تمرین‌های استخراج اطلاعات از وب

انجام تمرین‌های زیر

تمرین: هر کتابخانه‌ای که فکر می‌کنید برای استخراج اطلاعات از یک وبسایت نیاز دارید، وارد کنید.

In [28]:

```
# CODE HERE
```

In [1]:

```
import requests, bs4
```

تمرین: از کتابخانه requests و BeautifulSoup برای اتصال به http://quotes.toscrape.com/ (http://quotes.toscrape.com/) استفاده کنید و متن HTML صفحه اصلی را دریافت کنید.

In [30]:

```
# CODE HERE
```

In [2]:

```
res = requests.get("http://quotes.toscrape.com/")
```

In [3]:

```
res.text
```

```
\t<title>Quotes to Scrape</title>\n    <link rel="stylesheet" href="/st
atic/bootstrap.min.css">\n    <link rel="stylesheet" href="/static/mai
n.css">\n</head>\n<body>\n    <div class="container">\n        <div cla
ss="row header-box">\n            <div class="col-md-8">\n
<h1>\n                    <a href="/" style="text-decoration: none">Quo
tes to Scrape</a>\n                </h1>\n            </div>\n
<div class="col-md-4">\n                <p>\n                \n
<a href="/login">Login</a>\n                \n                </p>\n
</div>\n        </div>\n    \n\n<div class="row">\n    <div class="col-
md-8">\n\n    <div class="quote" itemscope itemtype="http://schema.org/
CreativeWork">\n        <span class="text" itemprop="text">“The world a
s we have created it is a process of our thinking. It cannot be changed
without changing our thinking.”</span>\n        <span>by <small class
="author" itemprop="author">Albert Einstein</small>\n        <a href="/
author/Albert-Einstein">(about)</a>\n        </span>\n        <div clas
s="tags">\n            Tags:\n            <meta class="keywords" itempr
op="keywords" content="change,deep-thoughts,thinking,world" /    > \n
\n            <a class="tag" href="/tag/change/page/1/">change</a>\n
\n            <a class="tag" href="/tag/deep-thoughts/page/1/">deep-tho
ughts</a>\n            \n            <a class="tag" href="/tag/thinkin
```

تمرین: نام نویسندگان موجود در صفحه اول را بدست آورید.

In [4]:

```
# CODE HERE
soup = bs4.BeautifulSoup(res.text, 'lxml')
```

In [5]:

```
soup.select('.author')
```

Out[5]:

```
[<small class="author" itemprop="author">Albert Einstein</small>,
 <small class="author" itemprop="author">J.K. Rowling</small>,
 <small class="author" itemprop="author">Albert Einstein</small>,
 <small class="author" itemprop="author">Jane Austen</small>,
 <small class="author" itemprop="author">Marilyn Monroe</small>,
 <small class="author" itemprop="author">Albert Einstein</small>,
 <small class="author" itemprop="author">André Gide</small>,
 <small class="author" itemprop="author">Thomas A. Edison</small>,
 <small class="author" itemprop="author">Eleanor Roosevelt</small>,
 <small class="author" itemprop="author">Steve Martin</small>]
```

In [6]:

```
authors = set()

for name in soup.select(".author"):
    authors.add(name.text)
```

In [7]:

```
authors
```

Out[7]:

```
{'Albert Einstein',
 'André Gide',
 'Eleanor Roosevelt',
 'J.K. Rowling',
 'Jane Austen',
 'Marilyn Monroe',
 'Steve Martin',
 'Thomas A. Edison'}
```

تمرین: یک لیست از تمامی نقل‌قول‌ها در صفحه اول ایجاد کنید.

In [13]:

```
#CODE HERE
```

In [8]:

```
quotes = []

for quote in soup.select(".text"):
    quotes.append(quote.text)
```

In [9]:

```
quotes
```

Out[9]:

```
['“The world as we have created it is a process of our thinking. It cannot
be changed without changing our thinking.”',
 '“It is our choices, Harry, that show what we truly are, far more than ou
r abilities.”',
 '“There are only two ways to live your life. One is as though nothing is
a miracle. The other is as though everything is a miracle.”',
 '“The person, be it gentleman or lady, who has not pleasure in a good nov
el, must be intolerably stupid.”',
 "“Imperfection is beauty, madness is genius and it's better to be absolut
ely ridiculous than absolutely boring.”",
 '“Try not to become a man of success. Rather become a man of value.”',
 '“It is better to be hated for what you are than to be loved for what you
are not.”',
 "“I have not failed. I've just found 10,000 ways that won't work.”",
 "“A woman is like a tea bag; you never know how strong it is until it's i
n hot water.”",
 '“A day without sunshine is like, you know, night.”']
```

تمرین: سایت را بررسی کنید و از Beautiful Soup برای استخراج ده برچسب برتر از متن درخواست نمایش داده شده در بالا سمت راست صفحه اصلی استفاده کنید (به عنوان مثال Love, Inspirational, Life و غیره). راهنمایی: به خاطر داشته باشید که تحت هر نقل‌قول همچنین برچسب‌ها وجود دارند، سعی کنید یک کلاس را پیدا کنید که فقط در برچسب‌های بالا سمت راست وجود دارد، شاید به span چک کنید.

