Python Scraping Exercise

Exercise 2 – Web Scraping

Student Name: **Student Id:**

Date:

Please use the screenshots ONLY as a reference. The written instructions have to be followed AS written.

Objective:

Web Scrapping is a technique of extracting information from websites using computer software or applications.

The objective of this exercise is to develop skills for acquiring data using a Python technique.

Step 1: Prerequisites

Before you can begin, your computer needs the following tools installed and working.

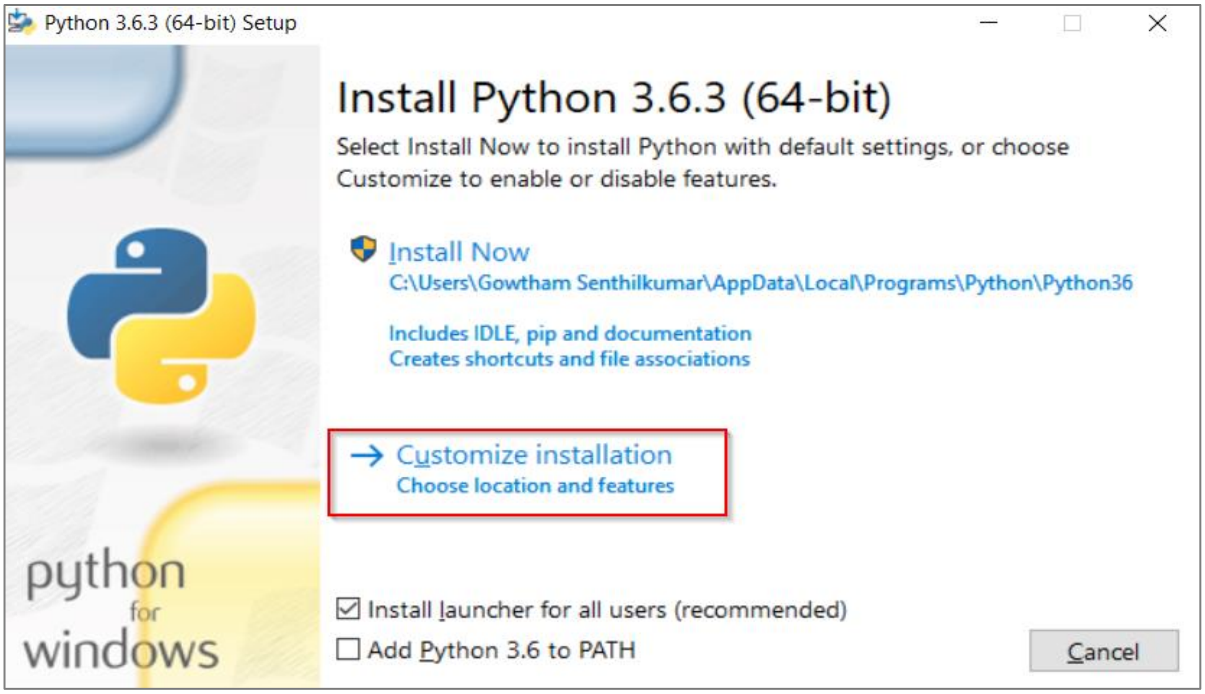
a. A command-line interface to interact with your computer
   On Windows you can find the command-line interface by opening the "command prompt".

b. A text editor to work with plain text files
   - For Windows, it is recommended to install Notepad++.

c. Python programming language version 3.6.3-amd64
   - If you don't have Python installed try downloading and installing it from the link below:
     https://www.python.org/downloads/release/python-363/
     Please install the **Windows x86-64 executable installer**.

Step 2: Installation of Python 3.6.3

- Select "Customize Installation"



Python 3.6.3 (64-bit) Setup

Install Python 3.6.3 (64-bit)

Select Install Now to install Python with default settings, or choose Customize to enable or disable features.

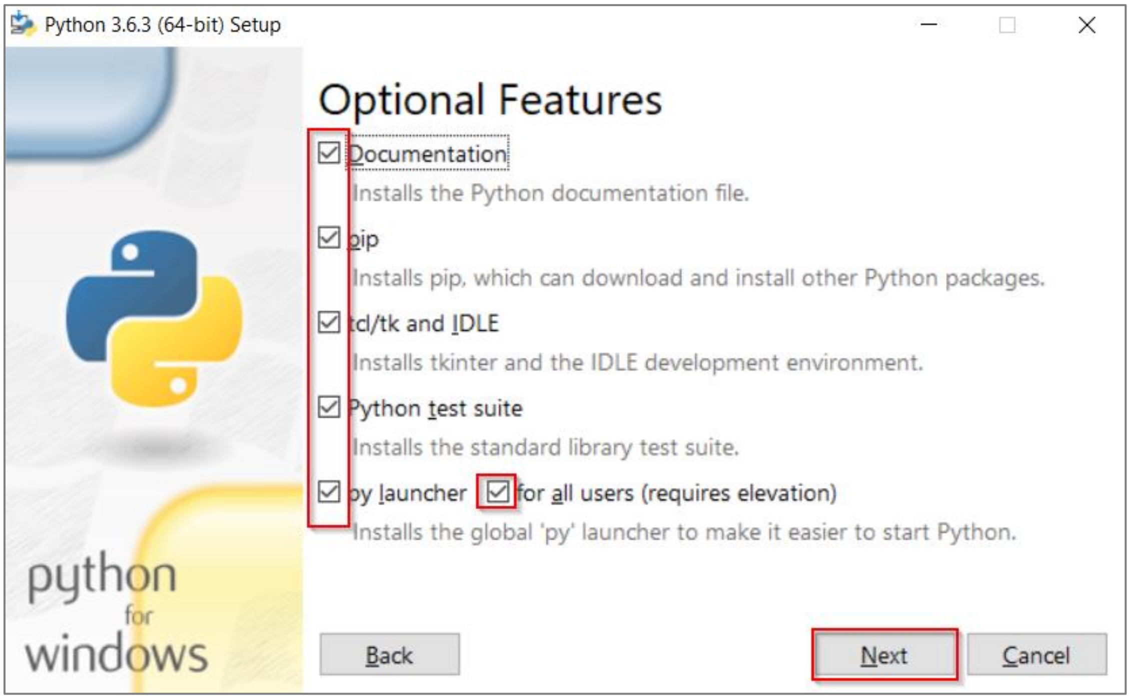
Install Now
C:\Users\Gowtham Senthilkumar\AppData\Local\Programs\Python\Python36

Includes IDLE, pip and documentation
Creates shortcuts and file associations

→ Customize installation
Choose location and features

☑ Install launcher for all users (recommended)
□ Add Python 3.6 to PATH

- Ensure that all the options are checked as shown in the screenshot, after which click on Next.



Python 3.6.3 (64-bit) Setup

Optional Features

☑ Documentation
Installs the Python documentation file.

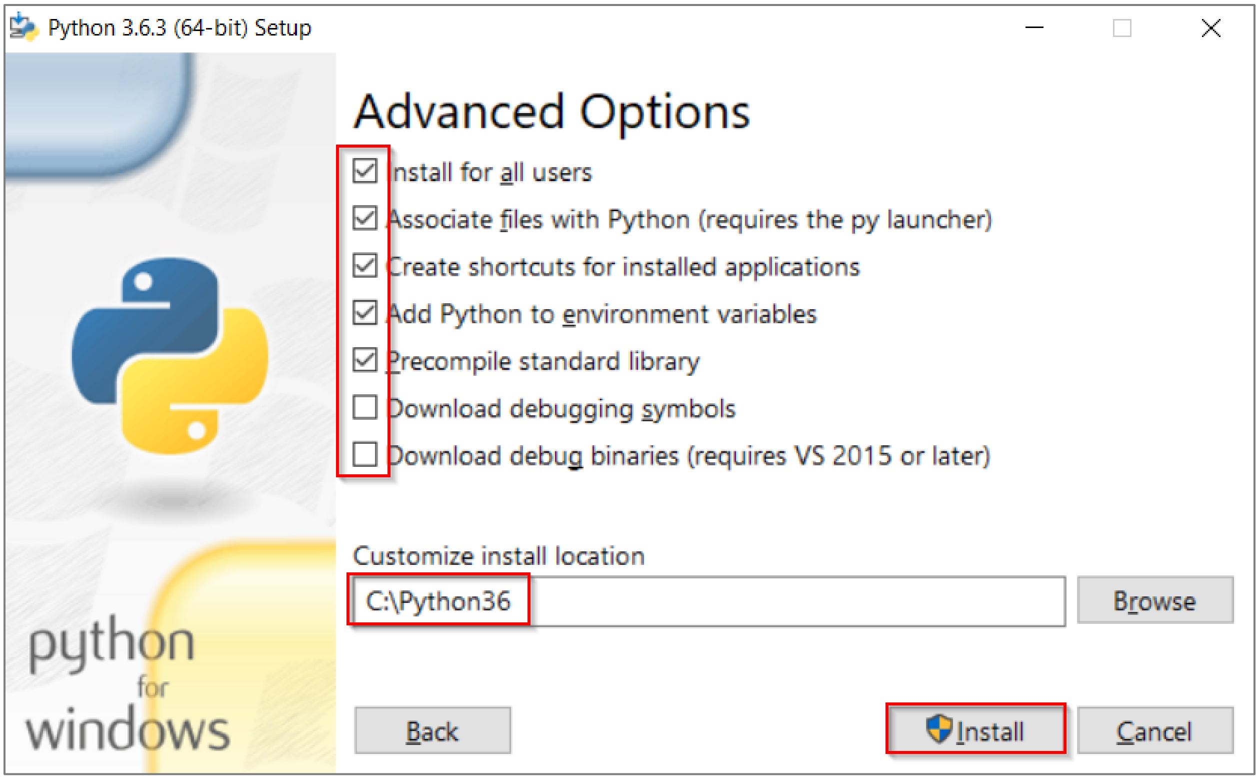
☑ pip
Installs pip, which can download and install other Python packages.

☑ tcl/tk and IDLE
Installs tkinter and the IDLE development environment.

☑ Python test suite
Installs the standard library test suite.

☑ py launcher ☑ for all users (requires elevation)
Installs the global 'py' launcher to make it easier to start Python.

- Ensure that all the options are checked as shown in the screenshot, change the custom install location to the one specified below. It is case sensitive. After which, click on Install.



Python 3.6.3 (64-bit) Setup

Advanced Options

☑ Install for all users

☑ Associate files with Python (requires the py launcher)

☑ Create shortcuts for installed applications

☑ Add Python to environment variables

☑ Precompile standard library

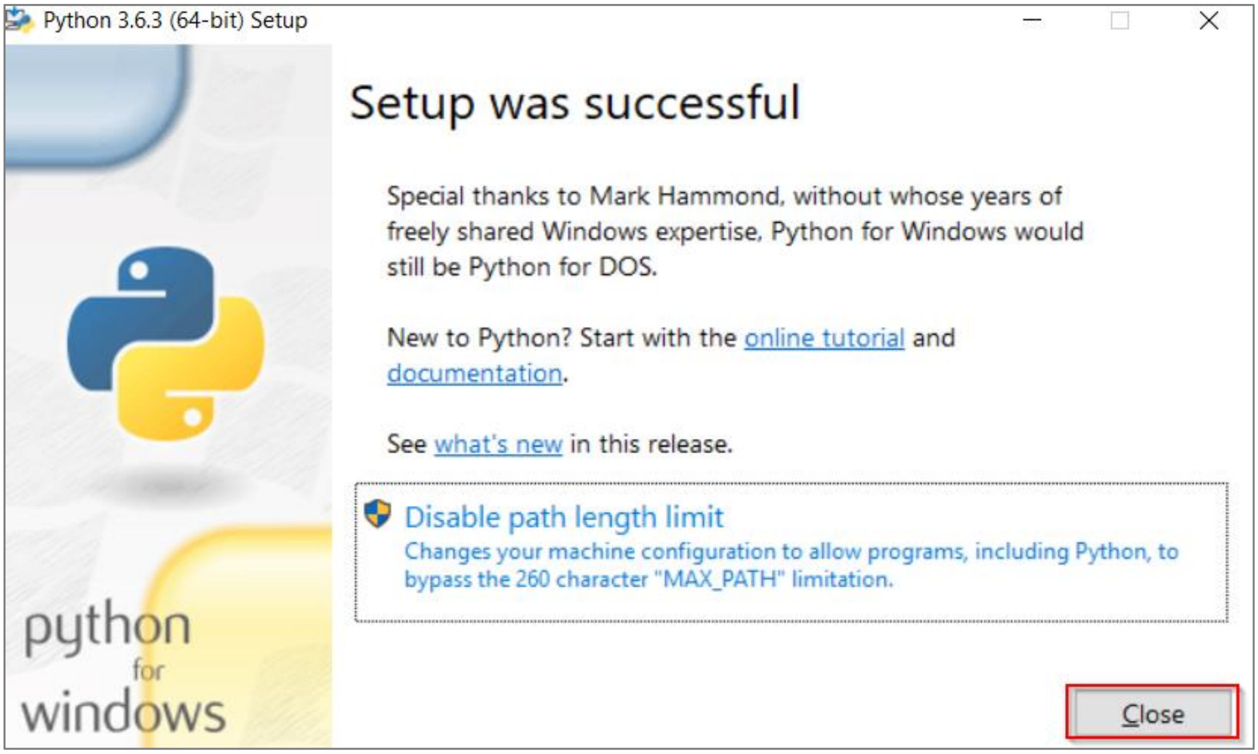
□ Download debugging symbols

□ Download debug binaries (requires VS 2015 or later)

