

Introduction to web scraping with Scrapy

Dr Caterina Constantinescu

Tesco Bank

July 7, 2020

Outline

- 1 About me
- 2 Web scraping options
- 3 Example 1: LookFantastic site
- 4 Example 2: National Careers Service
 - Creating a Scrapy project
 - Get overarching job categories
 - Get individual job names & descriptions
 - Get individual job salaries
- 5 Limitations

Outline

- 1 [About me](#)
- 2 [Webscraping options](#)
- 3 [Example 1: LookFantastic site](#)
- 4 [Example 2: National Careers Service](#)
 - [Creating a Scrapy project](#)
 - [Get overarching job categories](#)
 - [Get individual job names & descriptions](#)
 - [Get individual job salaries](#)
- 5 [Limitations](#)

About me



caterina.constantinescu@gmail.com





<https://datapowered.io/>



@c__constantine



CaterinaC

-
- Data scientist @ Tesco Bank
 - PhD & MSc in Psychology/-ical Research
 - R user since 2012
 - Ex-EdinbR organiser:  @edinb_r  EdinburghRusers
 - Previous webscraping experiences/attempts:
 - `import.io`
 - The Data Lab Accelerator: [Matching messy text to Standard Occupation Classifications](#)
 - Trying to expand my Python knowledge!

About me



Outline

- 1 About me
- 2 Webscraping options
- 3 Example 1: LookFantastic site
- 4 Example 2: National Careers Service
 - Creating a Scrapy project
 - Get overarching job categories
 - Get individual job names & descriptions
 - Get individual job salaries
- 5 Limitations

1 BeautifulSoup

- Module that can be used for pulling data out of HTML / XML documents.
- User friendly
- Relies on other packages to send requests and parse
- Pretty slow

1 BeautifulSoup

- Module that can be used for pulling data out of HTML / XML documents.
- User friendly
- Relies on other packages to send requests and parse
- Pretty slow

2 Selenium

- Tool for automated testing of web apps & automated browser actions
- Not for scraping, but can be used this way
- Unlike BeautifulSoup, can send web requests and also comes with a parser
- Loads JavaScript and can help access data behind JS
- Web scrapers that use either Scrapy or BeautifulSoup make use of Selenium if they require data that can only become available after JS is loaded.

1 BeautifulSoup

- Module that can be used for pulling data out of HTML / XML documents.
- User friendly
- Relies on other packages to send requests and parse
- Pretty slow

2 Selenium

- Tool for automated testing of web apps & automated browser actions
- Not for scraping, but can be used this way
- Unlike BeautifulSoup, can send web requests and also comes with a parser
- Loads JavaScript and can help access data behind JS
- Web scrapers that use either Scrapy or BeautifulSoup make use of Selenium if they require data that can only become available after JS is loaded.

3 Scrapy

- Web crawling and scraping framework
- For building complex scrapers and includes a lot of functionality
- Portable (does not rely on other libraries)
- Designed to feed into a pipeline and can handle proxies as well
- Best performance in the group
- Not beginner-friendly
- Does not handle JavaScript: must send Ajax requests to get data hidden behind JavaScript events or use Selenium

Outline

- 1 About me
- 2 Web scraping options
- 3 Example 1: LookFantastic site
- 4 Example 2: National Careers Service
 - Creating a Scrapy project
 - Get overarching job categories
 - Get individual job names & descriptions
 - Get individual job salaries
- 5 Limitations

Example 1: LookFantastic site

(<https://datapowered.io/post/2020-04-14-post-getting-stuck-in-with-scrapy/>)

Thomas Laetsch, [Web Scraping in Python](#)

LOOKFANTASTIC

[My Account](#) [My Basket](#)

[Brands](#) [Summer](#) [New](#) [MAC](#) [Hair](#) [Makeup](#) [Skin](#) [Body](#) [Tools](#) [Fragrance](#) [Men](#) [Offers](#) [Build a Routine](#) [Beauty Box](#) [Blog](#)

FREE UK Delivery Over £25 15% off your first order | Use code NEWLF FREE UK Next Day Delivery Over £100 Download Our App For Exclusive Offers


Save 10% on your order and 15% when you spend £60. Use code: **BEAUTY**

Home

Lookfantastic Discount & Voucher Codes

For the **best beauty deals**, take a look at the latest voucher codes from Lookfantastic and see if you can get a special discount today!

Browse below for the full list of our current offers- treat yourself with extra gifts with purchase, multi-buy offers and exclusive discounts. Whether you're looking for make-up, haircare, skincare or body products, you'll find the best beauty deals online at Lookfantastic. Make sure you're quick and shop now to take advantage of our special offers before they go!




OFFER

Ends: 26 Sep 2020 12:00 AM

Save 25% on selected La Roche Posay

Save 25% on selected La Roche Posay. Discount has been applied to the RRP. Offer valid for a limited time only.

SHOP THE OFFER



OFFER

Ends: 02 Sep 2020 12:00 AM

Save 25% on selected Belief The Body

Example 1: LookFantastic site

(<https://datapowered.io/post/2020-04-14-post-getting-stuck-in-with-scrapy/>)

```
1 from scrapy import Selector
2 import requests
3 import pandas as pd
4
5 url = "https://www.lookfantastic.com/voucher-codes.list"
6 html = requests.get(url).content
7 response = Selector(text=html)
8
9 # The start of the output using get() method looks something like this:
10 response.get()
11 # '<html lang="en-gb" xml:lang="en-gb" dir="ltr" xmlns:og="http://opengraphprotocol.org...'
12
13 # Let's use Selectors to extract information:
14 response.xpath('//*') # Grabs all elements and returns a list of Selectors (a SelectorList)
15 response.css('title::text').get() # Gets page title
16 response.css('title::text').extract_first()
17
18 # response.get() / response.getall()
19 # response.extract_first() / response.extract()
```

