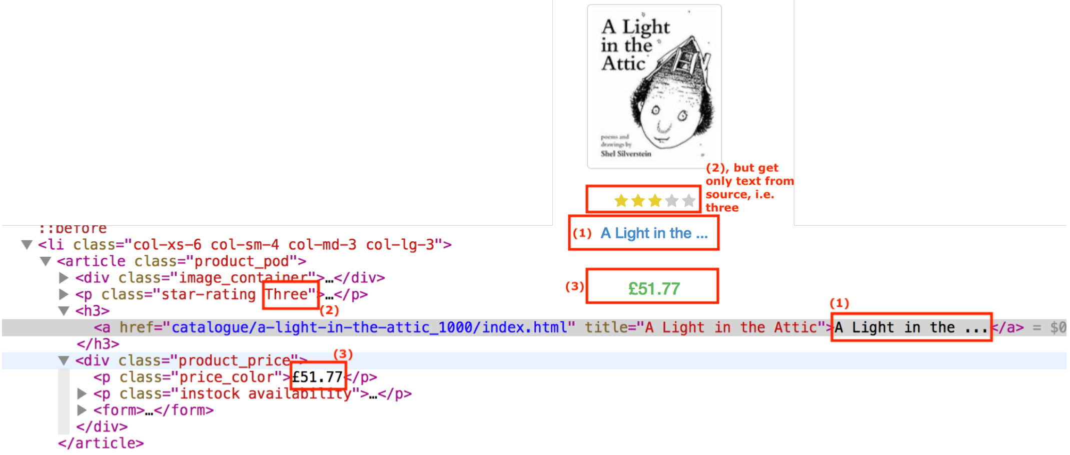
Assignment 2: Web Scraping

Q1. Scrape Book Catalog

- Scape content of http://books.toscrape.com (http://books.toscrape.com)
- Write a function getData() to scrape **title** (see (1) in Figure), **rating** (see (2) in Figure), **price** (see (3) in Figure) of all books (i.e. 20 books) listed in the page.
  - For example, the figure shows one book and the corresponding html code. You need to scrape the highlighted content.
  - For star ratings, you can simply scrape One, Two, Three, ...
- The output is a list of 20 tuples, e.g. [('A Light in the ...','Three','£51.77'), ...]



A Light in the Attic
poems and drawings by Shel Silverstein
(2), but get only text from source, i.e. three
★★★☆☆
(1) A Light in the ...
(3) £51.77
(1)
(2)
(3)

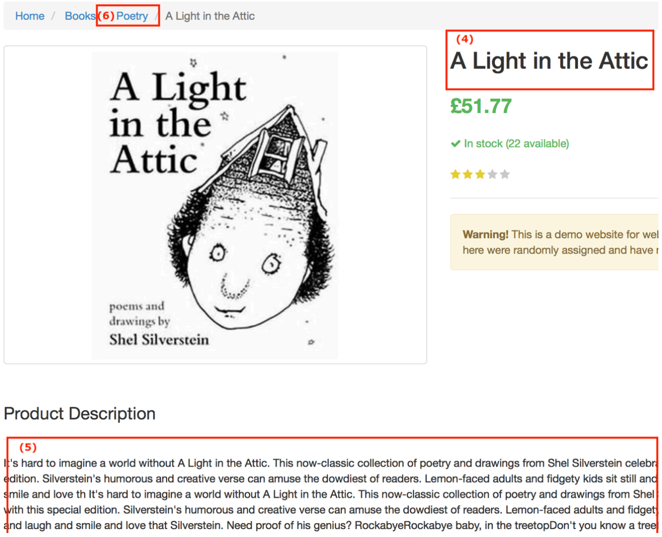
```
::before
▼ <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
  ▼ <article class="product_pod">
    ▶ <div class="image_container">…</div>
    ▶ <p class="star-rating Three">…</p>
    ▼ <h3>
        <a href="catalogue/a-light-in-the-attic_1000/index.html" title="A Light in the Attic">A Light in the ...</a> == $0
      </h3>
    ▼ <div class="product_price">
        <p class="price_color">£51.77</p>
      ▶ <p class="instock availability">…</p>
      ▶ <form>…</form>
      </div>
    </article>
```

Q2. Data Analysis

- Create a function preprocess_data which
  - takes the list of tuples from Q1 as an input
  - converts the price strings to numbers
  - calculates the average price of books by ratings
  - plots a bar chart to show the average price by ratings.

