```
scrapy startproject Scrapy
```

- cd Scrapy

```
scrapy genspider quotes quotes.com
```

```
-  quotes is the name of spider, hence inside the class we give quotes as the name of spider
-  Refer quotes_spider.py file
 ![[Pasted image 20210906062247.png]]
```

- To run spider,

```
scrapy crawl quotes
```

- To extract data with scrapy, scrapy is trying selectorsusing the scrapy shell. Run:

```
scrapy shell http://quotes.toscrape.com/page/1/
```

- quotes_spider.py
    - start_requests contain the urls to be scraped on
    - parse method in quotes_spider.py file is used to handle the responses which is downloaded for each request that is made

```
(base) deepcompute@deepcompute-ThinkPad-E470:~$ scrapy startproject Scrapy
New Scrapy project 'Scrapy', using template directory '/home/deepcompute/anaconda3/lib/python3.7/site-packages/scrapy/templat
es/project', created in:
    /home/deepcompute/Scrapy

You can start your first spider with:
    cd Scrapy
    scrapy genspider example example.com
```

```
(base) deepcompute@deepcompute-ThinkPad-E470:~$ cd Scrapy
(base) deepcompute@deepcompute-ThinkPad-E470:~/Scrapy$ scrapy genspider quotes quotes.com
Created spider 'quotes' using template 'basic' in module:
  Scrapy.spiders.quotes
(base) deepcompute@deepcompute-ThinkPad-E470:~/Scrapy$ scrapy crawl quotes
2021-09-06 06:27:21 [scrapy.utils.log] INFO: Scrapy 2.4.1 started (bot: Scrapy)
2021-09-06 06:27:21 [scrapy.utils.log] INFO: Versions: lxml 4.5.0.0, libxml2 2.9.9, cssselect 1.1.0, parsel 1.5.2, w3lib 1.21
.0, Twisted 21.2.0, Python 3.7.6 (default, Jan  8 2020, 19:59:22) - [GCC 7.3.0], pyOpenSSL 19.1.0 (OpenSSL 1.1.1d  10 Sep 201
9), cryptography 2.8, Platform Linux-5.11.0-27-generic-x86_64-with-debian-bullseye-sid
2021-09-06 06:27:21 [scrapy.utils.log] DEBUG: Using reactor: twisted.internet.epollreactor.EPollReactor
2021-09-06 06:27:21 [scrapy.crawler] INFO: Overridden settings:
{'BOT_NAME': 'Scrapy',
 'NEWSPIDER_MODULE': 'Scrapy.spiders',
 'ROBOTSTXT_OBEY': True,
 'SPIDER_MODULES': ['Scrapy.spiders']}
2021-09-06 06:27:21 [scrapy.extensions.telnet] INFO: Telnet Password: 967439feaffdecad
2021-09-06 06:27:21 [scrapy.middleware] INFO: Enabled extensions:
['scrapy.extensions.corestats.CoreStats',
 'scrapy.extensions.telnet.TelnetConsole',
 'scrapy.extensions.memusage.MemoryUsage',
 'scrapy.extensions.logstats.LogStats']
2021-09-06 06:27:22 [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']
2021-09-06 06:27:22 [scrapy.middleware] INFO: Enabled spider middlewares:
['scrapy.spidermiddlewares.httperror.HttpErrorMiddleware',
```

```
(base) deepcompute@deepcompute-ThinkPad-E470:~/Scrapy$ scrapy shell https://quotes.toscrape.com/page/1/
2021-09-06 06:29:11 [scrapy.utils.log] INFO: Scrapy 2.4.1 started (bot: Scrapy)
2021-09-06 06:29:11 [scrapy.utils.log] INFO: Versions: lxml 4.5.0.0, libxml2 2.9.9, cssselect 1.1.0, parsel 1.5.2, w3lib 1.21
.0, Twisted 21.2.0, Python 3.7.6 (default, Jan  8 2020, 19:59:22) - [GCC 7.3.0], pyOpenSSL 19.1.0 (OpenSSL 1.1.1d  10 Sep 201
9), cryptography 2.8, Platform Linux-5.11.0-27-generic-x86_64-with-debian-bullseye-sid
2021-09-06 06:29:11 [scrapy.utils.log] DEBUG: Using reactor: twisted.internet.epollreactor.EPollReactor
2021-09-06 06:29:11 [scrapy.crawler] INFO: Overridden settings:
