Data Wrangling homework assigment + Bonus

Problem 2: Scrape CCNY calendar

Create a private repo on github called calscrape. Share it with mdogy. Create an ipython notebook and commit it. I want you to keep committing it as you edit it so I can see it build up. Your mission is to scrape the cuny fall 2018 academic calendar site using the requests library, beautiful soup and pandas, just like in class. You should wind up with a pandas data frame where the index column is a python date. There should be a column for "day of the week" with variable lable dow and a column called text with the explanation.

Print out this data frame. 5 extra credit points if you figure out how to use the google api to programatically add these dates to your google calendar with the request library.

In [1]:

```
import requests
from bs4 import BeautifulSoup as BS
from IPython.display import HTML
import pandas as pd
```

Inspect Calendar URL

In [2]:

```
req = requests.get("https://www.ccny.cuny.edu/registrar/fall-2018-academic-calendar")
```

In [3]:

```
code = req.status_code

if code == 200:
    print('It\'s working')
print(code)
```

```
It's working
200
```

In [4]:

```
req.text;
```

Obtain Beatiful Soup object

In [5]:

```
soup = BS(req.text,'html.parser')
#print(soup.prettify)
```

In [6]:

```
dir(soup);
```

In [7]:

```
soup.title
```

Out[7]:

```
<title>Fall 2018 Academic Calendar | The City College of New York</t
itle>
```

In [8]:

```
soup.title.name
```

Out[8]:

```
'title'
```

In [9]:

```
soup.title.text
```

Out[9]:

```
'Fall 2018 Academic Calendar | The City College of New York'
```

NOTICE THERE IS NO PARAGRAPH TAGS IN THIS DATA

In [10]:

```
soup.p
```

Out[10]:

```
<p> </p>
```

A good refactor would decipher that _p_ string search doesnt correspond to a paragraph search.

In [11]:

```
all_p = soup.find_all('p')
len(all_p)
```

Out[11]:

```
141
```

Inspect wanted table(s)

In [12]:

```
all_tables = soup.find_all('table')
len(all_tables)
```

Out[12]:

```
1
```

In [13]:

```
all_tables[0];
```

In [14]:

```
print(type(all_tables[0]))
```

```
<class 'bs4.element.Tag'>
```

MODIFIED SCRIPT

In [15]:

```
from IPython.display import HTML
```

In [16]:

```
try:
    table_classes = [table['class'] for table in all_tables]
    print(table_classes)
    wikitables = [table for table in all_tables if 'wikitable' in table['class']
]
    print('WikiTables first \n',HTML(str(wikitables[0])))
    print('WikiTables  second \n', HTML(str(wikitables[0])))
except:
    print('The table tag has unnammed html class(es)')
```

```
The table tag has unnammed html class(es)
```

In [17]:

```
cuny_calendar = [table for table in all_tables][0]
rows = cuny_calendar.find_all('tr')
HTML(str(rows));
```

In [18]:

```
HTML(str(rows[0]))
```

Out[18]:

DATES

DAYS

In [19]:

```
value_rows = [[data.text.strip() for data in row.find_all('td')] for row in rows
[1:]]
series_list = list(zip(*value_rows))
```

In [20]:

```
col_head = [head.text.strip().lower().replace(' ', '_').replace('.', '') for hea
d in rows[0].find_all('td')]
```

In [21]:

```
df = pd.DataFrame(dict(zip(col_head, series_list)))
df.columns = ['date', 'day of the week', 'Note']
df.reindex(df['date'])
df
```

Out[21]:

	date	day of the week	Note
0	August		
1	August 01	Wednesday	Application for degree for January and Februar...
2	August 20	Monday	Last day to apply for Study Abroad
3	August 26	Sunday	Last day of Registration;Last day to file ePer...
4	August 27	Monday	Classes begin;Initial Registration Appeals beg...
5	August 27 –September 02	Monday –Sunday	Change of program period; late fees apply
6	August 28	Tuesday	Last day to submit a request for Independent S...

7	September			
8	September 01	Saturday	First day of Saturday Classes	
9	September 02	Sunday	Last day to add a class to an existing enrollm...	
10	September 03	Monday	Course Withdrawal drop period begins (A grade ...	
11	September 05	Wednesday	Classes follow a Monday schedule	
12	September 09	Sunday	Last day to drop for 50% tuition refund	
13	September 10 –September 11	Monday –Tuesday	No Classes scheduled	
14	September 16	Sunday	Last day for 25% tuition refund;Census date – ...	
15	September 17	Monday	Course withdrawal period begins (A grade of “W...	
16	September 18 –September 19	Tuesday – Wednesday	No classes scheduled	
17	September 20	Thursday	Freshman Convocation	
18	September 25	Tuesday	Last day to submit proof of immunization for N...	
19	September 26	Wednesday	Assignment of “WA” grades for immunization non...	
20	October			
21	October 08	Monday	College Closed	
22	October 12	Friday	Deadline for filing application for degree for...	
23	November			
24	November 01	Thursday	INC grades for Spring 2018 and Summer 2018 for...	
25	November 05	Monday	60% date of the term	
26	November 06	Tuesday	Course withdrawal period ends. Last day to wi...	
27	November 22 –November 25	Thursday –Sunday	College Closed	
28	December			
29	December 08	Saturday	Last day of Saturday classes	
30	December 12	Wednesday	Last day of Classes	
31	December 13	Thursday	Reading Day	
32	December 14	Friday	Reading Day / Final Examinations	
33	December 15 –December 16	Saturday – Sunday	Final Examinations – Weekend Classes	
34	December 17 –December 21	Monday – Friday	Final Examinations – Day/Evening Classes	