Customize install location

C:\Python36

- If prompted for Administrator permission for Installation, click on “Yes”
- Click on close



Python 3.6.3 (64-bit) Setup

Setup was successful

Special thanks to Mark Hammond, without whose years of freely shared Windows expertise, Python for Windows would still be Python for DOS.

New to Python? Start with the online tutorial and documentation.

See what's new in this release.

Disable path length limit

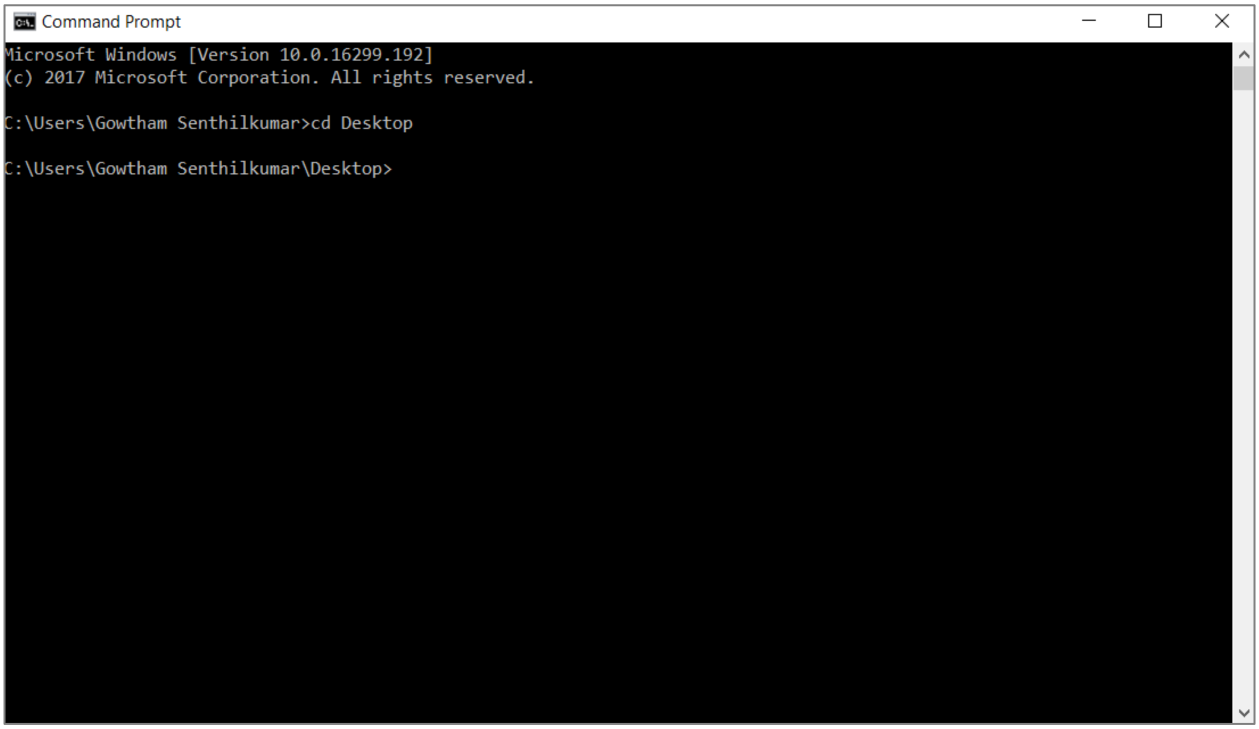
Changes your machine configuration to allow programs, including Python, to bypass the 260 character "MAX_PATH" limitation.

Step 3: How to use the command line

Open the command-line program for your operating system and let’s get started.

