ai for all *Web Scraper* | Installation Guide

This web scraper was coded using the framework *Scrapy*. Scrapy allowed us to focus our time coding the actual data extraction part of the web scraper without having to worry about other things such as threads, processes, or synchronization. It allowed us to be efficient with our time.

In order to be able to run the web scraper, a few dependencies must be installed.

Requirements

- Python 3.5+

- Works on Linux, Windows, Mac OSX, BSD

Installation

The quick way:

```
  $ pip install scrapy
```

See the install section in the documentation at

https://docs.scrapy.org/en/latest/intro/install.html for more details.

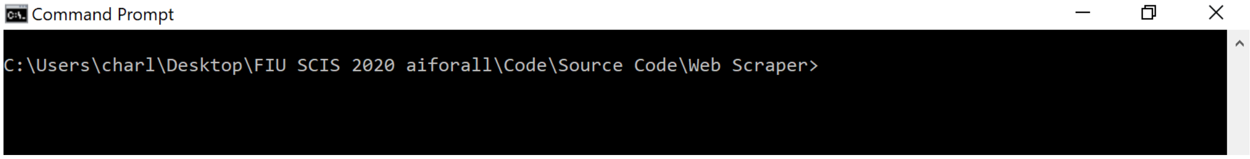
Documentation

Documentation is available online at https://docs.scrapy.org/

How to run the scraper

1. Open the terminal

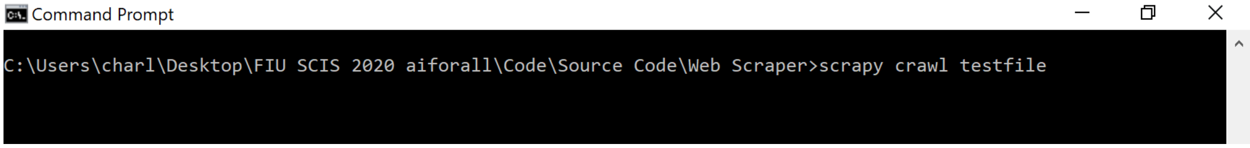
2. Navigate to the folder "Web Scraper" inside the "Source Code" directory



```
C:\Users\charl\Desktop\FIU SCIS 2020 aiforall\Code\Source Code\Web Scraper>
```

3. Run the scraper (the spider we are working with is called *testfile*):

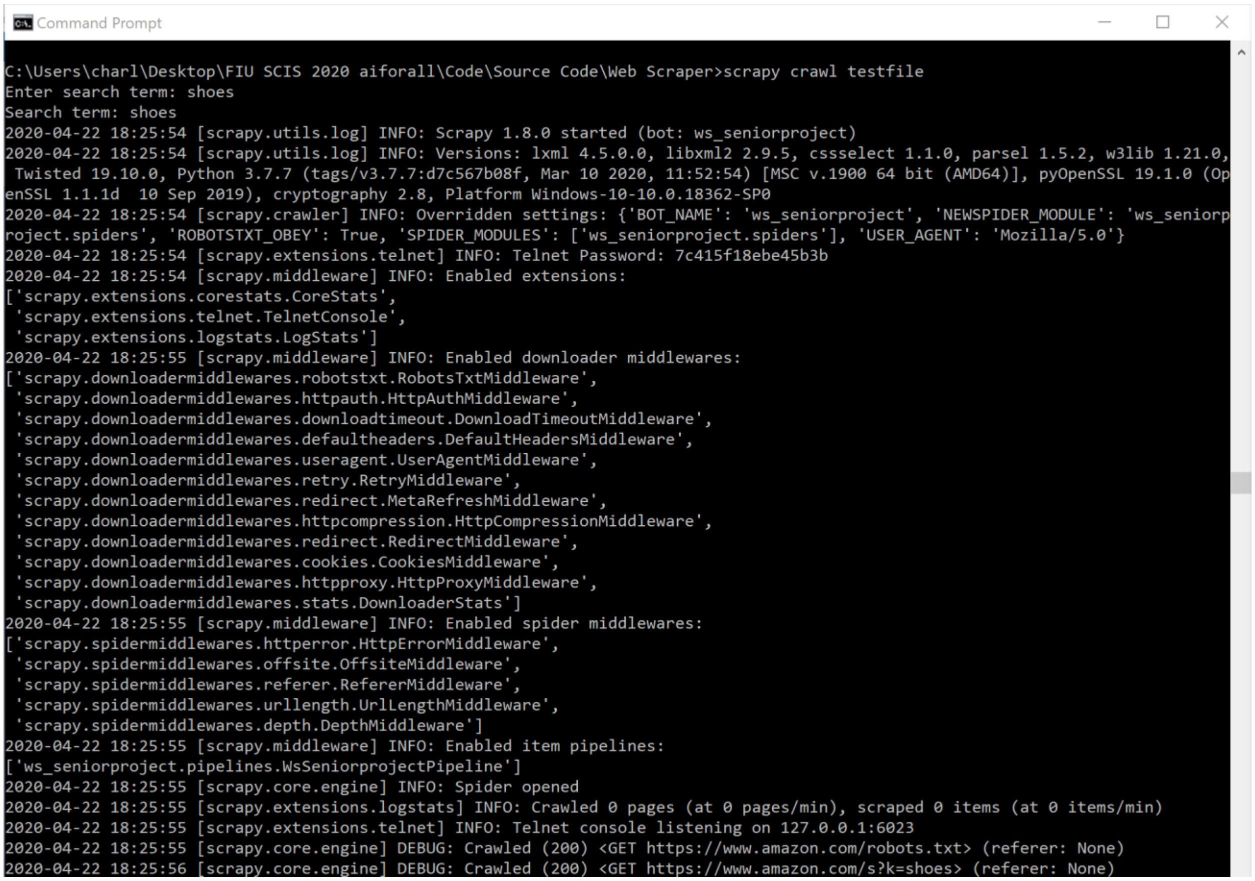
$ scrapy crawl testfile



```
C:\Users\charl\Desktop\FIU SCIS 2020 aiforall\Code\Source Code\Web Scraper>scrapy crawl testfile
```

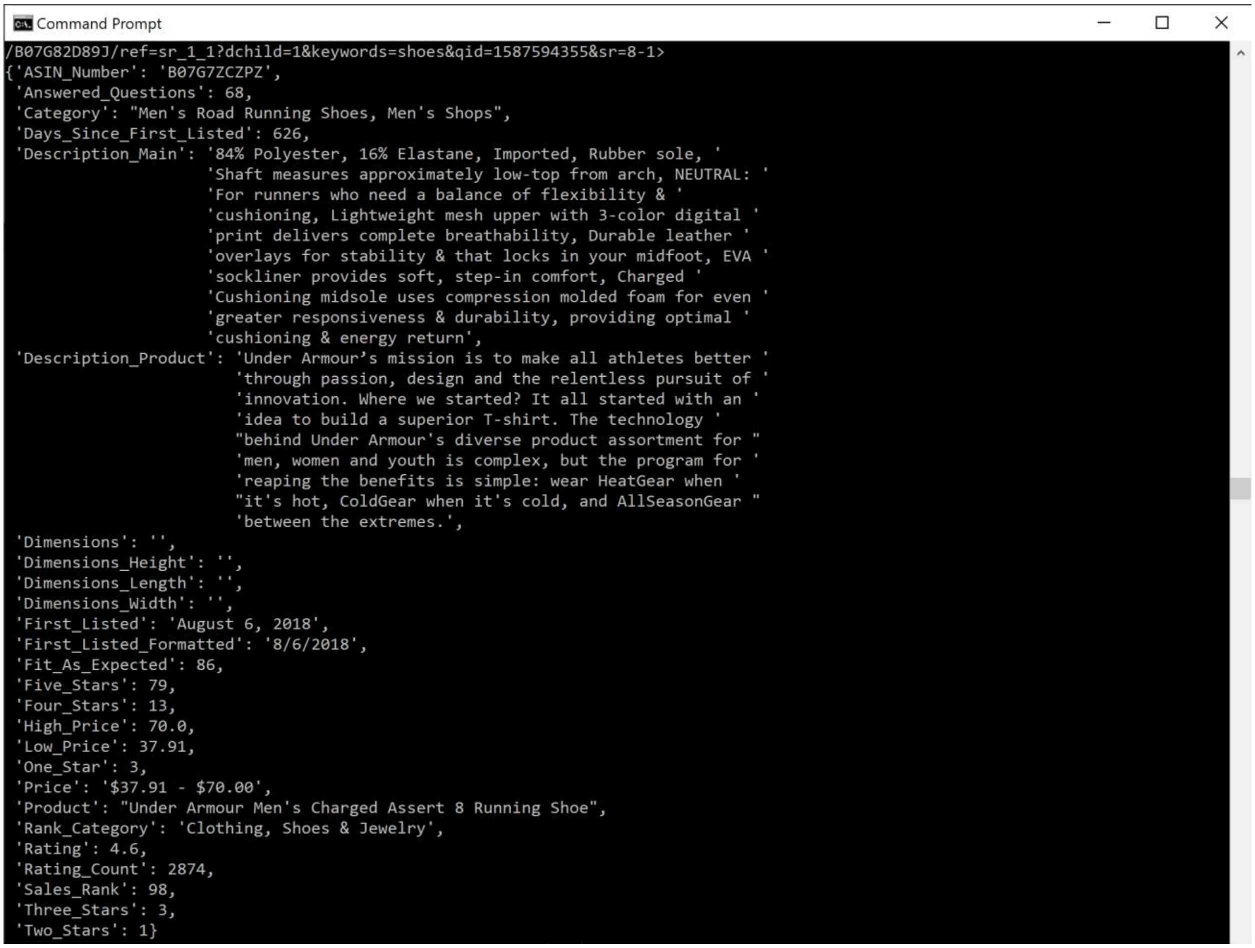
This will run the spider

4a. Input a product category such as “shoes”, “watches”, “pants”, “phone case”, etc.



```
C:\Users\charl\Desktop\FIU SCIS 2020 aiforall\Code\Source Code\Web Scraper>scrapy crawl testfile
Enter search term: shoes
Search term: shoes
2020-04-22 18:25:54 [scrapy.utils.log] INFO: Scrapy 1.8.0 started (bot: ws_seniorproject)
2020-04-22 18:25:54 [scrapy.utils.log] INFO: Versions: lxml 4.5.0.0, libxml2 2.9.5, cssselect 1.1.0, parsel 1.5.2, w3lib 1.21.0,
 Twisted 19.10.0, Python 3.7.7 (tags/v3.7.7:d7c567b08f, Mar 10 2020, 11:52:54) [MSC v.1900 64 bit (AMD64)], pyOpenSSL 19.1.0 (Op
enSSL 1.1.1d  10 Sep 2019), cryptography 2.8, Platform Windows-10-10.0.18362-SP0
2020-04-22 18:25:54 [scrapy.crawler] INFO: Overridden settings: {'BOT_NAME': 'ws_seniorproject', 'NEWSPIDER_MODULE': 'ws_seniorp
roject.spiders', 'ROBOTSTXT_OBEY': True, 'SPIDER_MODULES': ['ws_seniorproject.spiders'], 'USER_AGENT': 'Mozilla/5.0'}
2020-04-22 18:25:54 [scrapy.extensions.telnet] INFO: Telnet Password: 7c415f18ebe45b3b
2020-04-22 18:25:54 [scrapy.middleware] INFO: Enabled extensions:
['scrapy.extensions.corestats.CoreStats',
 'scrapy.extensions.telnet.TelnetConsole',
 'scrapy.extensions.logstats.LogStats']
2020-04-22 18:25:55 [scrapy.middleware] INFO: Enabled downloader middlewares:
['scrapy.downloadermiddlewares.robotstxt.RobotsTxtMiddleware',
 'scrapy.downloadermiddlewares.httpauth.HttpAuthMiddleware',
 'scrapy.downloadermiddlewares.downloadtimeout.DownloadTimeoutMiddleware',