35	December 21	Friday	End of Term
36	December 24-25	Monday –Tuesday	College Closed
37	December 28	Friday	Final Grade Submission Deadline
38	December 31	Monday	College Closed
39	January		
40	January 1, 2019	Tuesday	College Closed

In [22]:

```
!pip install --upgrade google-api-python-client google-auth-httplib2 google-auth
-oauthlib
```

```
Requirement already up-to-date: google-api-python-client in /anacond
a3/lib/python3.7/site-packages (1.7.11)
Requirement already up-to-date: google-auth-httplib2 in /anaconda3/l
ib/python3.7/site-packages (0.0.3)
Requirement already up-to-date: google-auth-oauthlib in /anaconda3/l
ib/python3.7/site-packages (0.4.0)
Requirement already satisfied, skipping upgrade: httplib2<1dev,>=0.9
.2 in /anaconda3/lib/python3.7/site-packages (from google-api-python
-client) (0.13.1)
Requirement already satisfied, skipping upgrade: six<2dev,>=1.6.1 in
/anaconda3/lib/python3.7/site-packages (from google-api-python-clien
t) (1.12.0)
Requirement already satisfied, skipping upgrade: google-auth>=1.4.1
in /anaconda3/lib/python3.7/site-packages (from google-api-python-cl
ient) (1.6.3)
Requirement already satisfied, skipping upgrade: uritemplate<4dev,>=
3.0.0 in /anaconda3/lib/python3.7/site-packages (from google-api-pyt
hon-client) (3.0.0)
Requirement already satisfied, skipping upgrade: requests-oauthlib>=
0.7.0 in /anaconda3/lib/python3.7/site-packages (from google-auth-oa
uthlib) (1.2.0)
Requirement already satisfied, skipping upgrade: rsa>=3.1.4 in /anac
onda3/lib/python3.7/site-packages (from google-auth>=1.4.1->google-a
pi-python-client) (4.0)
Requirement already satisfied, skipping upgrade: cachetools>=2.0.0 i
n /anaconda3/lib/python3.7/site-packages (from google-auth>=1.4.1->g
oogle-api-python-client) (3.1.1)
Requirement already satisfied, skipping upgrade: pyasn1-modules>=0.2
.1 in /anaconda3/lib/python3.7/site-packages (from google-auth>=1.4.
1->google-api-python-client) (0.2.6)
Requirement already satisfied, skipping upgrade: requests>=2.0.0 in
/anaconda3/lib/python3.7/site-packages (from requests-oauthlib>=0.7.
0->google-auth-oauthlib) (2.21.0)
Requirement already satisfied, skipping upgrade: oauthlib>=3.0.0 in
/anaconda3/lib/python3.7/site-packages (from requests-oauthlib>=0.7.
0->google-auth-oauthlib) (3.1.0)
Requirement already satisfied, skipping upgrade: pyasn1>=0.1.3 in /a
naconda3/lib/python3.7/site-packages (from rsa>=3.1.4->google-auth>=
1.4.1->google-api-python-client) (0.4.7)
Requirement already satisfied, skipping upgrade: certifi>=2017.4.17
in /anaconda3/lib/python3.7/site-packages (from requests>=2.0.0->req
uests-oauthlib>=0.7.0->google-auth-oauthlib) (2018.11.29)
Requirement already satisfied, skipping upgrade: urllib3<1.25,>=1.21
.1 in /anaconda3/lib/python3.7/site-packages (from requests>=2.0.0->
requests-oauthlib>=0.7.0->google-auth-oauthlib) (1.24.1)
Requirement already satisfied, skipping upgrade: idna<2.9,>=2.5 in /
anaconda3/lib/python3.7/site-packages (from requests>=2.0.0->request
s-oauthlib>=0.7.0->google-auth-oauthlib) (2.8)
Requirement already satisfied, skipping upgrade: chardet<3.1.0,>=3.0
.2 in /anaconda3/lib/python3.7/site-packages (from requests>=2.0.0->
requests-oauthlib>=0.7.0->google-auth-oauthlib) (3.0.4)
```

Extra Credit script to add Google Calendar events

Create 'credentials.json' with your username and password
https://developers.google.com/calendar/quickstart/python (https://developers.google.com/calendar/quickstart/python)

In [23]:

```
!python quickstart.py
```

```
Getting the upcoming 10 events
2019-09-26T11:30:00-04:00 ORCA Symposium
2019-09-26T12:30:00-04:00 ORCA Symposium
2019-09-26T15:30:00-04:00 AIAA
2019-09-27T21:00:00-04:00 EEG results
2019-09-28 Bubble Run
2019-09-28T08:00:00-04:00 Fun Run-5K - Individual Registration - 8:0
0AM Heat
2019-09-28T17:00:00-04:00 Play meeting
2019-10-01T08:30:00-04:00 Adomako x Mark| run coach
2019-10-01T19:00:00-04:00 Register for race SIH
2019-10-08T08:30:00-04:00 Adomako x Mark| run coach
```

In [ ]:

```

```