Libraries

In [1]:
```
import pandas as pd
import requests
from scrapy.http import TextResponse
```

No of pages to scrape

In [22]:
```
pages=int(input('How Many Pages Do You Want to Scrape: '))
```

Scraper code

In [34]:
```
dictionary = {'One':'1', 'Two':'2', 'Three':'3', 'Four':'4', 'Five':'5'}
data={'Title':[],'Price':[],'Stock':[],'Star':[]}
for i in range(pages):
    url = 'http://books.toscrape.com/catalogue/page-'+str(i+1)+'.html'
    #url = 'http://books.toscrape.com/catalogue/page-1.html'
    res = requests.get(url)
    response = TextResponse(res.url, body=res.text, encoding='utf-8')
    print("Scaning page -> {0}".format(i+1))
    books=response.css('ol.row')
    for book in books:
        for b in book.css('article.product_pod'):
            #print(b.css('a::attr(title)').extract_first())
            data['Title'].append(b.css('a::attr(title)').extract_first
())
            data['Price'].append(b.css('div.product_price p.price_colo
r::text').extract_first().split('Â')[1])
            data['Stock'].append(b.css('div.product_price p.instock.avai
lability::text').getall()[1].strip())
            data['Star'].append(''.join([v for k,v in dictionary.items()
if k in b.css('p::attr(class)').getall()[0].split()[-1]]))
```

```
Scaning page -> 1
```

In [ ]:
```
## Convert the dictionary to dataframe
```

In [39]:
```
book_df=pd.DataFrame(data)
book_df.head(5)
```

Out[39]:

	Price	Star	Stock	Title
0	£51.77	3	In stock	A Light in the Attic
1	£53.74	1	In stock	Tipping the Velvet
2	£50.10	1	In stock	Soumission
3	£47.82	4	In stock	Sharp Objects
4	£54.23	5	In stock	Sapiens: A Brief History of Humankind

Save dataframe to notebook

In [36]:
```
from platform_sdk.models import Dataset
from platform_sdk.dataset_writer import DatasetWriter
dataset = Dataset(get_platform_sdk_client_context()).get_by_id(dataset_id="602a6dce2dbf29194906f069")
dataset_writer = DatasetWriter(get_platform_sdk_client_context(), dataset)
write_tracker = dataset_writer.write(book_df, file_format='json')
```

```
INFO:azure.datalake.store.core:closing stream
INFO:azure.datalake.store.transfer:Transferred tempFile.parquet -> /foundation/data/stage/users/OrgID@AdobeID/01EYJVWBY327CWMXHPKSVHMN62/602a6dce2dbf29194906f06a/1613393442081.json
INFO:PlatformSDKPython:dataset_writer: 20 rows written. 101.77 MB memory used for this process
```

Read the saved data

In [40]:
```
from platform_sdk.dataset_reader import DatasetReader
from datetime import date
dataset_reader = DatasetReader(get_platform_sdk_client_context(), dataset_id="602a6dce2dbf29194906f069")
df0 = dataset_reader.limit(100).read()
df0.head(5)
```

```
INFO:PlatformSDKPython:dataset_reader: seconds taken to get dataset details from catalog and make PQS connection: 0.28
INFO:PlatformSDKPython:dataset_id: 602a6dce2dbf29194906f069, limit: 100
INFO:PlatformSDKPython:dataset_reader: seconds taken to execute query: 15.56
INFO:PlatformSDKPython:dataset_reader: 21 rows read. 103.18 MB memory used for this process
INFO:PlatformSDKPython:dataset_reader: seconds taken to format data of dataframe: 0.01
```

Out[40]:

	Price	Star	Stock	Title
0	£51.77	3	In stock	A Light in the Attic
1	£53.74	1	In stock	Tipping the Velvet
2	£50.10	1	In stock	Soumission
3	£47.82	4	In stock	Sharp Objects
4	£54.23	5	In stock	Sapiens: A Brief History of Humankind

In [ ]: