In [1]:
```
import requests
from bs4 import BeautifulSoup as bs
import pandas as pd
import re
import numpy as np
```

In [2]:
```
web=requests.get('https://books.toscrape.com/catalogue/page-1.html')
web.status_code
```

Out[2]:
```
200
```

In [3]:
```
doc=bs(web.text,'html.parser')
doc.head()
```

Out[3]:
```
[<title>
     All products | Books to Scrape - Sandbox
 </title>,
 <meta content="text/html; charset=utf-8" http-equiv="content-type"/>,
 <meta content="24th Jun 2016 09:30" name="created"/>,
 <meta content="" name="description"/>,
 <meta content="width=device-width" name="viewport"/>,
 <meta content="NOARCHIVE,NOCACHE" name="robots"/>,
 <link href="../static/oscar/favicon.ico" rel="shortcut icon"/>,
 <link href="../static/oscar/css/styles.css" rel="stylesheet" type="text/css"/>,
 <link href="../static/oscar/js/bootstrap-datetimepicker/bootstrap-datetimepicker.css" rel="stylesheet"/>,
 <link href="../static/oscar/css/datetimepicker.css" rel="stylesheet" type="text/css"/>]
```

In [4]:
```
book_links=[]
for i in doc.find_all('h3'):
  if i.find('a')!=None:
      book_links.append('https://books.toscrape.com/catalogue/'+i.find('a')['href'])
book_links
```

Out[4]:
```
['https://books.toscrape.com/catalogue/a-light-in-the-attic_1000/index.html',
 'https://books.toscrape.com/catalogue/tipping-the-velvet_999/index.html',
 'https://books.toscrape.com/catalogue/soumission_998/index.html',
 'https://books.toscrape.com/catalogue/sharp-objects_997/index.html',
 'https://books.toscrape.com/catalogue/sapiens-a-brief-history-of-humankind_996/index.html',
 'https://books.toscrape.com/catalogue/the-requiem-red_995/index.html',
 'https://books.toscrape.com/catalogue/the-dirty-little-secrets-of-getting-your-dream-job_994/index.html',
 'https://books.toscrape.com/catalogue/the-coming-woman-a-novel-based-on-the-life-of-the-infamous-feminist-victoria-woodhull_993/index.html',
 'https://books.toscrape.com/catalogue/the-boys-in-the-boat-nine-americans-and-their-epic-quest-for-gold-at-the-1936-berlin-olympics_992/index.htm
l',
 'https://books.toscrape.com/catalogue/the-black-maria_991/index.html',
 'https://books.toscrape.com/catalogue/starving-hearts-triangular-trade-trilogy-1_990/index.html',
 'https://books.toscrape.com/catalogue/shakespeares-sonnets_989/index.html',
 'https://books.toscrape.com/catalogue/set-me-free_988/index.html',
 'https://books.toscrape.com/catalogue/scott-pilgrims-precious-little-life-scott-pilgrim-1_987/index.html',
 'https://books.toscrape.com/catalogue/rip-it-up-and-start-again_986/index.html',
 'https://books.toscrape.com/catalogue/our-band-could-be-your-life-scenes-from-the-american-indie-underground-1981-1991_985/index.html',
 'https://books.toscrape.com/catalogue/olio_984/index.html',
 'https://books.toscrape.com/catalogue/mesaerion-the-best-science-fiction-stories-1800-1849_983/index.html',
 'https://books.toscrape.com/catalogue/libertarianism-for-beginners_982/index.html',
 'https://books.toscrape.com/catalogue/its-only-the-himalayas_981/index.html']
```

In [5]:
```
title=[]
a=requests.get(book_links[0])
ahtml=bs(a.text,'html.parser')
for i in ahtml.find_all('h1'):
    title.append(i.text)
title
```

Out[5]:
```
['A Light in the Attic']
```

In [6]:
```
columns=[]
for j in ahtml.find_all('th'):
  columns.append(j.text)
columns
```

Out[6]:
```
['UPC',
 'Product Type',
 'Price (excl. tax)',
 'Price (incl. tax)',
 'Tax',
 'Availability',
 'Number of reviews']
```

In [7]:
```
values=[]
for k in ahtml.find_all('td'):
  if k.text=="":
    values.append(np.nan)
  elif 'Â' in k.text:
    values.append(k.text.replace('Â',""))
  else:
    values.append(k.text)
values
```

Out[7]:
```
['a897fe39b1053632',
 'Books',
 '£51.77',
 '£51.77',
 '£0.00',
 'In stock (22 available)',
 '0']
```

In [8]:
```
dict1={}
dict1[title[0]]=values
dict1
```

Out[8]:
```
{'A Light in the Attic': ['a897fe39b1053632',
  'Books',
  '£51.77',
  '£51.77',
  '£0.00',
  'In stock (22 available)',
  '0']}
```

In [9]:
```
df=pd.DataFrame(dict1.values(),index=dict1.keys(),columns=columns)
df
```

Out[9]:

	UPC	Product Type	Price (excl. tax)	Price (incl. tax)	Tax	Availability	Number of reviews
A Light in the Attic	a897fe39b1053632	Books	£51.77	£51.77	£0.00	In stock (22 available)	0

In [10]:
```
def scrap(number_of_page):
  dict1={}
  for p in range(1,int(number_of_page)+1):
    web=requests.get('https://books.toscrape.com/catalogue/page-'+str(p)+'.html')
    doc=bs(web.text,'html.parser')
    b_links=[]
    for i in doc.find_all('h3'):
      if i.find('a')!=None:
          b_links.append('https://books.toscrape.com/catalogue/'+i.find('a')['href'])
    for t in b_links:
      a=requests.get(t)
      ahtml=bs(a.text,'html.parser')
      for i in ahtml.find_all('h1'):
        columns=[]
        for j in ahtml.find_all('th'):
          columns.append(j.text)
        values=[]
        for k in ahtml.find_all('td'):
          if k.text=="":
            values.append(np.nan)
          elif 'Â' in k.text:
            values.append(k.text.replace('Â',""))
          else:
            values.append(k.text)

        dict1[i.text]=values
  return pd.DataFrame(dict1.values(),index=dict1.keys(),columns=columns).reset_index().rename(columns={'index':'Book name'})
```

In [ ]:
```
books_info=scrap(50)
```

In [ ]:
```
books_info
```

In [ ]:
```
books_info.to_csv('./books_info.csv')
```

In [ ]:
```
books_info.info()
```