We'll go to a website, decide what information we want, see where and how it is stored, then scrape it and set it as a pandas dataframe.

Some thing that should be considered before scraping a website:

1. We should check a site's terms and conditions before you scrape.
2. Space out you requests so you don't overload the site's server, doing this could get us blocked.
3. Scrapers break after time - web pages change their layout all the time, we'll more likely have to rewrite your code.
4. Web pages are usually inconsistent, more than likely we'll have to clean up after scraping it.
5. Every web page and situation is different, we'll have to spend time configuring your scraper.

In [2]:

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

In [20]:

```python
import pandas as pd
from pandas import Series, DataFrame
```

We'll look at some legislative reports from the University of California Web Page.

In [6]:

```python
url = 'https://www.ucop.edu/operating-budget/budgets-and-reports/legislative-reports/2013-1
```

In [7]:

```python
result = requests.get( url )

c = result.content
```

In [8]:

```python
soup = BeautifulSoup(c)
```

Now using BeautifulSoup to search for the table we want to grab

In [11]:

```python
# Moving to the section of interest
summary = soup.find('div' , {'class':'list-land','id':'content'})

# Find the tables in teh HTML
tables = summary.find_all('table')
```

Now we'll find table entries . A 'td' tag defines a standard cell in a html table. The 'tr' tag defines a row in an html table.
We'll parse through our tables object and try to find each cell using the findALL('td') method.

In [16]:

```
# Set up empty data list
data = []

rows = tables[0].findAll('tr')

for tr in rows :
    cols = tr.findAll('td')

    for td in cols :

        text = td.find(text=True)
        print (text)
        data.append(text)
```

```
1
08/01/13
2013-14 (EDU 92495) Proposed Capital Outlay Projects (2013-14 only) (pdf)
2
09/01/13
2014-15  (EDU 92495) Proposed Capital Outlay Projects (pdf)
3
11/01/13
Utilization of Classroom and Teaching Laboratories (pdf)
4
11/01/13
Instruction and Research Space Summary & Analysis (pdf)
5
11/15/13
Statewide Energy Partnership Program (pdf)
6
11/30/13
2013-23 Capital Financial Plan (pdf)
7
11/30/13
Projects Savings Funded from Capital Outlay Bond Funds (pdf)
8
12/01/13
Streamlined Capital Projects Funded from Capital (pdf)
9
01/01/14
Annual General Obligation Bonds Accountability (pdf)
10
01/01/14
Small Business Utilization (pdf)
11
01/01/14
Institutional Financial Aid Programs - Preliminary report (pdf)
12
01/10/14
Summer Enrollment (pdf)
13
01/15/14
Contracting Out for Services at Newly Developed Facilities (pdf)
14
03/01/14
Performance Measures (pdf)
15
03/01/14
Entry Level Writing Requirement (pdf)
```

```
16
03/31/14
Annual Report on Student Financial Support (pdf)
17
04/01/14
Unique Statewide Pupil Identifier (pdf)
18
04/01/14
Riverside School of Medicine (pdf)
19
04/01/14
SAPEP Funds and Outcomes - N/A
20
05/15/14
Receipt and Use of Lottery Funds (pdf)
21
07/01/14
Cogeneration and Energy Consv Major Capital Projects (pdf)







Future Reports



24
12-
Breast Cancer Research Fund
25
12-31-15
Cigarette and Tobacco Products Surtax Research Program
26
01-01-16
Best Value Program
27
01-01-16
California Subject Matter Programs
28
04-01-16
COSMOS Program Outcomes
```