Q3 (Bonus) Expand your solution to Q1 to scrape the full details of all books on http://books.toscrape.com (http://books.toscrape.com)

- Write a function getFullData() to do the following:
  - Besides scraping title, rating, and price of each book as stated in Q1, also scrape the **full title** (see (4) in Figure), **description** (see (5) in Figure), and **category** (see (6) in Figure) in each individual book page.
    - An example individual book page is shown in the figure below.



Home / Books (6) Poetry / A Light in the Attic

(4)
A Light in the Attic

£51.77

✓ In stock (22 available)

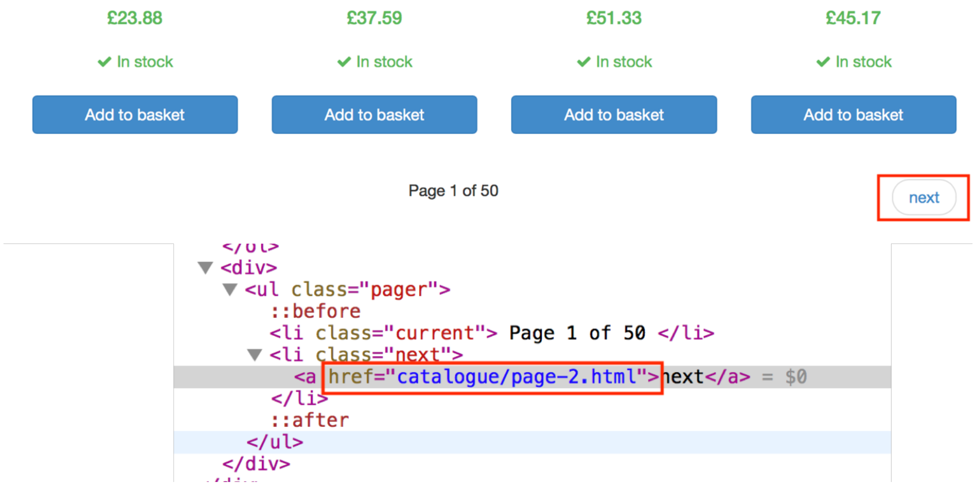
★★★☆☆

Warning! This is a demo website for wel
here were randomly assigned and have r

Product Description

(5)
It's hard to imagine a world without A Light in the Attic. This now-classic collection of poetry and drawings from Shel Silverstein celebr
edition. Silverstein's humorous and creative verse can amuse the dowdiest of readers. Lemon-faced adults and fidgety kids sit still and
smile and love th It's hard to imagine a world without A Light in the Attic. This now-classic collection of poetry and drawings from Shel
with this special edition. Silverstein's humorous and creative verse can amuse the dowdiest of readers. Lemon-faced adults and fidget
and laugh and smile and love that Silverstein. Need proof of his genius? RockabyeRockabye baby, in the treetopDon't you know a tree

  - Scape all book listing pages following the "next" link at the bottom. The figure below gives an screenshot of the "next" link and its corresponding html code.
  - **Do not hardcode page URLs** (except http://books.toscrape.com (http://books.toscrape.com)) in your code.



Page 1 of 50 next

```
</ul>
▼ <div>
  ▼ <ul class="pager">
      ::before
      <li class="current"> Page 1 of 50 </li>
    ▼ <li class="next">
        <a href="catalogue/page-2.html">next</a> == $0
      </li>
      ::after
    </ul>
  </div>
```

  - The output is a list containing 1000 tuples,
    - e.g. [('A Light in the ...','Three','£51.77', 'A Light in the Attic', "It's hard to imagine a world without A Light in the Attic. This now-classic collection ...",'Poetry'), ...]

