

OfficeTricks.Com Tutorial- Python

Source: <https://officetricks.com/how-to-scrap-web-data-python/>

Python 2.7/3 Tutorial - How to scrape a simple website using BeautifulSoup and requests

Data scraping is a method to get data from the websites for business purposes, analyzing and countless purposes used by businessmen and programmers. Python provides best of tools and libraries for carrying out data scraping work. In this tutorial, We'll see how to scrape a table on a website using Python's libraries and with a very simple coding. I assume you have python installed, and the basic knowledge of Python, to execute a program and install libraries with pip. These tutorials are based on Python 2.7.12, However the Python 3.X versions shows the same syntax on this parts.

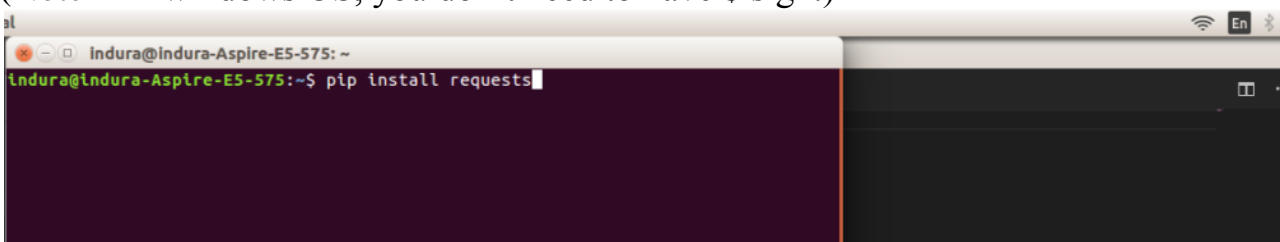
01. First, We have to install libraries Bs4, requests and lxml. You can install these libraries with this command run on the command prompt.

```
$ pip install requests
```

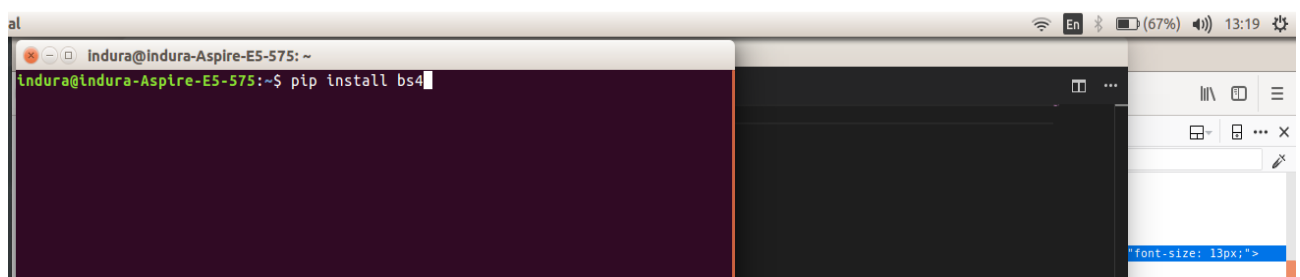
```
$ pip install bs4
```

```
$ pip install lxml
```

(Note – In windows OS, you don't need to have \$ sign.)

A screenshot of a terminal window on a Linux system. The terminal title is 'Indura@Indura-Aspire-E5-575: ~'. The prompt is 'Indura@Indura-Aspire-E5-575:~\$' and the command 'pip install requests' has been entered. The terminal background is dark purple.