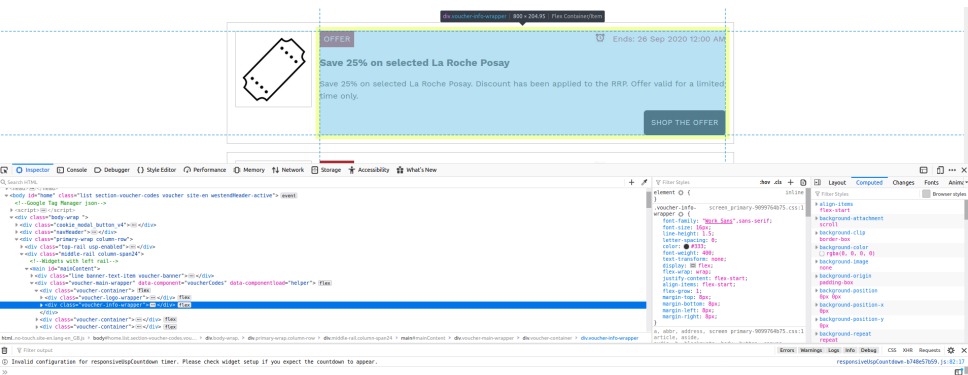
Example 1: LookFantastic site

(<https://datapowered.io/post/2020-04-14-post-getting-stuck-in-with-scrapy/>)

```
1 len(response.xpath('//*[@*]').re(r'3 for 2')) # extract using regular expressions
2 len(response.xpath('//*[@*]').re(r'Omorovicza'))
3 # Watch out - this will capture everything including the menu header!
4 # e.g., it will pick up on Omorovicza, even if it is not listed as part of the offers in the page body.
5
6 # xpath version:
7 response.xpath('/html/body/div[1]/div[3]/section/div[2]/div[1]').extract()
8 # And the css selector version:
9 response.css('html > body > div:nth-of-type(1) > div:nth-of-type(3) > section > div:nth-of-type(2) >
   div:nth-of-type(1)').extract()
10
11
12 # We could also have gone directly by class to identify the same info.
13 # This method will pick up on anything that has this specific class (and no others!):
14 response.xpath('//*[@div[@class="voucher-main-wrapper"]')
15 # In case the same element could have had multiple classes, we could have picked up on it as:
16 response.xpath('//*[@div[contains(@class, "voucher-main-wrapper")]')
17 # CSS selector version, targeting the desired class directly:
18 response.css('.voucher-main-wrapper')
```

Example 1: LookFantastic site

(<https://datapowered.io/post/2020-04-14-post-getting-stuck-in-with-scrapy/>)



Firefox's Web developer tools: Inspector

Example 1: LookFantastic site

(<https://datapowered.io/post/2020-04-14-post-getting-stuck-in-with-scrapy/>)

```
1 individual_offers = response.css('voucher-info-wrapper') # Grab all the individual offer boxes
2
3 # 1. Title
4 all_offer_titles = individual_offers.css(".voucher-title ::text").getall()
5
6 # 2. Text
7 individual_offers.css(".voucher-message ::text").getall() # Wrong length! (!= 245)
8 # Hence we need to deal with each element at a time, and keep output the same length as other info we are pulling
   out:
9 all_offer_messages = []
10 for offer_text in individual_offers:
11     # Sometimes the same message is split across multiple list elements. So we'll concatenate them, with no
       separating space between them:
12     current_offer_text = ''.join(offer_text.css(".voucher-message ::text").getall())
13     all_offer_messages.append(current_offer_text)
14
15 # 3. Offer type
16 # Most offers have a specific class (e.g., 'offer-products-money-off'), always starting with 'offer-'.
17 # But also a generic 'voucher-label' class.
18 # Loop needed here - otherwise any offers that happen to NOT have a second class get left out of output
19 offer_types = []
20 for offer in individual_offers:
21     current_offer_label = offer.css('voucher-label').xpath("@class").re(r'offer-.*')
22     offer_types.append(current_offer_label)
```

Example 1: LookFantastic site

(<https://datapowered.io/post/2020-04-14-post-getting-stuck-in-with-scrapy/>)

```
1 # 4. End date
2 all_offers_end_date = response.css(".voucher-end-date::text").getall()
3 all_offers_end_date = [deadline.replace("\t", "") for deadline in all_offers_end_date]
4 all_offers_end_date = [deadline.replace("\n", "") for deadline in all_offers_end_date]
5 all_offers_end_date = [deadline.replace("Ends:", "") for deadline in all_offers_end_date]
6
7 # 5. URL
8 response.css(".voucher-button::attr(href)").getall() # Still wrong length, so here we go again:
9
10 def xstr(s):
11     if s is None:
12         return ''
13     return str(s)
14
15 all_offers_URL = []
16 for link in individual_offers:
17     current_offer_label = "https://www.lookfantastic.com" +
18         xstr(link.css(".voucher-button::attr(href)").get())
19     all_offers_URL.append(current_offer_label)
```


Example 1: LookFantastic site

(<https://datapowered.io/post/2020-04-14-post-getting-stuck-in-with-scrapy/>)

```
1 # Now bring all the info together:
2 offer_df = pd.DataFrame({
3     'Title': all_offer_titles,
4     'Text': all_offer_messages, # messages_stripped_of_html
5     # Flattening list of lists, also allowing for empty list elements:
6     'Type': pd.DataFrame(offer_types).fillna("").squeeze().tolist(),
7     'End': all_offers_end_date,
8     'URL': all_offers_URL
9 })
10
11 offer_df['Type'] = offer_df['Type'].str.replace('offer-products-', '')
12 offer_df['Type'] = offer_df['Type'].str.replace('-', ' ').str.capitalize()
```

So what does offer_df look like?

Example 1: LookFantastic site

(<https://datapowered.io/post/2020-04-14-post-getting-stuck-in-with-scrapy/>)

