumhq.org/docs/03_webdriver.jsp
- https://towardsdatascience.com/scraping-the-internets-most-popular-websites-a4c6f0be382d
- https://github.com/cs109/content
  - http://nbviewer.jupyter.org/github/cs109/content/blob/master/lec_04_scraping.ipynb