 'scrapy.downloadermiddlewares.defaultheaders.DefaultHeadersMiddleware',
 'scrapy.downloadermiddlewares.useragent.UserAgentMiddleware',
 'scrapy.downloadermiddlewares.retry.RetryMiddleware',
 'scrapy.downloadermiddlewares.redirect.MetaRefreshMiddleware',
 'scrapy.downloadermiddlewares.httpcompression.HttpCompressionMiddleware',
 'scrapy.downloadermiddlewares.redirect.RedirectMiddleware',
 'scrapy.downloadermiddlewares.cookies.CookiesMiddleware',
 'scrapy.downloadermiddlewares.httpproxy.HttpProxyMiddleware',
 'scrapy.downloadermiddlewares.stats.DownloaderStats']
2020-04-22 18:25:55 [scrapy.middleware] INFO: Enabled spider middlewares:
['scrapy.spidermiddlewares.httperror.HttpErrorMiddleware',
 'scrapy.spidermiddlewares.offsite.OffsiteMiddleware',
 'scrapy.spidermiddlewares.referer.RefererMiddleware',
 'scrapy.spidermiddlewares.urllength.UrlLengthMiddleware',
 'scrapy.spidermiddlewares.depth.DepthMiddleware']
2020-04-22 18:25:55 [scrapy.middleware] INFO: Enabled item pipelines:
['ws_seniorproject.pipelines.WsSeniorprojectPipeline']
2020-04-22 18:25:55 [scrapy.core.engine] INFO: Spider opened
2020-04-22 18:25:55 [scrapy.extensions.logstats] INFO: Crawled 0 pages (at 0 pages/min), scraped 0 items (at 0 items/min)
2020-04-22 18:25:55 [scrapy.extensions.telnet] INFO: Telnet console listening on 127.0.0.1:6023
2020-04-22 18:25:55 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.amazon.com/robots.txt> (referer: None)
2020-04-22 18:25:56 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.amazon.com/s?k=shoes> (referer: None)
```