	Title	Text	Type	End	
161	Receive a complimentary MAC Fix + Setting Spray ...	Receive a complimentary Fix+ sample with £45 spend on M-A-C, and receive a complimentary MAC S...	Offer min spend free gift	Sat May 30 00:00:00 BST 20...	https://www.lookfantastic.com/brands/mac/vi
162	Save 30% L'Oréal Men Expert...	Save 30% L'Oréal Men Expert. Discount has been applied to the RRP. Offer valid for a limited time...	Money off	Wed Apr 22 00:00:00 BST 2...	https://www.lookfantastic.com/brands/l-oreal
163	Save 20% on selected Moschino.	Save 20% on selected Moschino. Discount has been applied to the RRP. Offer valid for a limited tim...	Money off	Wed Apr 22 00:00:00 BST 2...	https://www.lookfantastic.com/brands/mosch
164	Enjoy a complimentary Hairburst washbag when y...	Enjoy a complimentary Hairburst washbag when you spend £80 on the brand. Complimentary gift...	Offer min spend free gift	Sat May 23 00:00:00 BST 20...	https://www.lookfantastic.com/brands/hairbu
165	Receive a complimentary Giorgio Armani SI LF Exc...	Receive a complimentary Giorgio Armani SI LF Exclusive when you buy any SI Perfume. Compliment...	Free gift	Sat May 30 00:00:00 BST 20...	https://www.lookfantastic.com/brands/giorgio
166	50% off selected Elizabeth Arden	50% off selected Elizabeth Arden. Discount has been applied to the RRP. Offer valid for a limited ti...	Money off	Sat Apr 25 00:00:00 BST 2020	https://www.lookfantastic.com/brands/elizabi
167	Save 25% on selected TriPollar.	Save 25% on selected TriPollar. Discount has been applied to the RRP. Offer valid for a limited time...	Money off	Sat Jun 27 00:00:00 BST 20...	https://www.lookfantastic.com/brands/tripolli
168	Save 20% on selected Caudalie.	Save 20% on selected Caudalie, plus receive a complimentary Hydration Perfection GWP (worth £2...	Money off	Fri Aug 21 00:00:00 BST 2020	https://www.lookfantastic.com/brands/caudal
169	Ultrasun 20% off selected products.	Receive 20% off selected Ultrasun products. Discount has been applied to the RRP. Offer valid for ...	Money off	Wed Apr 22 00:00:00 BST 2...	https://www.lookfantastic.com/brands/ultrasu
170	Receive a complimentary Paul Mitchell Girl Power ...	Receive a complimentary Paul Mitchell Girl Power Thermal Bag when you spend £50 on the brand. C...	Offer min spend free gift	Wed May 13 00:00:00 BST 2...	https://www.lookfantastic.com/brands/paul-m
171	Save 25% on selected Vichy.	Save 25% on selected Vichy. Discount has been applied to the RRP. Offer valid for a limited time only.	Money off	Tue Jul 28 00:00:00 BST 2020	https://www.lookfantastic.com/brands/vichy/s
172	Save 30% on selected Decléor	Save 30% on selected Decléor. Discount has been applied to the RRP. Offer valid for a limited time...	Money off	Wed May 06 00:00:00 BST 2...	https://www.lookfantastic.com/brands/decléor
173	Enjoy a complimentary Grow Gorgeous Volume R...	Enjoy a complimentary Grow Gorgeous Volume Root Stimulating Primer 200ml when you spend £6...	Offer min spend free gift	Sat Jun 27 00:00:00 BST 20...	https://www.lookfantastic.com/brands/grow-g
174	L'Oréal Professionnel Série Expert Pro Longer L...	L'Oréal Professionnel Série Expert Pro Longer Lengths Renewing Cream 150ml when you spend £4...	Offer min spend free gift	Sat Jun 27 00:00:00 BST 20...	https://www.lookfantastic.com/brands/l-oreal
175	Enjoy a complimentary Wella hair ring when you s...	Enjoy a complimentary Wella hair ring when you spend £80 on the brand. Complimentary gift will b...	Offer min spend free gift	Sat May 23 00:00:00 BST 20...	https://www.lookfantastic.com/brands/wella-s
176	Receive a complimentary Yves Saint Laurent Pure ...	Complimentary gift will be awarded at the basket. Offer valid for a limited time only, while stocks L...	Offer min spend free gift	Sat Jun 27 00:00:00 BST 20...	https://www.lookfantastic.com/brands/yves-sa

The coveted output

Outline

- 1 About me
- 2 Webscraping options
- 3 Example 1: LookFantastic site
- 4 Example 2: National Careers Service**
 - Creating a Scrapy project
 - Get overarching job categories
 - Get individual job names & descriptions
 - Get individual job salaries
- 5 Limitations

Outline

- 1 About me
- 2 Webscraping options
- 3 Example 1: LookFantastic site
- 4 Example 2: National Careers Service
 - Creating a Scrapy project
 - Get overarching job categories
 - Get individual job names & descriptions
 - Get individual job salaries
- 5 Limitations

Setting up a Scrapy project

- Scrapy has no need for the requests package, but it's handy for:
 - Learning purposes (J. Laetsch's approach)
 - To check what we are doing as we progress
- Here is a better way:

```
1 cd /path/to/my/desired/proj/location
2 scrapy startproject MyProjName
3 # Work on your spiders in that folder, then:
4 scrapy crawl mySpidersClassName -o /path/to/my/upcoming/scraped/output.csv
```

Setting up a Scrapy project

The screenshot displays the PyCharm IDE interface with a Scrapy project named 'ScrapingOnlineOffers' open. The left sidebar shows the project structure, with the 'spiders' directory highlighted. The main editor window shows the 'JobSpider.py' file, which contains the following code:

```
1 import scrapy
2
3 class JobSpider(scrapy.Spider):
4     name = "jobs"
5     start_urls = [
6         "https://nationalcareers.service.gov.uk/explore-careers"
7     ]
8
9     def parse(self, response):
10         for category in response.css('h2.homepage-jobcategories > li'):
11             yield {
12                 'category_name': category.css('::text').get(),
13                 'link': category.css("a::attr(href)").get()
14             }
15
16 # In terminal: scrapy crawl jobs -o jobs.json
```

The bottom of the IDE shows the Python Console with the following output:

```
Python Console
/home/caterina/anaconda3/bin/python3.7 /home/caterina/.local/share/JetBrains/Toolbox/apps/PyCharm-E/ch-0/193.6494.60/plugins/python-ce/helpers/pydev/pydevco
Python 3.7.6 (default, Jan 8 2020, 10:59:22)
In[21]:
```

The right sidebar shows the Event Log with the following message:

```
Event Log
05/07/2020
12:54 PyCharm and
Plugin: EduTool
Update...
```

Outline

- 1 About me
- 2 Web scraping options
- 3 Example 1: LookFantastic site
- 4 Example 2: National Careers Service
 - Creating a Scrapy project
 - Get overarching job categories
 - Get individual job names & descriptions
 - Get individual job salaries
- 5 Limitations

Example 2: National Careers Service

National Careers Service

[Explore careers](#) [Skills assessment](#) [Find a course](#) [Contact us](#) [About us](#) [Help to get a job](#)

BETA Complete [Ipsos MORI survey](#) to give us your feedback about the service.

Explore careers

Find out what a job involves and if it's right for you.