In [17]:

```
data
```

Out[17]:

```
['1',
 '08/01/13',
 '2013-14 (EDU 92495) Proposed Capital Outlay Projects (2013-14 only) (pd
f)',
 '2',
 '09/01/13',
 '2014-15\xa0 (EDU 92495) Proposed Capital Outlay Projects (pdf)',
 '3',
 '11/01/13',
 'Utilization of Classroom and Teaching Laboratories (pdf)',
 '4',
 '11/01/13',
 'Instruction and Research Space Summary & Analysis (pdf)',
 '5',
 '11/15/13',
 'Statewide Energy Partnership Program (pdf)',
 '6',
 '11/30/13',
 '2013-23 Capital Financial Plan (pdf)',
 '7',
 '11/30/13',
 'Projects Savings Funded from Capital Outlay Bond Funds (pdf)',
 '8',
 '12/01/13',
 'Streamlined Capital Projects Funded from Capital (pdf)',
 '9',
 '01/01/14',
 'Annual General Obligation Bonds Accountability (pdf)',
 '10',
 '01/01/14',
 'Small Business Utilization (pdf)',
 '11',
 '01/01/14',
 'Institutional Financial Aid Programs - Preliminary report (pdf)',
 '12',
 '01/10/14',
 'Summer Enrollment (pdf)',
 '13',
 '01/15/14',
 'Contracting Out for Services at Newly Developed Facilities (pdf)',
 '14',
 '03/01/14',
 'Performance Measures (pdf)',
 '15',
 '03/01/14',
 'Entry Level Writing Requirement (pdf)',
 '16',
 '03/31/14',
 'Annual Report on Student\xa0Financial Support (pdf)',
 '17',
 '04/01/14',
 'Unique Statewide Pupil Identifier (pdf)',
 '18',
 '04/01/14',
 'Riverside School of Medicine (pdf)',
 '19',
 '04/01/14',
 'SAPEP Funds and Outcomes - N/A',
 '20',
 '05/15/14',
 'Receipt and Use of Lottery Funds (pdf)',
 '21',
 '07/01/14',
 'Cogeneration and Energy Consv Major Capital Projects (pdf)',
 '\n',
 '\n',
 '\n',
 '\xa0',
 'Future Reports',
 '\n',
 '24',
 '12-',
 'Breast Cancer Research Fund',
 '25',
 '12-31-15',
 'Cigarette and Tobacco Products Surtax Research Program',
 '26',
 '01-01-16',
 'Best Value Program',
 '27',
 '01-01-16',
 'California Subject Matter Programs',
 '28',
 '04-01-16',
 'COSMOS Program Outcomes']
```

In [18]:

```
reports = []
date =[]

index = 0

for item in data:
    if 'pdf' in item:
        date.append(data[index-1])

        reports.append(item.replace(u'\xa0',u' '))

    index +=1
```

In [21]:

```
date = Series(date)
reports = Series(reports)
```

In [24]:

```
legislative_df = pd.concat([date,reports], axis=1)
```

In [25]:

```
legislative_df.columns = ['Date' , 'Report']
```

In [26]:

```
legislative_df
```

Out[26]:

	Date	Report
0	08/01/13	2013-14 (EDU 92495) Proposed Capital Outlay Pr...
1	09/01/13	2014-15 (EDU 92495) Proposed Capital Outlay P...
2	11/01/13	Utilization of Classroom and Teaching Laborato...
3	11/01/13	Instruction and Research Space Summary & Analy...
4	11/15/13	Statewide Energy Partnership Program (pdf)
5	11/30/13	2013-23 Capital Financial Plan (pdf)
6	11/30/13	Projects Savings Funded from Capital Outlay Bo...
7	12/01/13	Streamlined Capital Projects Funded from Capit...
8	01/01/14	Annual General Obligation Bonds Accountability...
9	01/01/14	Small Business Utilization (pdf)
10	01/01/14	Institutional Financial Aid Programs - Prelimi...
11	01/10/14	Summer Enrollment (pdf)
12	01/15/14	Contracting Out for Services at Newly Develope...
13	03/01/14	Performance Measures (pdf)
14	03/01/14	Entry Level Writing Requirement (pdf)
15	03/31/14	Annual Report on Student Financial Support (pdf)
16	04/01/14	Unique Statewide Pupil Identifier (pdf)
17	04/01/14	Riverside School of Medicine (pdf)
18	05/15/14	Receipt and Use of Lottery Funds (pdf)
19	07/01/14	Cogeneration and Energy Consv Major Capital Pr...