Data scraping with

Requests, BeautifulSoup and Selenium

Installing Beautiful Soup

```
$ pip install beautifulsoup4
```

or

install with PyCharm

Test your module.

```
>>> from bs4 import BeautifulSoup
```

HTML Document - Requests

```
>>> import requests
>>> req = requests.get(“http://samblog.org/?lang=en”)
>>> html_doc = req.content
>>> from bs4 import BeautifulSoup
>>> soup = BeautifulSoup(html_doc, 'html.parser')
>>> soup.title
<title>
        SamBlog – Şehir Araştırmaları Merkezi
</title>
>>> soup.title.name
title
>>> soup.title.string
SamBlog – Şehir Araştırmaları Merkezi
>>> soup.title.parent.name
head
>>> soup.a
<a class="header-9" id="close-sidebar-nav"><i class="penci-faicon fa fa-close"></i></a>
>>> soup.a[‘class']
[‘header-9’’]
```

```
>>> soup.find_all('a')
```

```
[<a class="header-9" id="close-sidebar-nav"><i class="penci-faicon fa fa-close"></i></a>, <a
href="http://samblog.org/?lang=en"><img alt="SamBlog" class="penci-lazy"
data-src="https://samblog.sehir.edu.tr/wp-content/uploads/2018/06/TR_BEYAZZEMIN-2.png"
src="http://samblog.org/wp-content/themes/soledad/images/penci-holder.png"/></a>, <a
href="https://www.facebook.com/sehirsam/" rel="nofollow" target="_blank"><i class="penci-faicon fa
fa-facebook"></i></a>, <a href="https://twitter.com/SamSehir" rel="nofollow" target="_blank"><i
class="penci-faicon fa fa-twitter"></i></a>, <a href="mailto:sam@sehir.edu.tr"><i class="penci-faicon fa
fa-envelope"></i></a>, <a href="http://samblog.org/who-we-are/?lang=en">Who We’re</a>, … ]
```

```
>>> soup.find(class_=‘retweet')
```

```
<a class="retweet" href="https://twitter.com/intent/retweet?tweet_id=1250712848227078149"
target="_blank">Retweet</a>
```

For loop in links...

```
>>> for link in soup.find_all('a'):
        print(link.get('href'))
None
http://samblog.org/?lang=en
https://www.facebook.com/sehirsam/
https://twitter.com/SamSehir
mailto:sam@sehir.edu.tr
http://samblog.org/who-we-are/?lang=en
http://samblog.org/projects/?lang=en
…
```

Searching by CSS class

```
>>> soup.find("a", class_="favorite")
<a class="favorite"
href="https://twitter.com/intent/favorite?tweet_id=1250712848227078149"
target="_blank">Favorite</a>
```

```
>>> soup.find_all("a", class_=“favorite")
```

or

```
>>> soup.find_all("a", attrs={"class": "favorite"})
[<a class="favorite"
href="https://twitter.com/intent/favorite?tweet_id=1250712848227078149"
target="_blank">Favorite</a>, <a class="favorite"
href="https://twitter.com/intent/favorite?tweet_id=1249741205895864323"
target="_blank">Favorite</a>, <a class="favorite"
href="https://twitter.com/intent/favorite?tweet_id=1247085574261035009"
target="_blank">Favorite</a>, … ]
```

Selenium

Selenium uses your browser. Therefore, if webpage is written in JavaScript, you can scrape it without any problem.

Installation

```
>>> pip install selenium
```

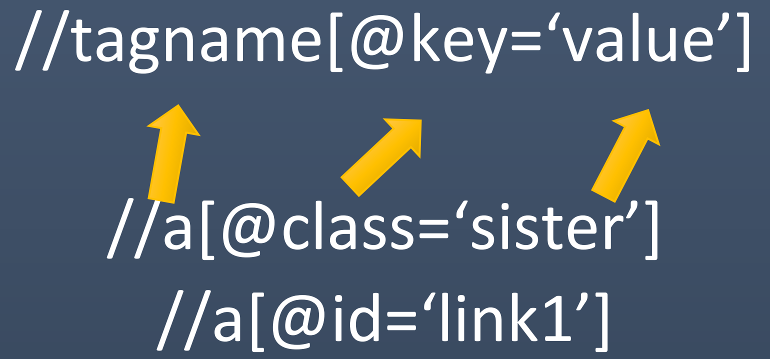
```
from selenium import webdriver


driver = webdriver.Chrome(‘chromedriver.exe’)  <---- it uses Chrome Browser
driver.get("http://www.python.org")
elem = driver.find_element_by_name("q")
html = driver.page_source
driver.close()
```

Selenium

```
element = driver.find_element_by_id("passwd-id")
element = driver.find_element_by_name("passwd")
element = driver.find_element_by_xpath("//input[@id='passwd-id']")
```

```
<a href="http://example.com/elsie" class="sister" id="link1">Elsie</a>
```



//tagname[@key=‘value’]

//a[@class=‘sister’]

//a[@id=‘link1’]

Selenium

You can use

- Click action
- Drag and drop
- Fill forms
- Parsing iframes
- Collect cookies
- Solve CAPTCHA
- Whatever the browser can do.