Explore by job category

[Administration](#)

[Animal care](#)

[Beauty and wellbeing](#)

[Business and finance](#)

[Computing, technology and digital](#)

[Construction and trades](#)

[Creative and media](#)

[Delivery and storage](#)

[Emergency and uniform services](#)

[Engineering and maintenance](#)

[Environment and land](#)

[Government services](#)

[Healthcare](#)

[Home services](#)

[Hospitality and food](#)

[Law and legal](#)

[Managerial](#)

[Manufacturing](#)

[Retail and sales](#)

[Science and research](#)

[Social care](#)

[Sports and leisure](#)

[Teaching and education](#)

[Transport](#)

[Travel and tourism](#)



Example 2: National Careers Service

```
1 import scrapy
2
3 class jobsSpider(scrapy.Spider):
4     name = "jobs"
5     start_urls = [
6         "https://nationalcareers.service.gov.uk/explore-careers"
7     ]
8     def parse(self, response):
9         for category in response.css('.homepage-jobcategories > li'):
10             yield {
11                 'category_name' : category.css(":text").get(),
12                 'link' : category.css("a::attr(href)").get()
13             }
14
15 # In terminal: scrapy crawl jobs -o jobs.json
```

Example 2: National Careers Service

```
ScrapingLookFantastic.py x JobClassifications.py x JobSpider.py x jobs.json x JobSpiderNestedLinks.py x
1  [
2  {"category_name": "Administration", "link": "/job-categories/administration"},
3  {"category_name": "Animal care", "link": "/job-categories/animal-care"},
4  {"category_name": "Beauty and wellbeing", "link": "/job-categories/beauty-and-wellbeing"},
5  {"category_name": "Business and finance", "link": "/job-categories/business-and-finance"},
6  {"category_name": "Computing, technology and digital", "link": "/job-categories/computing-technology-and-digital"},
7  {"category_name": "Construction and trades", "link": "/job-categories/construction-and-trades"},
8  {"category_name": "Creative and media", "link": "/job-categories/creative-and-media"},
9  {"category_name": "Delivery and storage", "link": "/job-categories/delivery-and-storage"},
10 {"category_name": "Emergency and uniform services", "link": "/job-categories/emergency-and-uniform-services"},
11 {"category_name": "Engineering and maintenance", "link": "/job-categories/engineering-and-maintenance"},
12 {"category_name": "Environment and land", "link": "/job-categories/environment-and-land"},
13 {"category_name": "Government services", "link": "/job-categories/government-services"},
14 {"category_name": "Healthcare", "link": "/job-categories/healthcare"},
15 {"category_name": "Home services", "link": "/job-categories/home-services"},
16 {"category_name": "Hospitality and food", "link": "/job-categories/hospitality-and-food"},
17 {"category_name": "Law and legal", "link": "/job-categories/law-and-legal"},
18 {"category_name": "Managerial", "link": "/job-categories/managerial"},
19 {"category_name": "Manufacturing", "link": "/job-categories/manufacturing"},
20 {"category_name": "Retail and sales", "link": "/job-categories/retail-and-sales"},
21 {"category_name": "Science and research", "link": "/job-categories/science-and-research"},
22 {"category_name": "Social care", "link": "/job-categories/social-care"},
23 {"category_name": "Sports and leisure", "link": "/job-categories/sports-and-leisure"},
24 {"category_name": "Teaching and education", "link": "/job-categories/teaching-and-education"},
25 {"category_name": "Transport", "link": "/job-categories/transport"},
26 {"category_name": "Travel and tourism", "link": "/job-categories/travel-and-tourism"}
27 ]
```

Outline

- 1 About me
- 2 Webscraping options
- 3 Example 1: LookFantastic site
- 4 Example 2: National Careers Service
 - Creating a Scrapy project
 - Get overarching job categories
 - Get individual job names & descriptions
 - Get individual job salaries
- 5 Limitations

Example 2: National Careers Service

National Careers Service

[Explore careers](#) [Skills assessment](#) [Find a course](#) [Contact us](#) [About us](#) [Help to get a job](#)

BETA Complete [Josias MORI survey](#) to give us your feedback about the service.

[Home](#): [Explore careers](#) > [Government services](#)

Government services

Air accident investigator

Air accident engineering inspector, air accident operations inspector

Air accident investigators search for the causes of accidents and serious incidents, involving civilian aircraft.

Army officer

Professionally qualified officer, officer reserve, commissioned officer

Army officers command, manage and motivate teams of soldiers.

Assistant immigration officer

Assistant immigration officers check that people have the right to visit or stay in the UK.

Bodyguard

Close protection officer, CPO

Bodyguards protect individuals or groups from the risk of violence, kidnapping and other harmful situations.

Bomb disposal technician

Bomb disposal technicians identify, defuse and destroy explosive devices.

Border Force officer

Other job categories

[Administration](#)

[Animal care](#)

[Beauty and wellbeing](#)

[Business and finance](#)

[Computing, technology and digital](#)

[Construction and trades](#)

[Creative and media](#)

[Delivery and storage](#)

[Emergency and uniform services](#)

[Engineering and maintenance](#)

[Environment and land](#)

[Healthcare](#)

[Home services](#)

[Hospitality and food](#)

[Law and legal](#)

[Managerial](#)

[Manufacturing](#)

[Retail and sales](#)

[Science and research](#)

[Social care](#)

[Sports and leisure](#)

[Teaching and education](#)

[Transport](#)

[Travel and tourism](#)

About me
Web scraping options
Example 1: LookFantastic site
Example 2: National Careers Service
Limitations

Creating a Scrapy project
Get overarching job categories
Get individual job names & descriptions
Get individual job salaries

Example 2: National Careers Service

The screenshot shows the National Careers Service website. The main heading is "Government services". Below it, there are several job categories listed, each with a brief description. The categories are: Air accident investigator, Army officer, and Assistant immigration officer. To the right, there is a section titled "Other job categories" with a list of links: Administration, Animal care, Beauty and wellbeing, Business and finance, Computing, technology and digital, Construction and trades, Creative and media, Delivery and storage, Emergency and uniform services, Engineering and maintenance, Environment and land, Healthcare, Home services, and Hospitality and food.

