WEB CRAWLING AND SCRAPING USING: SCRAPY

WEB CRAWLING :::: [everything is done from CMD/TERMINAL]

1. Installing Scrapy :
   1. pip install: To install using pip, open the terminal and run the following command:

**pip install scrapy**

1. Creating a Scrapy Project
   1. Scrapy is a framework which has a standardized structure.
   2. To create a new project in scrapy use command:
      1. scrapy startproject <name of project>

eg : scrapy startproject webscrapy

Running this command will create a project directory with the below contents:

webscrapy

2├── scrapy.cfg -- deploy configuration file of scrapy project

3└── webscrapy -- your scrapy project module.

4 ├── \_\_init\_\_.py -- module initializer(empty file)

5 ├── items.py -- project item definition py file

6 ├── middlewares.py -- project middleware py file

7 ├── pipelines.py -- project pipeline py file

8 ├── settings.py -- project settings py file

9 └── spiders -- directory where spiders are kept

10 ├── \_\_init\_\_.py

1. Create a Spider
   1. Use the command “genspider” – takes the name of the spider and the URL it will crawl:
      1. cd webscrapy (webscrapy is the name of the project)
      2. scrapy genspider <name of spider> <URL>

eg: scrapy genspider quotes <http://quotes.toscrape.com>

After running this command, Scrapy will automatically create a Python file named **quotes** in the spider folder. When you open that spider quotes.py, you will see a class named quotesSpider that inherits scrapy.Spider class and contains a method named parse.

\*\*Sample .py file\*\*\*\*

import scrapy

2

3class ImdbSpider(scrapy.Spider):

4 name = 'imdb'

5 allowed\_domains = ['www.imdb.com']

6 start\_urls = ['http://www.imdb.com/']

7

8 def parse(self, response):

9 pass

* name: The name of the spider. In this case, it is ImdbSpider. Naming spiders properly is essential when you have to maintain hundreds of spiders.
* allowed\_domains: An optional list of strings containing domains that this spider is allowed to crawl. Requests for URLs not belonging to the domain names specified in this list won’t be followed.
* parse(self, response): This function is called whenever the crawler successfully crawls a URL.

1. To run the spider, use the below command,
   1. Scrapy crawl quotes [ make sure to be in the correct directory]

BASIC WEB SCRAPING   
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

* + - 1. Open the quotes.py file which was created.
      2. Copy the code in github and paste in the .py file and save it

Note: The output is in dictionary format as per the code, saving the output (scraped data) in the csv format [Step 2 has the code for this as well]

* + - 1. After saving the .py file, run the below command:
         1. scrapy crawl quotes -o quotes.csv

This creates a quotes.csv file within the spiders directory with the scraped data.