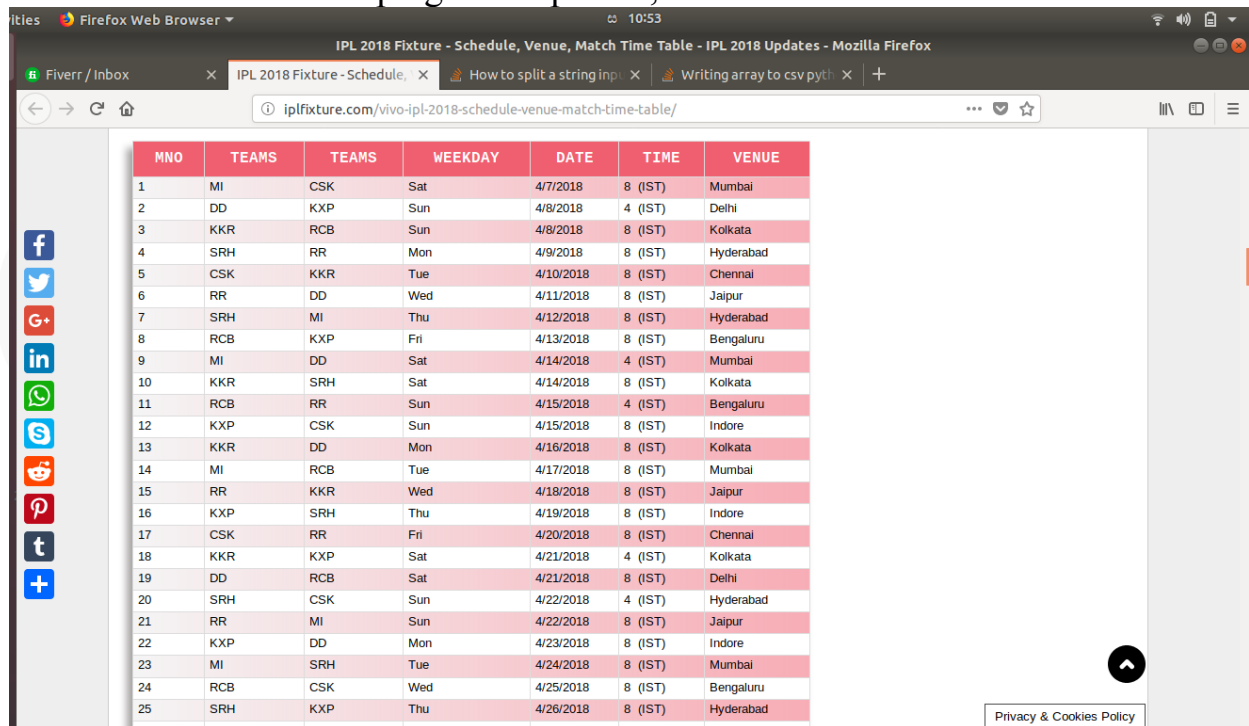
Installing requests

A screenshot of a terminal window on a Linux system. The terminal title is 'Indura@Indura-Aspire-E5-575: ~'. The prompt is 'Indura@Indura-Aspire-E5-575:~\$' and the command 'pip install bs4' has been entered. The terminal background is dark purple. A system tray at the bottom right shows the time as 13:19 and battery level at 67%.

Installing bs4

02. Once you've completed installing all libraries, We can move to the coding part.

This is the **table** We're hoping to scrape and, save it as csv file as a table.



MNO	TEAMS	TEAMS	WEEKDAY	DATE	TIME	VENUE
1	MI	CSK	Sat	4/7/2018	8 (IST)	Mumbai
2	DD	KXP	Sun	4/8/2018	4 (IST)	Delhi
3	KKR	RCB	Sun	4/8/2018	8 (IST)	Kolkata
4	SRH	RR	Mon	4/9/2018	8 (IST)	Hyderabad
5	CSK	KKR	Tue	4/10/2018	8 (IST)	Chennai
6	RR	DD	Wed	4/11/2018	8 (IST)	Jaipur
7	SRH	MI	Thu	4/12/2018	8 (IST)	Hyderabad
8	RCB	KXP	Fri	4/13/2018	8 (IST)	Bengaluru
9	MI	DD	Sat	4/14/2018	4 (IST)	Mumbai
10	KKR	SRH	Sat	4/14/2018	8 (IST)	Kolkata
11	RCB	RR	Sun	4/15/2018	4 (IST)	Bengaluru
12	KXP	CSK	Sun	4/15/2018	8 (IST)	Indore
13	KKR	DD	Mon	4/16/2018	8 (IST)	Kolkata
14	MI	RCB	Tue	4/17/2018	8 (IST)	Mumbai
15	RR	KKR	Wed	4/18/2018	8 (IST)	Jaipur
16	KXP	SRH	Thu	4/19/2018	8 (IST)	Indore
17	CSK	RR	Fri	4/20/2018	8 (IST)	Chennai
18	KKR	KXP	Sat	4/21/2018	4 (IST)	Kolkata
19	DD	RCB	Sat	4/21/2018	8 (IST)	Delhi
20	SRH	CSK	Sun	4/22/2018	4 (IST)	Hyderabad
21	RR	MI	Sun	4/22/2018	8 (IST)	Jaipur
22	KXP	DD	Mon	4/23/2018	8 (IST)	Indore
23	MI	SRH	Tue	4/24/2018	8 (IST)	Mumbai
24	RCB	CSK	Wed	4/25/2018	8 (IST)	Bengaluru
25	SRH	KXP	Thu	4/26/2018	8 (IST)	Hyderabad