The developer tools are open, showing the HTML structure of the page. The selected element is a link with the text "Air accident investigator". The HTML structure is as follows:

```
<ul class="job-categories_items">
  <li class="job-categories_item">
    <h3 class="text-secondary font-size-18">Air accident investigator</h3>
    <p class="font-size-14">Air accident engineering inspector, air accident operations inspector</p>
    <p class="font-size-14">Air accident investigators search for the causes of accidents and serious incidents, involving civilian aircraft.</p>
  </li>
  <li class="job-categories_item">
    <h3 class="text-secondary font-size-18">Army officer</h3>
    <p class="font-size-14">Professionally qualified officer, officer reserve, commissioned officer</p>
    <p class="font-size-14">Army officers command, manage and motivate teams of soldiers.</p>
  </li>
  <li class="job-categories_item">
    <h3 class="text-secondary font-size-18">Assistant immigration officer</h3>
    <p class="font-size-14">Assistant immigration officers help to process applications for visas and entry clearance</p>
  </li>
</ul>
```

Example 2: National Careers Service

The screenshot shows the National Careers Service website. The main heading is "Government services". Below it, there are three job categories listed: "Air accident investigator", "Army officer", and "Assistant immigration officer". Each category has a brief description. To the right, there is a section titled "Other job categories" with a list of links: Administration, Animal care, Beauty and wellbeing, Business and finance, Computing, technology and digital, Construction and trades, Creative and media, Delivery and storage, Emergency and uniform services, Engineering and maintenance, Environment and land, Healthcare, Home services, and Hospitality and food.

The developer tools are open, showing the HTML structure of the page. The selected element is a list of job categories, with the following HTML structure:

```
<ul class="job-categories_items">
  <li class="job-categories_item">
    <h2>Air accident investigator</h2>
    <p>Air accident engineering inspector, air accident operations inspector</p>
    <p>Air accident investigators search for the causes of accidents and serious incidents, involving civilian aircraft.</p>
  </li>
  <li class="job-categories_item">
    <h2>Army officer</h2>
    <p>Professionally qualified officer, officer reserve, commissioned officer</p>
    <p>Army officers command, manage and motivate teams of soldiers.</p>
  </li>
  <li class="job-categories_item">
    <h2>Assistant immigration officer</h2>
    <p>Assistant immigration officers help to control the entry of people into the United Kingdom.</p>
  </li>
</ul>
```

The console shows a message: "JOMIGRATE: Migrate is installed, version 1.4.1".

Example 2: National Careers Service

The screenshot shows the National Careers Service website. The header includes navigation links: Explore careers, Skills assessment, Find a course, Contact us, About us, and Help to get a job. A BETA banner promotes a survey. The main heading is "Government services". Below it, there are sections for "Air accident" (with a text search link), "Army officer", and "Assistant immigration officer". A sidebar on the right lists "Other job categories" such as Administration, Animal care, Beauty and wellbeing, Business and finance, Computing, technology and digital, Construction and trades, Creative and media, Delivery and storage, Emergency and uniform services, Engineering and maintenance, Environment and land, Healthcare, Home services, and Hospitality and food. The bottom of the image shows a web browser interface with developer tools open, displaying the HTML structure of the page. The HTML shows a container for job categories with a search link and a list of categories. The developer tools also show the CSS styles for the page, including a computed style for the job categories list.

National Careers Service
Explore careers Skills assessment Find a course Contact us About us Help to get a job

BETA Complete [James MORI survey](#) to give us your feedback about the service.

Home: Explore careers > Government services

Government services

Air accident [Find text secondary font small meta dfc code search goAllTitle](#) 630 x 20

Air accident engineering inspector, air accident operations inspector
Air accident investigators search for the causes of accidents and serious incidents, involving civilian aircraft.

Army officer
Professionally qualified officer, officer reserve, commissioned officer
Army officers command, manage and motivate teams of soldiers.

Assistant immigration officer

Other job categories

- [Administration](#)
- [Animal care](#)
- [Beauty and wellbeing](#)
- [Business and finance](#)
- [Computing, technology and digital](#)
- [Construction and trades](#)
- [Creative and media](#)
- [Delivery and storage](#)
- [Emergency and uniform services](#)
- [Engineering and maintenance](#)
- [Environment and land](#)
- [Healthcare](#)
- [Home services](#)
- [Hospitality and food](#)

Inspector Console Debugger Style Editor Performance Memory Network Storage Accessibility What's New

Search HTML

```
<div class="job-categories-items">
  <div class="job-categories-item">
    <h3>Air accident</h3>
    <p>Air accident engineering inspector, air accident operations inspector</p>
    <p>Air accident investigators search for the causes of accidents and serious incidents, involving civilian aircraft.</p>
  </div>
  <div class="job-categories-item">
    <h3>Army officer</h3>
    <p>Professionally qualified officer, officer reserve, commissioned officer</p>
    <p>Army officers command, manage and motivate teams of soldiers.</p>
  </div>
  <div class="job-categories-item">
    <h3>Assistant immigration officer</h3>
  </div>
</div>
```

body.js-enabled main > divPublicWrapper > divPublicWrapper > divGrid-row > divMainContent_74475353DF001_Cat000_of_... > ul.job-categories-items > li.job-categories-item > p.dfc-code-search-goAllTitle

JOMIGRATE: Migrate is installed, version 1.4.1

Filter output

Errors: Warnings: Logs: Info: Debug: CSS: XHR: Requests: jquerybundle.min.js:1:89827

Example 2: National Careers Service

```
1 import scrapy
2
3 root = "https://nationalcareers.service.gov.uk/explore-careers"
4
5 class nestedJobSpider(scrapy.Spider):
6     name = "jobDetails"
7
8     def start_requests(self):
9         yield scrapy.Request(url=root, callback=self.parse)
10
11     def parse(self, response):
12         links = response.css('.homepage-jobcategories > li a::attr(href)').extract()
13         for link in links:
14             yield response.follow(url = link, callback = self.parse2)
```

Example 2: National Careers Service

```
1 def parse2(self, response):
2     parent_job_category = response.css('.heading-xlarge::text').extract()
3
4     job_list_items = response.css('.job-categories_item')
5     for job in job_list_items:
6         j_name = job.css('.dfc-code-search-jpTitle::text').extract_first()
7
8         alt_j_name = job.css(".dfc-code-search-jpAltTitle::text").extract_first()
9         if not alt_j_name:
10             alt_j_name = "None"
11
12         j_descr = job.css('.dfc-code-search-jpOverview::text').extract_first()
13
14         # print(j_name)
15         yield {
16             'ParentCat': parent_job_category,
17             'JobName': j_name,
18             'AltJobName': alt_j_name,
19             'JobDescr': j_descr
20         }
21
22 # In terminal: scrapy crawl jobDetails -o /path/to/my/jobDetails.json
```