4b. Receive an output. The following is a sample output. If Amazon does not include some information on its page (such as dimensions in this sample), the particular attribute will be empty. The scraper can be modified to output any number of items ranging from 1 to as many as it finds (usually around 50-60).



```
/B07G82D89J/ref=sr_1_1?dchild=1&keywords=shoes&qid=1587594355&sr=8-1>
{'ASIN_Number': 'B07G7ZCZPZ',
 'Answered_Questions': 68,
 'Category': "Men's Road Running Shoes, Men's Shops",
 'Days_Since_First_Listed': 626,
 'Description_Main': '84% Polyester, 16% Elastane, Imported, Rubber sole, '
                     'Shaft measures approximately low-top from arch, NEUTRAL: '
                     'For runners who need a balance of flexibility & '
                     'cushioning, Lightweight mesh upper with 3-color digital '
                     'print delivers complete breathability, Durable leather '
                     'overlays for stability & that locks in your midfoot, EVA '
                     'sockliner provides soft, step-in comfort, Charged '
                     'Cushioning midsole uses compression molded foam for even '
                     'greater responsiveness & durability, providing optimal '
                     'cushioning & energy return',
 'Description_Product': 'Under Armour’s mission is to make all athletes better '
                        'through passion, design and the relentless pursuit of '
                        'innovation. Where we started? It all started with an '
                        'idea to build a superior T-shirt. The technology '
                        "behind Under Armour's diverse product assortment for "
                        'men, women and youth is complex, but the program for '
                        'reaping the benefits is simple: wear HeatGear when '
                        "it's hot, ColdGear when it's cold, and AllSeasonGear "
                        'between the extremes.',
 'Dimensions': '',
 'Dimensions_Height': '',
 'Dimensions_Length': '',
 'Dimensions_Width': '',
 'First_Listed': 'August 6, 2018',
 'First_Listed_Formatted': '8/6/2018',
 'Fit_As_Expected': 86,
 'Five_Stars': 79,
 'Four_Stars': 13,
 'High_Price': 70.0,
 'Low_Price': 37.91,
 'One_Star': 3,
 'Price': '$37.91 - $70.00',
 'Product': "Under Armour Men's Charged Assert 8 Running Shoe",
 'Rank_Category': 'Clothing, Shoes & Jewelry',
 'Rating': 4.6,
 'Rating_Count': 2874,
 'Sales_Rank': 98,
 'Three_Stars': 3,
 'Two_Stars': 1}
```

Alternatively, store the results in a json file by running this command:

```
$ scrapy crawl testfile -o results.json
```

Or in a .csv file:

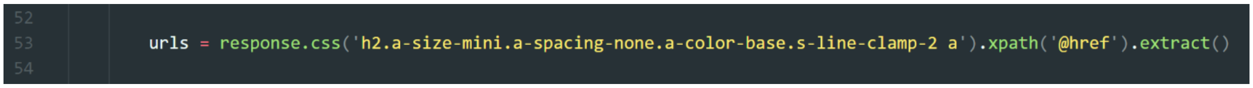
```
$ scrapy crawl testfile -o results.csv
```

How to modify number of outputs

1. Navigate to **testfile.py** following this path:

*Source Code -> Web Scraper -> ws_senioproject -> spiders -> **tesfile.py***

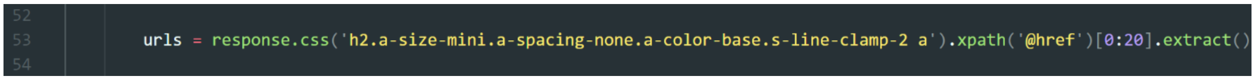
2. In the testfile.py file, navigate to line 53 where you'll find this code:



```
52
53          urls = response.css('h2.a-size-mini.a-spacing-none.a-color-base.s-line-clamp-2 a').xpath('@href').extract()
54
```

3. Add brackets after ('@href') such that it looks like this: ('@href')[0:9].extract()

This means that the web scraper will return 9 items. Change it to any number such as:



```
52
53          urls = response.css('h2.a-size-mini.a-spacing-none.a-color-base.s-line-clamp-2 a').xpath('@href')[0:20].extract()
54
```

This will cause it to return 20 items. Remove the brackets and it will return as many items as it finds (usually 50-60 items)