Web Scraping using BeautifulSoup

Objectives:

- Using `requests` to download server-side rendered HTML code
- Using `BeautifulSoup` to parse HTML code

Scrape for Chinese Novels

We will demonstrate following website to demonstrate how to download and save articles.

- http://www.shuku.net:8082/novels/jinyong/yi/yi.html (http://www.shuku.net:8082/novels/jinyong/yi/yi.html)

For example, first article is in following website.

- http://www.shuku.net:8082/novels/jinyong/yi/yitian01.html (http://www.shuku.net:8082/novels/jinyong/yi/yitian01.html)

In [1]:

```
!pip install beautifulsoup4
```

```
Requirement already satisfied: beautifulsoup4 in c:\users\isszq\anaconda3\l
ib\site-packages (4.8.2)
Requirement already satisfied: soupsieve>=1.2 in c:\users\isszq\anaconda3\l
ib\site-packages (from beautifulsoup4) (1.9.5)
```

Extract List of Titles

Import libraries.

In [4]:

```
from bs4 import BeautifulSoup
import requests
```

Use `requests` to send GET request to server and download HTML.

- Use status code to make sure request is successful.

In [14]:

```
base_url = 'http://www.shuku.net:8082/novels/jinyong/yi/'

resp = requests.get(base_url + 'yi.html')
print(resp.status_code)
```

```
200
```

Make a soup from HTML code, which is in `resp.text` .

In [10]:

```python
soup = BeautifulSoup(resp.text)
```

Examine the article items in the HTML code. Each title are contained in a `<td>` element with attribute `align='center'` .

In [12]:

```python
items = soup.find_all('td', {'align':'center'})
```

Extract title and link for one item.

In [23]:

```python
item = items[0]
print(item.a)
print(item.a.text)
link = item.find('a').attrs['href']
print(link)
```

```
<a href="yitian01.html">天涯思君不可忘</a>
天涯思君不可忘
yitian01.html
```

Use for-loop to extract title and link pairs into a list.

In [25]:

```python
result = []
for item in items:
    if item.find('a'):
        title = item.a.text
        link = item.a.attrs['href']
        result.append((title, link))
```

Examine the final `result` list.

In [1]:

```python
print(len(result))
result[:5]
```

```
---------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-1-930a380903a4> in <module>
----> 1 print(len(result))
      2 result[:5]

NameError: name 'result' is not defined
```

Extract One Article

Try to extract first article.

- Use relative link to form the complete link to the article
- Use `requests` to download HTML and make a soup.

In [47]:

```python
title, link = result[0]

url = base_url + link
html = requests.get(url)
soup = BeautifulSoup(html.text)
```

The article is contained in `<pre>` tag.

Extract text from the tag and examine it.

- There are some extra text in the article. This is because of some redundant tags within the `<pre>` tag, e.g. `<script>` and `<a>` tags.

In [48]:

```python
item = soup.find('pre')
print(item.text)
```

```
                        第一回　天涯思君不可忘

　　“春游浩荡，是年年寒食，梨花时节。白锦无纹香烂漫，玉树琼苞堆雪。静夜
沉沉，浮光霭霭，冷浸溶溶月。人间天上，烂银霞照通彻。浑似姑射真人，天姿灵
秀，意气殊高洁。万蕊参差谁信道，不与群芳同列。浩气清英，仙才卓荦，下土难
分别。瑶台归去，洞天方看清绝。”

　　作这一首《无俗念》词的，乃南宋末年一位武学名家，有道之士。此人姓丘，
名处机，道号长春子，名列全真七子之一，是全真教中出类拔萃的人物。《词品》
评论此词道：“长春，世之所谓仙人也，而词之清拔如此”。这首词诵的似是梨花，
其实词中真意却是赞誉一位身穿白衣的美貌少女，说她“浑似姑射真人，天姿灵秀，
意气殊高洁”，又说她“浩气清英，仙才卓荦”，“不与群芳同列”。词中所颂这
美女，乃古墓派传人小龙女。她一生爱穿白衣，当真如风拂玉树，雪裹琼苞，兼之
生性清冷，实当得起“冷浸溶溶月”的形容，以“无俗念”三字赠之，可说十分贴
切。长春子丘处机和她在终南山上比邻而居，当年一见，便写下这首词来。

　　这时丘处机逝世已久，小龙女也已嫁与神雕大侠杨过为妻。在河南少室山山道
之上，却见有一个少女，正在低低念诵此词。这少女十八九岁年纪，身穿淡黄衣衫
```

Remove those redundant tags from existing soup.

In [49]:

```python
for s in soup.select('script'):
    s.extract()

for s in soup.select('a'):
    s.extract()
```

Extract the text again. The text is what we need to save to a text file.

In [50]:

```
item = soup.find('pre')
print(item.text)
```

```


                                第一回　天涯思君不可忘

     “春游浩荡，是年年寒食，梨花时节。白锦无纹香烂漫，玉树琼苞堆雪。静夜
沉沉，浮光霭霭，冷浸溶溶月。人间天上，烂银霞照通彻。浑似姑射真人，天姿灵
秀，意气殊高洁。万蕊参差谁信道，不与群芳同列。浩气清英，仙才卓荦，下土难
分别。瑶台归去，洞天方看清绝。”

     作这一首《无俗念》词的，乃南宋末年一位武学名家，有道之士。此人姓丘，
名处机，道号长春子，名列全真七子之一，是全真教中出类拔萃的人物。《词品》
评论此词道：“长春，世之所谓仙人也，而词之清拔如此”。这首词诵的似是梨花，
其实词中真意却是赞誉一位身穿白衣的美貌少女，说她“浑似姑射真人，天姿灵秀，
意气殊高洁”，又说她“浩气清英，仙才卓荦”，“不与群芳同列”。词中所颂这
美女，乃古墓派传人小龙女。她一生爱穿白衣，当真如风拂玉树，雪裹琼苞，兼之
生性清冷，实当得起“冷浸溶溶月”的形容，以“无俗念”三字赠之，可说十分贴
切。长春子丘处机和她在终南山上比邻而居，当年一见，便写下这首词来。

     这时丘处机逝世已久，小龙女也已嫁与神雕大侠杨过为妻。在河南少室山山道
之上，却见有一个少女，正在低低念诵此词。这少女十八九岁年纪，身穿淡黄衣衫
```

Save text to file using `<title>.text` as file name.

In [51]:

```
with open(title+'.txt', 'w', encoding='utf-8') as f:
    f.write(item.text)
```

Complete Solution

In [ ]:

```
import requests
from bs4 import BeautifulSoup

base_url = 'http://www.shuku.net:8082/novels/jinyong/yi'

# Step 1
def get_title_list():
    resp = requests.get(base_url + '/yi.html')
    html = resp.text
    soup = BeautifulSoup(html)
    items = soup.find_all('td', {'align':'center'})
    result = []

    for item in items:
        t = item.find('a')
        title = t.text
        link = t.attrs['href']
        result.append((title, link))
    return result

result = get_title_list()
print(result)

# Step 2

for item in result:
    title, link = item
    url = base_url + '/' + link
    resp = requests.get(url)
    html = resp.text
    soup = BeautifulSoup(html)
    item = soup.find('pre')
    article = item.text
    print('writing... ', title)
    with open(title+'.txt', 'w', encoding='utf-8') as f:
        f.write(article)
```