Example 2: National Careers Service

```
ScrapingLookFantastic.py x JobClassifications.py x JobSpider.py x jobs.json x JobDetails.json x JobDetails.csv x JobSpiderNestedLinks.py x JobDetailsWithSalary.json x
1 ParentCat,JobName,AltJobName,JobDescr
2 Administration,Accounting technician,None,Accounting technicians handle day-to-day financial matters in all types of business.
3 Administration,Admin assistant,Office administrator, clerical assistant, administrative assistant,Admin assistants give support to offices by organising meetings, typing documents and updating compa
4 Administration,Arts administrator,None,Arts administrators help to organise events and exhibitions, manage staff, and look after buildings like theatres or museums.
5 Administration,Assistant immigration officer,None,Assistant immigration officers check that people have the right to visit or stay in the UK.
6 Administration,Auditor,None,Internal and external auditors check organisations' financial records and procedures to make sure they are accurate and efficient.
7 Administration,Bid writer,None,Bid writers prepare documents used to pitch for contracts to provide services, or to apply for project funding.
8 Administration,Bilingual secretary,None,Bilingual secretaries provide administrative services in English and one or more foreign languages.
9 Administration,Bookkeeper,Accounts clerk,Bookkeepers keep financial records up to date and help prepare accounts.
10 Administration,Border force officer,None,Border force officers protect UK border entry points like ports and airports, by enforcing immigration and customs regulations.
11 Administration,Car rental agent,Car rental assistant, vehicle reservation agent,Car rental agents hire out and lease vehicles to businesses and the public.
12 Administration,Charity fundraiser,None,Charity fundraisers organise events and activities to encourage people to donate to causes and organisations.
13 Administration,Civil Service administrative officer,None,Civil Service administrative officers work in government departments, carrying out policies and running services for the public.
14 Administration,Civil Service executive officer,None,Civil Service executive officers work in government departments that develop policies and provide services to the public.
15 Administration,Conference and exhibition manager,Conference organiser,Conference and exhibition managers plan and run events like trade shows, conferences and exhibitions.
16 Administration,Credit controller,Debt collection agent,Credit controllers help firms get the money they are owed from businesses and individuals.
17 Administration,Data entry clerk,Audio typist, copy typist,Data entry clerks type information into databases and systems and create letters, reports and other documents.
18 Administration,Diplomatic Service officer,None,Diplomatic Service officers help to promote and protect British interests, businesses and citizens overseas.
19 Administration,Estates officer,None,Estates officers are responsible for the management and upkeep of land and property belonging to local councils and public bodies.
20 Administration,European Union official,EU official,European Union (EU) officials work for institutions like the European Commission or the European Parliament.
21 Administration,Farm secretary,Agricultural business administrator, rural business administrator,Farm secretaries are responsible for the day-to-day running of the business side of farms.
22 Administration,Finance officer,Financial officer, finance clerk, treasurer,Finance officers help to manage the finances of an organisation by keeping track of its income and controlling its spending.
23 Business and finance,Accounting technician,None,Accounting technicians handle day-to-day financial matters in all types of business.
24 Business and finance,Actuary,Actuarial analyst,Actuaries work with companies and government departments, to help them forecast long-term financial costs and investment risks.
25 Business and finance,Auditor,None,Internal and external auditors check organisations' financial records and procedures to make sure they are accurate and efficient.
26 Business and finance,Bank manager,Building society manager, financial institution manager or director,Bank managers oversee the day-to-day operations of their branch, supervise staff and work to attract
27 Business and finance,Banking customer service adviser,None,Banking customer service advisers provide a face-to-face service in banks and building societies.
28 Business and finance,Bookkeeper,Accounts clerk,Bookkeepers keep financial records up to date and help prepare accounts.
29 Business and finance,Business adviser,Business consultant, enterprise adviser, business coach,Business advisers give advice and support to new business start-ups and help established businesses to grow
30 Business and finance,Business development manager,None,Business development managers find new customers, and persuade existing ones to buy extra services.
31 Business and finance,Business project manager,Commercial project lead, project manager,Business project managers plan and organise people, tasks and resources to complete a project on time and within
32 Business and finance,Chief executive,Chief executive officer, CEO, managing director, businessman, businesswoman,Chief executives plan and put into place policies to help their organisations be success
33 Construction and trades,Acoustics consultant,Acoustician, acoustics engineer,Acoustics consultants help manage and control noise and vibrations in homes, workplaces and other environments.
34 Construction and trades,Architect,None,Architects design new buildings and the spaces around them, and work on the restoration and conservation of existing buildings.
35 Construction and trades,Architectural technician,None,Architectural technicians work closely with architectural teams on the design process of building projects.
36 Construction and trades,Architectural technologist,Building technologist,Architectural technologists manage all stages of the technical design and planning process of building projects.
37 Construction and trades,Boat builder,Marine craftsman, shipwright,Boat builders build, repair and refit marine craft from small sailing boats to large sea-going vessels.
38 Construction and trades,Bricklayer,Mason, brickie,Bricklayers build houses, repair walls and chimneys, and refurbish decorative stonework. They also work on restoration projects.
39 Construction and trades,Builders' merchant,None,Builders' merchants sell building and do-it-yourself products and materials to the building trade and the public.
40 Construction and trades,Building control officer,Building control surveyor,Building control officers make sure building regulations are followed.
41 Administration,Financial services customer adviser,Sales adviser, contact centre agent,Financial services customer advisers work in contact centres for banks, insurance, investment and credit companies.
42 Administration,GP practice manager,GP surgery manager, general practitioner practice manager,GP practice managers run the business side of doctors' surgeries and health centres.
43 Administration,Health and safety adviser,Health and safety officer,Health and safety advisers work to reduce accidents, injury and health problems in the workplace.
44 Administration,Health records clerk,Medical records clerk,Health records clerk,Health records clerks keep people's medical records up to date.
45 Administration,Health service manager,Hospital manager, NHS hospital manager,Health service managers run local healthcare services like hospitals, GP practices and community health services.
```

Outline

- 1 About me
- 2 Webscraping options
- 3 Example 1: LookFantastic site
- 4 Example 2: National Careers Service**
 - Creating a Scrapy project
 - Get overarching job categories
 - Get individual job names & descriptions
 - Get individual job salaries
- 5 Limitations

About me
Web scraping options
Example 1: LookFantastic site
Example 2: National Careers Service
Limitations

Creating a Scrapy project
Get overarching job categories
Get individual job names & descriptions
Get individual job salaries

Example 2: National Careers Service

The screenshot shows the National Careers Service website. The main heading is "Government services". Below it, there's a section for "Air accident investigator" with a description: "Air accident engineering inspector, air accident operations inspector". To the right, there's a list of "Other job categories" including Administration, Animal care, Beauty and wellbeing, Business and finance, Computing, technology and digital, Construction and trades, Creative and media, Delivery and storage, Emergency and uniform services, Engineering and maintenance, Environment and land, Healthcare, Home services, and Hospitality and food.