{'BOT_NAME': 'Scrapy',
 'DUPEFILTER_CLASS': 'scrapy.dupefilters.BaseDupeFilter',
 'LOGSTATS_INTERVAL': 0,
 'NEWSPIDER_MODULE': 'Scrapy.spiders',
 'ROBOTSTXT_OBEY': True,
 'SPIDER_MODULES': ['Scrapy.spiders']}
2021-09-06 06:29:11 [scrapy.extensions.telnet] INFO: Telnet Password: afc72dd68fae7049
2021-09-06 06:29:11 [scrapy.middleware] INFO: Enabled extensions:
['scrapy.extensions.corestats.CoreStats',
 'scrapy.extensions.telnet.TelnetConsole',
 'scrapy.extensions.memusage.MemoryUsage']
2021-09-06 06:29:11 [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']
2021-09-06 06:29:12 [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']
```

```
In [1]: response.css('title')
Out[1]: [<Selector xpath='descendant-or-self::title' data='<title>Quotes to Scrape</title>'>]

In [2]: response.css('title::text').getall()
Out[2]: ['Quotes to Scrape']

In [3]: response.xpath('//title')
Out[3]: [<Selector xpath='//title' data='<title>Quotes to Scrape</title>'>]

In [4]: response.xpath('//title/text()').get()
Out[4]: 'Quotes to Scrape'

In [5]: response.css("div.quote")
Out[5]:
[<Selector xpath="descendant-or-self::div[@class and contains(concat(' ', normalize-space(@class), ' '), ' quote ')]" data='<
div class="quote" itemscope itemtype...'>,
 <Selector xpath="descendant-or-self::div[@class and contains(concat(' ', normalize-space(@class), ' '), ' quote ')]" data='<
div class="quote" itemscope itemtype...'>,
 <Selector xpath="descendant-or-self::div[@class and contains(concat(' ', normalize-space(@class), ' '), ' quote ')]" data='<
div class="quote" itemscope itemtype...'>,
 <Selector xpath="descendant-or-self::div[@class and contains(concat(' ', normalize-space(@class), ' '), ' quote ')]" data='<
div class="quote" itemscope itemtype...'>,
 <Selector xpath="descendant-or-self::div[@class and contains(concat(' ', normalize-space(@class), ' '), ' quote ')]" data='<
div class="quote" itemscope itemtype...'>,
 <Selector xpath="descendant-or-self::div[@class and contains(concat(' ', normalize-space(@class), ' '), ' quote ')]" data='<
div class="quote" itemscope itemtype...'>,
 <Selector xpath="descendant-or-self::div[@class and contains(concat(' ', normalize-space(@class), ' '), ' quote ')]" data='<
div class="quote" itemscope itemtype...'>,
 <Selector xpath="descendant-or-self::div[@class and contains(concat(' ', normalize-space(@class), ' '), ' quote ')]" data='<
div class="quote" itemscope itemtype...'>,
 <Selector xpath="descendant-or-self::div[@class and contains(concat(' ', normalize-space(@class), ' '), ' quote ')]" data='<
div class="quote" itemscope itemtype...'>,
 <Selector xpath="descendant-or-self::div[@class and contains(concat(' ', normalize-space(@class), ' '), ' quote ')]" data='<
div class="quote" itemscope itemtype...'>]
```

```
In [6]: response.css("div.quote")[0]
Out[6]: <Selector xpath="descendant-or-self::div[@class and contains(concat(' ', normalize-space(@class), ' '), ' quote ')]"
data='<div class="quote" itemscope itemtype...'>

In [7]: q1 = response.css("div.quote")[0]

In [8]: text = q1.css("span.text::text").get()

In [9]: text
Out[9]: '“The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking.”'

In [10]: a1 = q1.css("small.author::text").get()

In [11]: a1
Out[11]: 'Albert Einstein'

In [12]:
```