Beautiful Soup - Installation

As BeautifulSoup is not a standard python library, we need to install it first. We are going to install the BeautifulSoup 4 library (also known as BS4), which is the latest one.

To isolate our working environment so as not to disturb the existing setup, let us first create a virtual environment.

## **Creating a virtual environment (optional)**

A virtual environment allows us to create an isolated working copy of python for a specific project without affecting the outside setup.

Best way to install any python package machine is using pip, however, if pip is not installed already (you can check it using – “pip –version” in your command or shell prompt), you can install by giving below command −

### **Linux environment**

$sudo apt-get install python-pip

### **Windows environment**

To install pip in windows, do the following −

* Download the get-pip.py from <https://bootstrap.pypa.io/get-pip.py> or from the github to your computer.
* Open the command prompt and navigate to the folder containing get-pip.py file.
* Run the following command −

>python get-pip.py

That’s it, pip is now installed in your windows machine.

You can verify your pip installed by running below command −

>pip --version

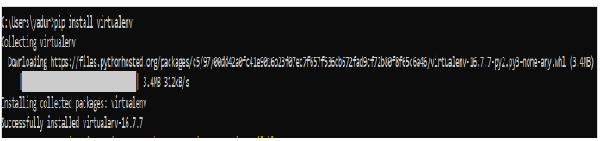
pip 19.2.3 from c:\users\yadur\appdata\local\programs\python\python37\lib\site-packages\pip (python 3.7)

## **Installing virtual environment**

Run the below command in your command prompt −

>pip install virtualenv

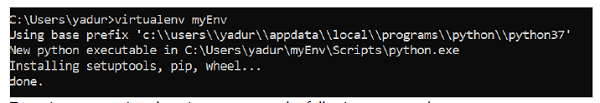
After running, you will see the below screenshot −



Below command will create a virtual environment (“myEnv”) in your current directory −

>virtualenv myEnv

### **Screenshot**



To activate your virtual environment, run the following command −

>myEnv\Scripts\activate



In the above screenshot, you can see we have “myEnv” as prefix which tells us that we are under virtual environment “myEnv”.

To come out of virtual environment, run deactivate.

(myEnv) C:\Users\yadur>deactivate

C:\Users\yadur>

As our virtual environment is ready, now let us install beautifulsoup.

## **Installing BeautifulSoup**

As BeautifulSoup is not a standard library, we need to install it. We are going to use the BeautifulSoup 4 package (known as bs4).

### **Linux Machine**

To install bs4 on Debian or Ubuntu linux using system package manager, run the below command −

$sudo apt-get install python-bs4 (for python 2.x)

$sudo apt-get install python3-bs4 (for python 3.x)

You can install bs4 using easy\_install or pip (in case you find problem in installing using system packager).

$easy\_install beautifulsoup4

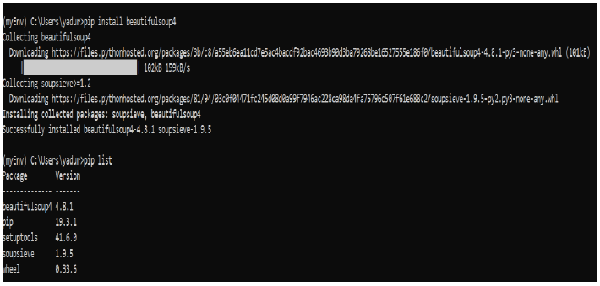
$pip install beautifulsoup4

(You may need to use easy\_install3 or pip3 respectively if you’re using python3)

### **Windows Machine**

To install beautifulsoup4 in windows is very simple, especially if you have pip already installed.

>pip install beautifulsoup4



So now beautifulsoup4 is installed in our machine. Let us talk about some problems encountered after installation.

## **Problems after installation**

On windows machine you might encounter, wrong version being installed error mainly through −

* error: ***ImportError “No module named HTMLParser”***, then you must be running python 2 version of the code under Python 3.
* error: ***ImportError “No module named html.parser”***error, then you must be running Python 3 version of the code under Python 2.

Best way to get out of above two situations is to re-install the BeautifulSoup again, completely removing existing installation.

If you get the ***SyntaxError “Invalid syntax”*** on the line ROOT\_TAG\_NAME = u’[document]’, then you need to convert the python 2 code to python 3, just by either installing the package −

$ python3 setup.py install

or by manually running python’s 2 to 3 conversion script on the bs4 directory −

$ 2to3-3.2 -w bs4

## **Installing a Parser**

By default, Beautiful Soup supports the HTML parser included in Python’s standard library, however it also supports many external third party python parsers like lxml parser or html5lib parser.

To install lxml or html5lib parser, use the command −

### **Linux Machine**

$apt-get install python-lxml

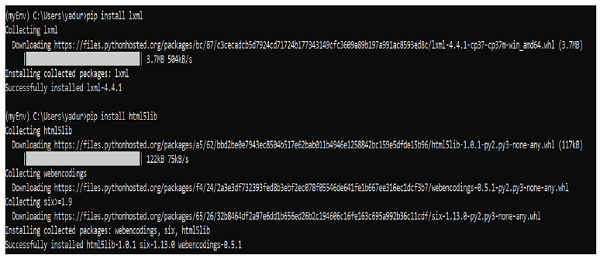
$apt-get insall python-html5lib

### **Windows Machine**

$pip install lxml

$pip install html5lib

$pip install requests



Generally, users use lxml for speed and it is recommended to use lxml or html5lib parser if you are using older version of python 2 (before 2.7.3 version) or python 3 (before 3.2.2) as python’s built-in HTML parser is not very good in handling older version.

## **Running Beautiful Soup**

It is time to test our Beautiful Soup package in one of the html pages (taking web page – <https://www.tutorialspoint.com/index.htm>, you can choose any-other web page you want) and extract some information from it.

In the below code, we are trying to extract the title from the webpage −

from bs4 import BeautifulSoup

import requests

url = "https://www.tutorialspoint.com/index.htm"

req = requests.get(url)

soup = BeautifulSoup(req.text, "html.parser")

print(soup.title)