The developer tools are open, showing the HTML structure of the "Air accident investigator" section. The HTML includes a heading, a description, and a list of job categories.

```
<h2>Air accident investigator</h2>
<p>Air accident engineering inspector, air accident operations inspector</p>
<p>Air accident investigators search for the causes of accidents and serious incidents, involving civilian aircraft.</p>
<h3>Army officer</h3>
<p>Professionally qualified officer, officer reserve, commissioned officer</p>
<p>Army officers command, manage and motivate teams of soldiers.</p>
<h3>Assistant immigration officer</h3>
```

Example 2: National Careers Service

BETA Complete [Josias MORI survey](#) to give us your feedback about the service.

[Home: Explore careers](#) > [Air accident investigator](#)

Air accident investigator

Air accident engineering inspector, air accident operations inspector

Air accident investigators search for the causes of accidents and serious incidents, involving civilian aircraft.

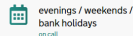
Average salary (a year)



Typical hours (a week)



You could work



How to become >

What it takes >

What you'll do >

Career path and progression >

Current opportunities >

Not what you're looking for?

Search further careers

Enter a job title



How to become an air accident investigator

You can get into this job through:

Is this page useful? [Yes](#) [No](#)

Related careers

Example 2: National Careers Service

[illegible]

Example 2: National Careers Service

```
1 import scrapy
2 from collections import defaultdict
3
4 root = "https://nationalcareers.service.gov.uk/explore-careers"
5
6 class nestedJobSpider(scrapy.Spider):
7     name = "jobDetailsWithSalary"
8
9     def start_requests(self):
10         yield scrapy.Request(url = root, callback = self.parse)
11
12     def parse(self, response):
13         links = response.css('.homepage-jobcategories > li a::attr(href)').extract()
14         for link in links:
15             yield response.follow(url = link, callback = self.parse2)
```


Example 2: National Careers Service

```
1 def parse2(self, response):
2
3     parent_job_category = response.css('.heading-xlarge::text').extract_first()
4
5     job_list_items = response.css('.job-categories_item')
6     for job in job_list_items:
7         j_name = job.css('.dfc-code-search-jpTitle::text').extract_first()
8
9         alt_j_name = job.css(".dfc-code-search-jpAltTitle::text").extract_first()
10        if not alt_j_name:
11            alt_j_name = "None"
12
13        j_descr = job.css('.dfc-code-search-jpOverview::text').extract_first()
14
15        item = defaultdict(list)
16        item['ParentJobCat'] = parent_job_category
17        item['JobName'] = j_name
18        item['AltJobName'] = alt_j_name
19        item['JobDescr'] = j_descr
20
21        fine_job_details = job.css('.dfc-code-search-jpTitle ::attr(href)').extract_first()
22
23        yield response.follow(url = fine_job_details,
24                              callback = self.parse3,
25                              meta={'item': item}) # Save item here to pick up later in parse3
26        # https://docs.scrapy.org/en/latest/topics/request-response.html#scrapy.http.Response.meta
```

Example 2: National Careers Service

```
1  def parse3(self, response):
2
3      item = response.meta['item'] # retrieve item generated in previous request
4
5      min_sal = response.css('.dfc-code-jpsstarter::text').extract_first()
6      max_sal = response.css('.dfc-code-jpsexperienced::text').extract_first()
7
8      if not min_sal:
9          min_sal = "Not specified"
10     else:
11         min_sal = min_sal.strip()
12
13     if not max_sal:
14         max_sal = "Not specified"
15     else:
16         max_sal = max_sal.strip()
17
18     # print(min_sal + " - " + max_sal)
19
20     item['MinSal'] = min_sal
21     item['MaxSal'] = max_sal
22     # print(item)
23
24     yield item
```