Website = <http://iplfixture.com/vivo-ipl-2018-schedule-venue-match-time-table/>

This is how it looks like on our csv file after extracting data.

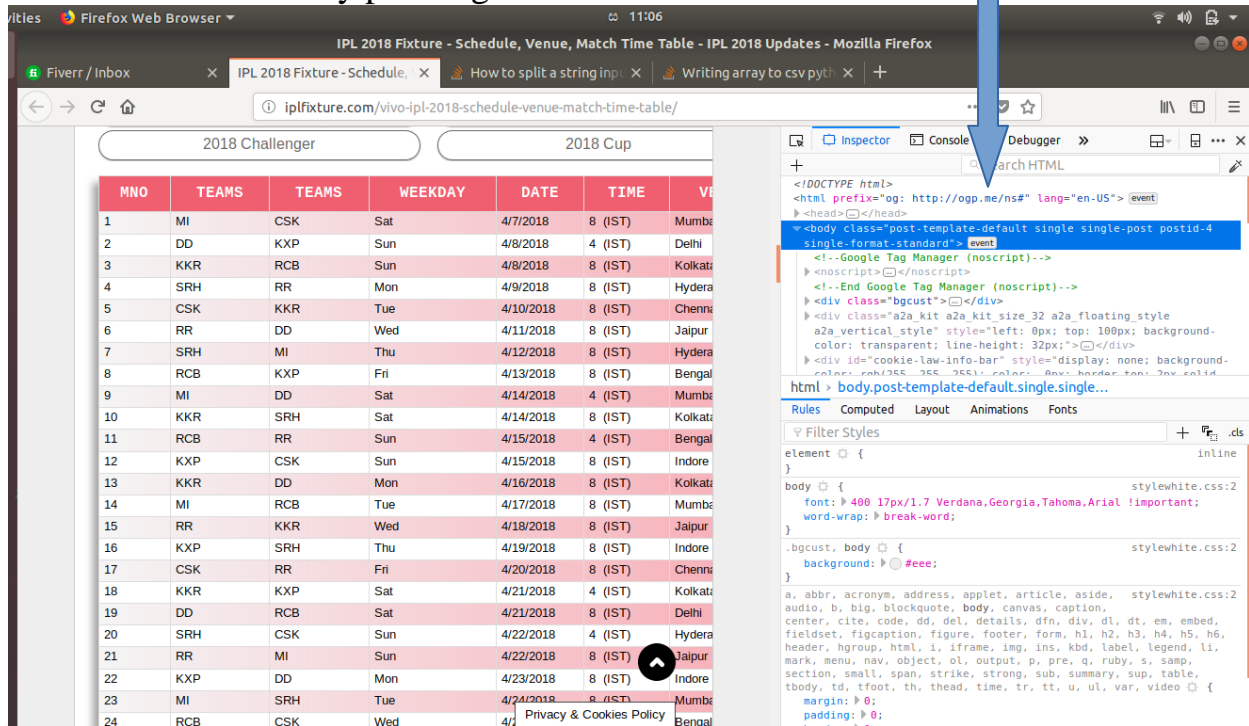
Mno	Teams	Teams	WeekDay	Date	Time	Venue	Venue
1	MI	CSK	Sat	4/7/2018	8 (IST)	Mumbai	
2	DD	KXP	Sun	4/8/2018	4 (IST)	Delhi	
3	KKR	RCB	Sun	4/8/2018	8 (IST)	Kolkata	
4	SRH	RR	Mon	4/9/2018	8 (IST)	Hyderabad	
5	CSK	KKR	Tue	4/10/2018	8 (IST)	Chennai	
6	RR	DD	Wed	4/11/2018	8 (IST)	Jaipur	
7	SRH	MI	Thu	4/12/2018	8 (IST)	Hyderabad	
8	RCB	KXP	Fri	4/13/2018	8 (IST)	Bengaluru	
9	MI	DD	Sat	4/14/2018	4 (IST)	Mumbai	
10	KKR	SRH	Sat	4/14/2018	8 (IST)	Kolkata	
11	RCB	RR	Sun	4/15/2018	4 (IST)	Bengaluru	
12	KXP	CSK	Sun	4/15/2018	8 (IST)	Indore	