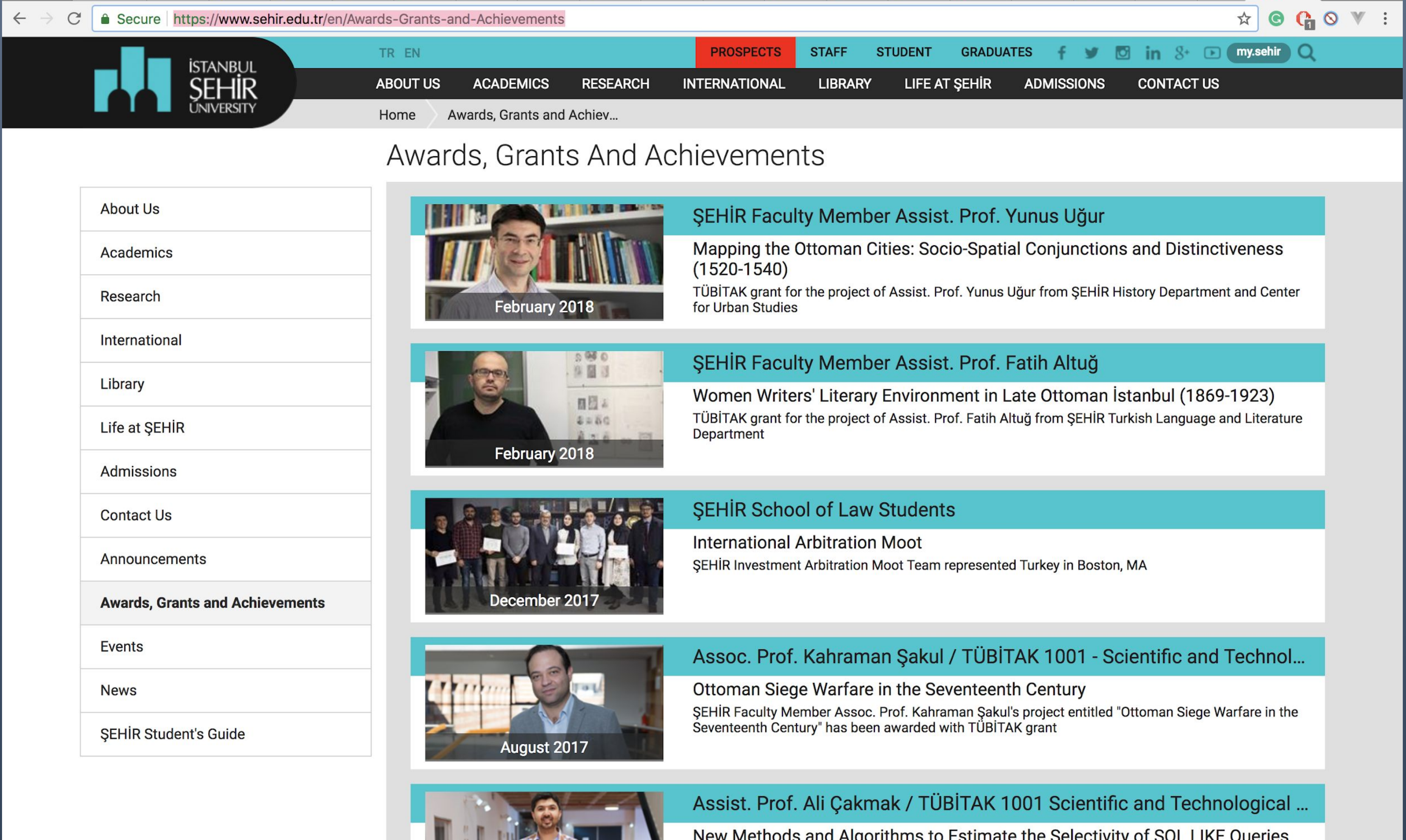
Searching in Selenium

- find_element(by='id', value=None)
- find_element_by_class_name(name)
- find_element_by_css_selector(css_selector)
- find_element_by_id(id_)
- find_element_by_link_text(link_text)
- find_element_by_name(name)
- find_element_by_partial_link_text(link_text)
- find_element_by_tag_name(name)
- find_element_by_xpath(xpath)
- And their plural versions (find_elements___)

Practice Session (Week 9) - BeautifulSoup

In this practice session you will become a web scraper from Sehir website. Using your BeautifulSoup skills that you learned a little time ago you will fetch Awards, Grants and Achievements about Istanbul Sehir University's web page in details that with title, description, date and link of the page.

https://www.sehir.edu.tr/en/Awards-Grants-and-Achievements



Awards, Grants And Achievements

February 2018

ŞEHİR Faculty Member Assist. Prof. Yunus Uğur

Mapping the Ottoman Cities: Socio-Spatial Conjunctions and Distinctiveness (1520-1540)

TÜBİTAK grant for the project of Assist. Prof. Yunus Uğur from ŞEHİR History Department and Center for Urban Studies

February 2018

ŞEHİR Faculty Member Assist. Prof. Fatih Altuğ

Women Writers' Literary Environment in Late Ottoman İstanbul (1869-1923)

TÜBİTAK grant for the project of Assist. Prof. Fatih Altuğ from ŞEHİR Turkish Language and Literature Department

December 2017

ŞEHİR School of Law Students

International Arbitration Moot

ŞEHİR Investment Arbitration Moot Team represented Turkey in Boston, MA

August 2017

Assoc. Prof. Kahraman Şakul / TÜBİTAK 1001 - Scientific and Technol...

Ottoman Siege Warfare in the Seventeenth Century

ŞEHİR Faculty Member Assoc. Prof. Kahraman Şakul's project entitled "Ottoman Siege Warfare in the Seventeenth Century" has been awarded with TÜBİTAK grant

Assist. Prof. Ali Çakmak / TÜBİTAK 1001 Scientific and Technological ...

New Methods and Algorithms to Estimate the Selectivity of SQL LIKE Queries

From LMS

Download *Practice.Session.Week11.pdf* file