Introduction to Web Scraping using Scrapy

What is web scraping?

Examples of web scraping

gathering…

≥ video game prices
≥ weather data for the week
≥ a list of conifers (pine trees)

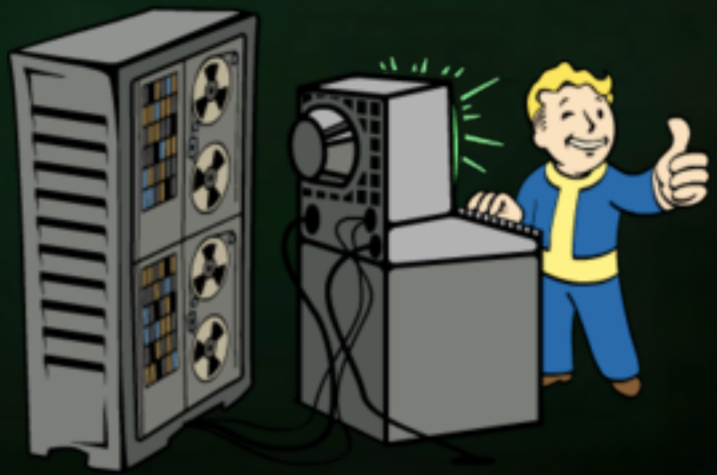
Installing Scrapy

Requirements!

≥ Python 2.7
≥ pip
≥ lxml
≥ OpenSSL

Installing Scrapy

```
≥ pip install scrapy
```



What is scrapy?

Scrapy commands

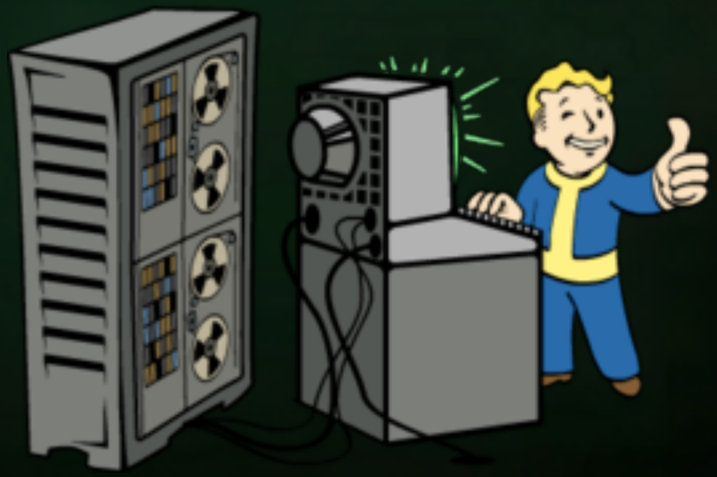
```
> scrapy <command> -h
```

Global commands:

```
> startproject
> settings
> runspider
> shell
> fetch
> view
> version
```

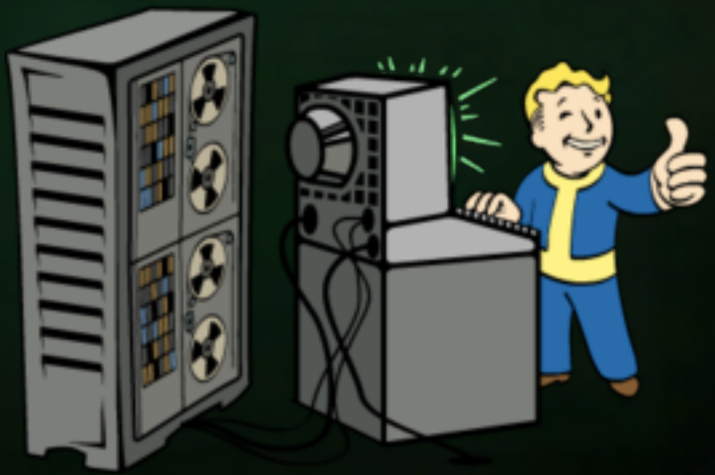
Project-only commands

```
> crawl
> check
> list
> edit
> parse
> genspider
> bench
```



Structure of Scrapy

```
tutorial/
    scrapy.cfg
    testing/
        __init__.py
        items.py
        pipelines.py
        settings.py
        spiders/
            __init__.py
```



Building a Scrapy bot to extract conifer plants

Creating a new project

```
≥ scrapy start project
```

http://www.greatplantpicks.org/plantlists/by_plant_type/conifer

Defining field items in items.py

>_

```
import scrapy

class ConifersItem(scrapy.Item):
    name = scrapy.Field()
    genus = scrapy.Field()
    species = scrapy.Field()
    pass
```

Building the bot

```
>_

import scrapy
from conifers.items import ConifersItem

class ConifersSpider(scrapy.Spider):
    name = "conifers"
    allowed_domains = ["greatplantpicks.org"]
    start_urls = [
    "http://www.greatplantpicks.org/plantlists/by_plant_type/conifer"]

    def parse(self, response):
        filename = response.url.split("/")[-2] + '.html'
        with open(filename, 'wb') as f:
            f.write(response.body)
```

Building the bot

```
> scrapy crawl conifers
```

Extracting HTML elements using XPath and CSS selectors

```
def parse(self, response):
  for sel in response.xpath('//tbody/tr'):
    item = ConifersItem()
    item['name'] = sel.xpath('td[@class="common-name"]/a/ text()').extract()
    item['genus'] = sel.xpath('td[@class="plantname"]/a/span[@class="genus"]/text()').extract()
    item['species'] = sel.xpath('td[@class="plantname"]/a/span[@class="species"]/text()').extract()
    yield item
```

Running the bot we built and exporting the data as a csv and JSON file

```
≥ scrapy crawl conifers -o trees_json.json

≥ scrapy crawl conifers -o trees_csv.csv
```

Scrape Away!

Images from Fallout 4 wikia