13	KKR	DD	Mon	4/16/2018	8 (IST)	Kolkata	
14	MI	RCB	Tue	4/17/2018	8 (IST)	Mumbai	
15	RR	KKR	Wed	4/18/2018	8 (IST)	Jaipur	
16	KXP	SRH	Thu	4/19/2018	8 (IST)	Indore	
17	CSK	RR	Fri	4/20/2018	8 (IST)	Chennai	
18	KKR	KXP	Sat	4/21/2018	4 (IST)	Kolkata	
19	DD	RCB	Sat	4/21/2018	8 (IST)	Delhi	
20	SRH	CSK	Sun	4/22/2018	4 (IST)	Hyderabad	
21	RR	MI	Sun	4/22/2018	8 (IST)	Jaipur	
22	KXP	DD	Mon	4/23/2018	8 (IST)	Indore	
23	MI	SRH	Tue	4/24/2018	8 (IST)	Mumbai	
24	RCB	CSK	Wed	4/25/2018	8 (IST)	Bengaluru	
25	SRH	KXP	Thu	4/26/2018	8 (IST)	Hyderabad	
26	DD	KKR	Fri	4/27/2018	8 (IST)	Delhi	
27	CSK	MI	Sat	4/28/2018	8 (IST)	Chennai	
28	RR	SRH	Sun	4/29/2018	4 (IST)	Jaipur	
29	RCB	KKR	Sun	4/29/2018	8 (IST)	Bengaluru	