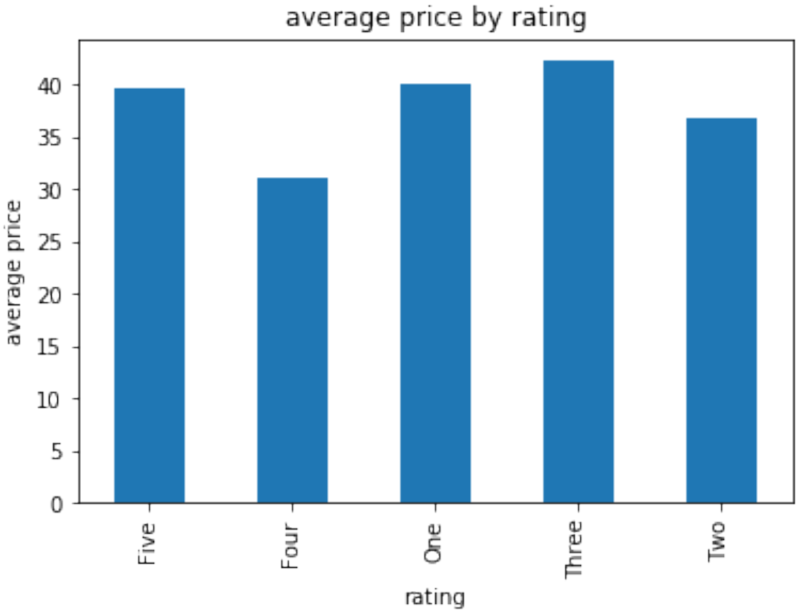
In [1]:

```
import requests
from bs4 import BeautifulSoup
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np


def preprocess_data(data):

    # add your code


def getData():


    data=[]  # variable to hold all reviews

    # add your code

    return data

def getFullData():


    data=[]  # variable to hold all book data

    page_url="http://books.toscrape.com"


    # add your code

    return data
```

In [2]:

```
if __name__ == "__main__":

    # Test Q1
    data=getData()

    # Test Q2
    preprocess_data(data)

    # Test Q3
    data=getFullData()
    print(len(data))

    # randomly select one book
    print(data[899])
```

```
20
                                       title rating  price
0                         A Light in the ...  Three  51.77
1                         Tipping the Velvet    One  53.74
2                                 Soumission    One  50.10
3                              Sharp Objects   Four  47.82
4              Sapiens: A Brief History ...   Five  54.23
5                            The Requiem Red    One  22.65
6              The Dirty Little Secrets ...   Four  33.34
7                   The Coming Woman: A ...  Three  17.93
8                      The Boys in the ...   Four  22.60
9                            The Black Maria    One  52.15
10  Starving Hearts (Triangular Trade ...     Two  13.99
11                     Shakespeare's Sonnets   Four  20.66
12                               Set Me Free   Five  17.46
13     Scott Pilgrim's Precious Little ...   Five  52.29
14                         Rip it Up and ...   Five  35.02
15                     Our Band Could Be ...  Three  57.25
16                                      Olio    One  23.88
17         Mesaerion: The Best Science ...    One  37.59
18             Libertarianism for Beginners    Two  51.33
19                  It's Only the Himalayas    Two  45.17
```



In [6]:

```
1000
('Girl Online On Tour ...', 'One', '£53.47', 'Girl Online On Tour (G
irl Online #2)', "The sequel to the number-one bestseller Girl Onlin
e. Penny joins her rock-star boyfriend, Noah, on his European music 
tour.Penny's bags are packed.When Noah invites Penny on his European
music tour, she can't wait to spend time with her rock-god-tastic bo
yfriend.But, between Noah's jam-packed schedule, less-than-welcoming
bandmates and threatening messages from jealous fan The sequel to th
e number-one bestseller Girl Online. Penny joins her rock-star boyfr
iend, Noah, on his European music tour.Penny's bags are packed.When 
Noah invites Penny on his European music tour, she can't wait to spe
nd time with her rock-god-tastic boyfriend.But, between Noah's jam-p
acked schedule, less-than-welcoming bandmates and threatening messag
es from jealous fans, Penny wonders whether she's really cut out for
life on tour. She can't help but miss her family, her best friend El
liot . . . and her blog, Girl Online.Can Penny learn to balance life
and love on the road, or will she lose everything in pursuit of the 
perfect summer? ...more", 'Young Adult')
```