In [16]:

```
# CODE HERE
```

In [10]:

```
soup = bs4.BeautifulSoup(res.text,'lxml')
```

In [11]:

```
soup.select('.tag-item')
```

Out[11]:

```
[<span class="tag-item">
 <a class="tag" href="/tag/love/" style="font-size: 28px">love</a>
 </span>,
 <span class="tag-item">
 <a class="tag" href="/tag/inspirational/" style="font-size: 26px">insp
irational</a>
 </span>,
 <span class="tag-item">
 <a class="tag" href="/tag/life/" style="font-size: 26px">life</a>
 </span>,
 <span class="tag-item">
 <a class="tag" href="/tag/humor/" style="font-size: 24px">humor</a>
 </span>,
 <span class="tag-item">
 <a class="tag" href="/tag/books/" style="font-size: 22px">books</a>
 </span>,
 <span class="tag-item">
 <a class="tag" href="/tag/reading/" style="font-size: 14px">reading</a
```

In [12]:

```
for item in soup.select(".tag-item"):
    print(item.text)
```

```
love

inspirational

life

humor

books

reading

friendship

friends

truth

simile
```

**تمرین: توجه کنید که بیش از یک صفحه وجود دارد و صفحات بعدی به این صورت به نظر می‌رسند: http://quotes.toscrape.com/page/2/ (http://quotes.toscrape.com/page/2/). با استفاده از آنچه که درباره حلقه for و اتصال رشته می‌دانید، از تمام صفحات عبور کرده و تمام نویسندگان منحصر به فرد در وبسایت را بدست آورید. به خاطر داشته باشید که راه‌های زیادی برای دست‌یابی به این هدف وجود دارد و همچنین توجه کنید که باید به طریقی بفهمید که حلقه شما در صفحه آخر نقل‌قول‌ها قرار دارد. برای اهداف اشکال

زدایی، به شما اطلاع می‌دهم که تنها 10 صفحه وجود دارد، بنابراین صفحه آخر به این صورت است: http://quotes.toscrape.com/page/10/ (http://quotes.toscrape.com/page/10/). اما سعی کنید یک حلقه‌ی قوی بسازید که به حدی قوی باشد که لازم نباشد تعداد صفحات را به طور پیشین بدانید، شاید از try/except برای این کار استفاده کنید، این به شما بستگی دارد!**

In [22]:

```
# CODE HERE
```

روش‌های دیگر زیادی وجود دارند که حتی قوی‌تر و انعطاف‌پذیرتر هستند، اما ایده اصلی همان است، استفاده از یک حلقه while برای گشتن در صفحات مختلف و شرط خروجی بر اساس صفحه نامعتبر است.

In [13]:

```
url = 'http://quotes.toscrape.com/page/'
```

راه حل اول

In [14]:

```
authors = set()

for page in range(1,10):

    # Concatenate to get new page URL
    page_url = url+str(page)
    # Obtain Request
    res = requests.get(page_url)
    # Turn into Soup
    soup = bs4.BeautifulSoup(res.text,'lxml')
    # Add Authors to our set
    for name in soup.select(".author"):
        authors.add(name.text)
```

راه حل دیگر

In [ ]:

```
page_url = url+str(9999999)

# Obtain Request
res = requests.get(page_url)

# Turn into Soup
soup = bs4.BeautifulSoup(res.text,'lxml')
```

In [ ]:

```
soup
```

In [ ]:

```
"No quotes found!" in res.text
```

In [ ]:

```
page_still_valid = True
authors = set()
page = 1

while page_still_valid:

    # Concatenate to get new page URL
    page_url = url+str(page)

    # Obtain Request
    res = requests.get(page_url)

    # Check to see if we're on the last page
    if "No quotes found!" in res.text:
        break

    # Turn into Soup
    soup = bs4.BeautifulSoup(res.text,'lxml')

    # Add Authors to our set
    for name in soup.select(".author"):
        authors.add(name.text)

    # Go to Next Page
    page += 1
```

In [ ]:

```
authors
```