03. Ready to do the trick? Well then, better start coding. First you have to choose an IDE for coding with python. You can use Sublime text or some text editor you're familiar with. (Hopefully, not the notepad.)

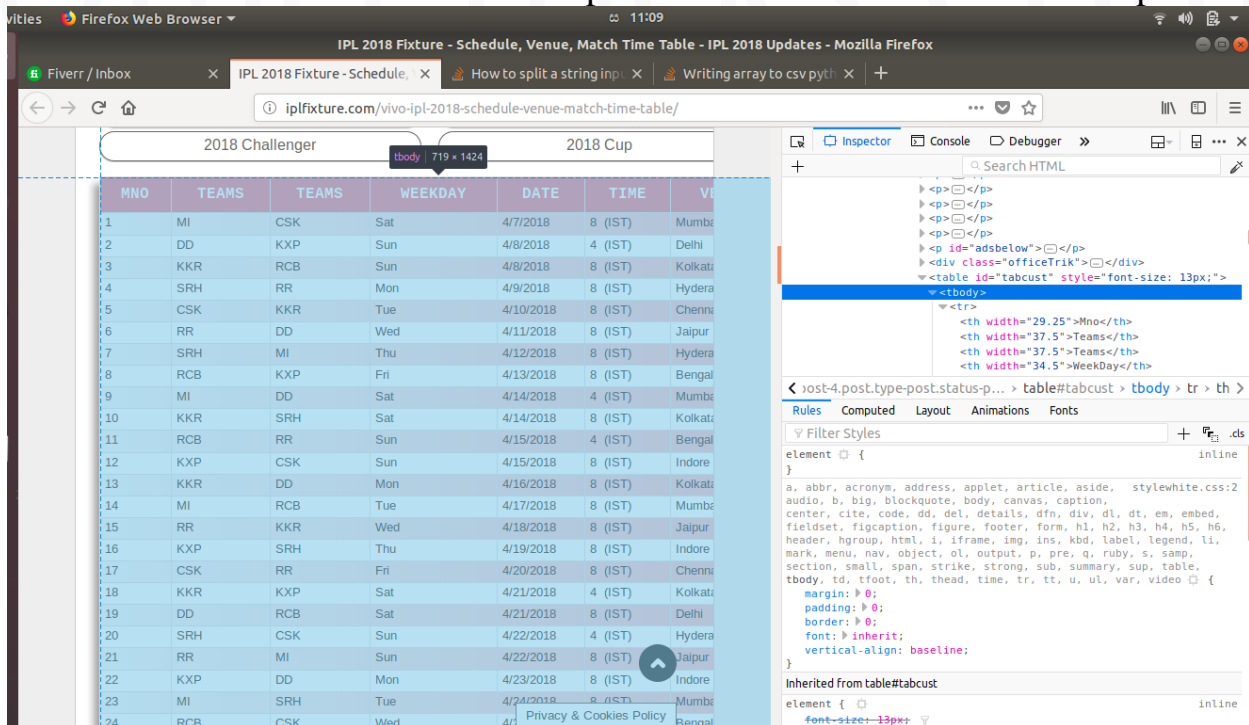
```
1 import requests #Python Library for send HTTP requests
2
3 from bs4 import BeautifulSoup #Python Library for handling parsed content
4
5 import csv #csv file handling library for python
6
7
8
9
10
11
12 link = 'http://iplfixture.com/vivo-ipl-2018-schedule-venue-match-time-table/'
13
14
15 def connection(link): #method to send HTTP request and get contents.
16     page = requests.get(link)
17     soup = BeautifulSoup(page.content, 'Lxml')
18
19     return soup
20
21
22 table = connection(link).find('table',{'id':'tabcust'}) #called the method
23
24 tr = table.find('tbody').find_all('tr')
25
26
27
28
29
30
31
32 for t in tr:
33     td = (str(t.text).replace('\n', ' ')).split()
34     print td
35     # We have the table's each row now.
36
37     # calling the writer function to write the row to csv file.
```

The IDE I'm using. **PyCharm Community**. If you're hoping to carry out more python developing, I suggest you install this.

04. Now it's time to have basic knowledge about inside of a website. As we know, a website is made of elements, which were defined with a language called HTML. In this website's elements, There is the one we need to scrape. You can view the elements of a website by pressing F12.



Now we have to find the element that represent the table that we need to scrape.



You can see the <tbody> tag surrounds the whole table we need. Now, let's go get it!

Here is the code snippet for the program, Let's get the code explained.

```
import requests #Python Library for send HTTP requests

from bs4 import BeautifulSoup #Python library for handling parsed content

import csv #csv file handling library for python

link = 'http://iplfixture.com/vivo-ipl-2018-schedule-venue-match-time-table/'

def connection(link): #method to send HTTP request and get contents.
    page = requests.get(link)
    soup = BeautifulSoup(page.content, 'xml')
    return soup

table = connection(link).find('table', {'id': 'tabcust'}) #called the method

tr = table.find('tbody').find_all('tr')

#Using a for loop for get each <td> tag within the <tr> tag.

for t in tr:
    td = (str(t.text).replace('\n', ' ')).split()
    print td
    # We have the table's each row now.
    #Now we have to get each <td> tag out separately to append to our excel table.

    #Writing data to the csv file.
    resultFile = open("out.csv", 'ab')
    wr = csv.writer(resultFile, dialect='excel')
    wr.writerow(td)
    resultFile.close()
```