Example 2: National Careers Service

```
ScrapingLookFantastic.py x JobClassifications.py x JobSpider.py x Jobs.json x JobDetails.json x JobDetails.csv x JobDetailsWithSalary.csv x JobSpiderNestedLinks.py x -
1 ParentJobCat,Jobhouse,AltJobName,JobDescr_MinSal,MaxSal
2 Beauty and wellbeing,Tattooist,Body artist,Tattooists create permanent artwork on their clients' bodies.,Not specified,Not specified
3 Delivery and storage,Warehouse manager,Manager,Warehouse managers plan and co-ordinate warehouse operations at distribution depots, retail superstores and manufacturing plants.,£18,000,"£40,000"
4 Animal care,Farm worker,Farm labourer,Farm workers raise livestock and plant and harvest crops, using agricultural machinery.,£13,000,"£25,000"
5 Social care,Palliative care assistant,Specialist healthcare assistant,Palliative care assistants provide specialist end of life care and support to patients and their families.,£18,005,"£21,142"
6 Beauty and wellbeing,Reflexologist,Reflexologists apply pressure to certain parts of the hands and feet (reflexes) which they believe can help with relaxation and wellbeing.,Not specified,Not specified
7 Delivery and storage,Warehouse worker,Warehouse operative,Warehouse workers take delivery of goods and pack orders for dispatch.,£12,700,"£27,000"
8 Delivery and storage,Yard person,Yard operative,Yard people load and unload deliveries and re-order stocks at factory warehouses, building suppliers and retail distribution companies.,£13,000,"£23,000"
9 Beauty and wellbeing,Reiki healer,Reiki practitioners use their hands to help people relax and improve their wellbeing.,Not specified,Not specified
10 Business and finance,Business project manager,Commercial project lead, project manager,Business project managers plan and organise people, tasks and resources to complete a project on time and within
11 Social care,Occupational therapy support worker,Occupational therapy (OT) assistant, OT technician, rehabilitation assistant, technical instructor,Occupational therapy support workers work with occupa
12 Social care,Drug and alcohol worker,Drug and alcohol workers help people tackle their drug, alcohol or solvent misuse problems.,£17,000,"£40,000"
13 Travel and tourism,Visitor attraction general manager,Manager,Visitor attraction general managers look after the operation and finances of an attraction, and the health and safety of staff and visitors.,
14 Manufacturing,Crane driver,Crane operator,Crane drivers operate lifting machinery on construction, quarrying and mining sites, at ports and in warehouses.,£20,000,"£36,000"
15 Manufacturing,Wood machinist,Wood machinist cut and prepare timber for use in wood products.,£16,000,"£20,000"
16 Manufacturing,Chemical plant process operator,Chemical process operator, chemical plant worker, chemical plant operator,Chemical plant process operators control machinery that makes chemical products,
17 Manufacturing,Agricultural engineering technician,Engineer,Agricultural engineering technicians help to solve practical engineering problems in land-based industries.,£20,000,"£36,000"
18 Manufacturing,Car manufacturing worker,Car maker, assembly line worker, motor vehicle assembler,Car manufacturing workers build motor vehicles by assembling parts on a production line.,£11,500,"£22,0
19 Construction and trades,Window fitter,Window fitters install windows, conservatories and glazed curtain walls in homes and businesses.,£15,000,"£25,000"
20 Construction and trades,Water network operative,Distribution technician, leakage operative, network service technician,Water network operatives look after the pipes, mains and pumping stations that su
21 Managerial,Town planner,Spatial planner, planner, urban designer, planning officer,Town planners help shape the way towns and cities develop, and balance the demands on land with the needs of the comm
22 Managerial,Wedding planner,Planner,wedding planners help couples organise their wedding.,£17,000,"£25,000"
23 Construction and trades,Tiler,Wall tiler, floor tiler, ceramic tiler,Tilers tile walls and floors in kitchens, bathrooms, shops, hotels and restaurants.,£17,000,"£30,000"
24 Construction and trades,Thermal insulation engineer,Pipework lagger, thermal insulation installer,Thermal insulation engineers install insulating materials around pipes, boilers and ductwork in facto
25 Construction and trades,Thatcher,Thatchers use traditional craft skills, materials and tools to replace and repair thatched roofs.,£13,000,"£26,000"
26 Construction and trades,Technical surveyor,Surveying technician,Technical surveyors carry out tasks to support chartered surveyors, architects and engineers.,£18,000,"£32,000"
27 Construction and trades,Steel fixer,Steel fixers install and tie together the steel bars and mesh used to strengthen concrete on construction projects.,£14,000,"£35,000"
28 Construction and trades,Structural engineer,Engineer,Structural engineers help to design and build large structures and buildings, like hospitals, sports stadiums and bridges.,£22,000,"£70,000"
29 Construction and trades,Steeplejack,Lightning conductor engineer,Steeplejacks carry out repairs on buildings and structures to make them safe.,£15,000,"£26,000"
30 Construction and trades,Refrigeration and air-conditioning installer,Air-con engineer, HVAC engineer,Refrigeration and air-conditioning installers work on air quality and cooling systems in buildings a
31 Construction and trades,Road worker,Road construction operative, highways operative,Road workers build and repair roads and motorways.,£16,000,"£40,000"
32 Sports and leisure,Sports physiotherapist,Physiotherapist,Sports physiotherapists diagnose and treat sports injuries.,£23,000,"£45,000"
33 Sports and leisure,Sports professional,Professional,Sports professionals are skilled and talented sportsmen and sportswomen, who are paid to compete in their chosen sport.,Not specified,Not specified
34 Sports and leisure,Sports coach,Coach,Sports coaches teach sports skills to individuals and teams of all abilities.,£14,000,"£35,000"
35 Construction and trades,Plumber,Plumber,Plumbers fix and service hot and cold water systems, heating systems and drainage networks.,£15,000,"£40,000"
36 Construction and trades,Steel erector,Engineer,A steel erector assembles the metal framework of new buildings, and structures like bridges and tunnels.,£14,000,"£35,000"
37 Managerial,Rural surveyor,Agricultural surveyor,Rural surveyors value the assets of farms and estates, advise clients on legal and tax issues, and plan and develop land use.,£20,000,"£45,000"
38 Construction and trades,Shopfitter,Shopfitters build and install fixtures and fittings in offices, restaurants, shops and bars.,£13,500,"£30,000"
39 Construction and trades,Roofer,Slater, Flat roofer,Roofers re-slate and tile roofs, fix skylight windows and replace lead sheeting and cladding.,£13,000,"£32,000"
40 Construction and trades,Scaffolder,Engineer,Scaffolders put up and take down scaffolding on buildings that allow workers to work safely at height.,£14,000,"£30,000"
41 Construction and trades,Quantity surveyor,Engineer,Quantity surveyors oversee construction projects, managing risks and controlling costs.,£18,000,"£80,000"
42 Managerial,Quality assurance manager,Quality manager, quality control manager, quality inspector,Quality assurance managers make sure a company's products and services meet and maintain set standards,
43 Construction and trades,Pipe fitter,Pipefitter,Pipe fitters install industrial pipework, valves and pumps in factories, commercial premises and large buildings like power stations.,£20,000,"£40,000"
44 Construction and trades,Plasterer,Plasterers prepare walls and ceilings for decoration and finishing.,£14,000,"£30,000"
45 Construction and trades,Painter and decorator,Decorator,Painters and decorators prepare and apply paint, wallpaper and finishes to different surfaces.,£15,000,"£30,000"
```

Outline

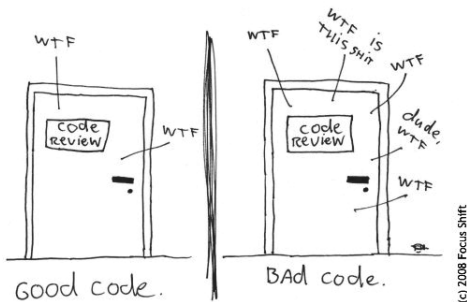
- 1 About me
- 2 Web scraping options
- 3 Example 1: LookFantastic site
- 4 Example 2: National Careers Service
 - Creating a Scrapy project
 - Get overarching job categories
 - Get individual job names & descriptions
 - Get individual job salaries
- 5 Limitations

Limitations

- Did not address dynamic content
- Fragility: structural changes on a site?
- Things I may not be aware of! Welcome to comment on:
<https://datapowered.io/post/2020-04-14-post-getting-stuck-in-with-scrapy/>
Another post to come on this topic.

Limitations

The ONLY valid measurement
OF code quality: WTFs/minute



Source: <https://www.osnews.com/story/19266/>