For Windows, you can click on Start button and type **cmd** in your search buttons

Command prompt window will look like the screenshot below:



```
Microsoft Windows [Version 10.0.16299.192]
(c) 2017 Microsoft Corporation. All rights reserved.

C:\Users\Gowtham Senthilkumar>cd Desktop

C:\Users\Gowtham Senthilkumar\Desktop>
```

- For Windows use command:
  cd (to change folder)

The command prompt should print out your current location relative to the root of your computer’s file system. In this case, you’re probably in the default directory for your user, also known as your home directory. It’s easy to lose track of which folder you’re in when you’re working from the command line, so this is a helpful tool for finding your way.

Change directories:
Now let’s move. In order to change directories from the command line, we’ll return to the cd command we saw earlier, which works for OSX, Linux and Windows.

The only thing you need to do is tell it which directory to move into. In this case, the following will probably drop you on your desktop.

cd Desktop

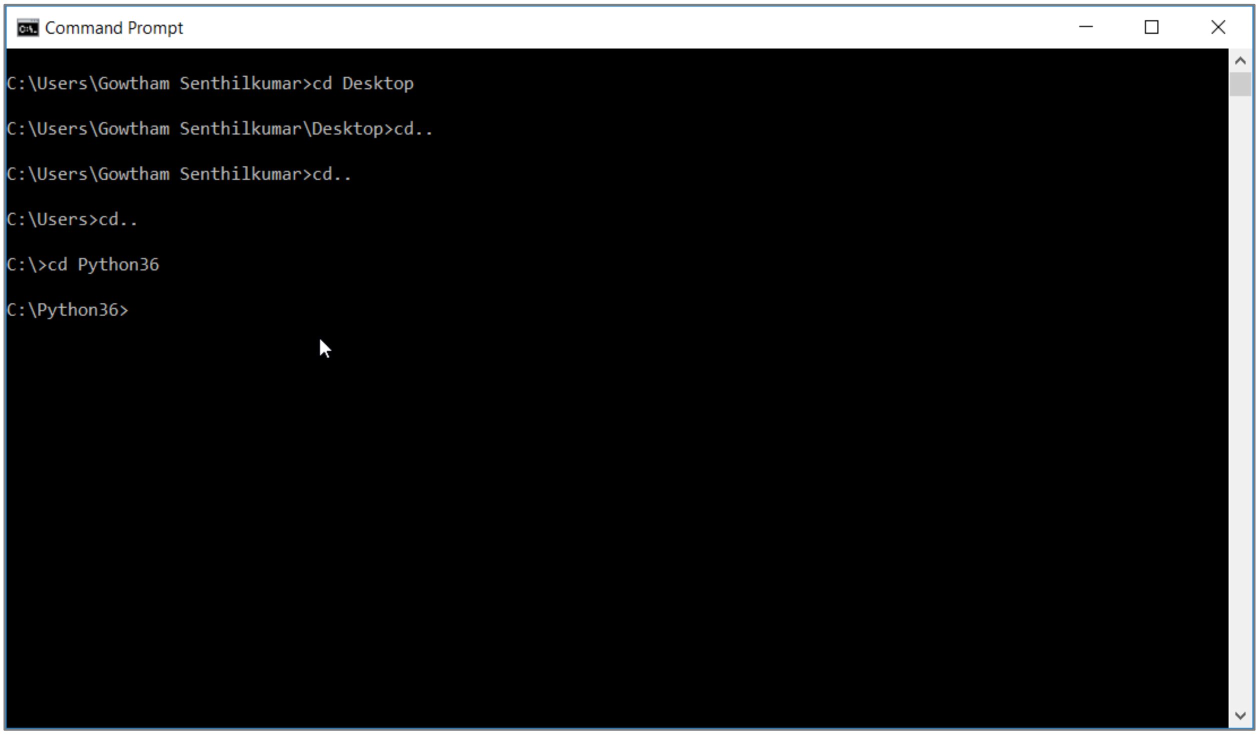
Step 3: Web Scraping using Python

a. Installing dependencies

The scraper will use Python’s BeautifulSoup toolkit to parse the site’s HTML and extract the data. We’ll also use the Requests library to open the URL, download the HTML and pass it to BeautifulSoup.

- Make sure to check the directory in which Python is installed. Generally, Python is installed in “C:\Python36”. If not, then please navigate to the correct directory for Python36.
- In order to go to the above directory, follow the commands given below on the command line-
  - Command (two times) - cd..
  - This command will bring you to the C drive.
  - To point to the C:/Python36, follow the command-

    cd Python36



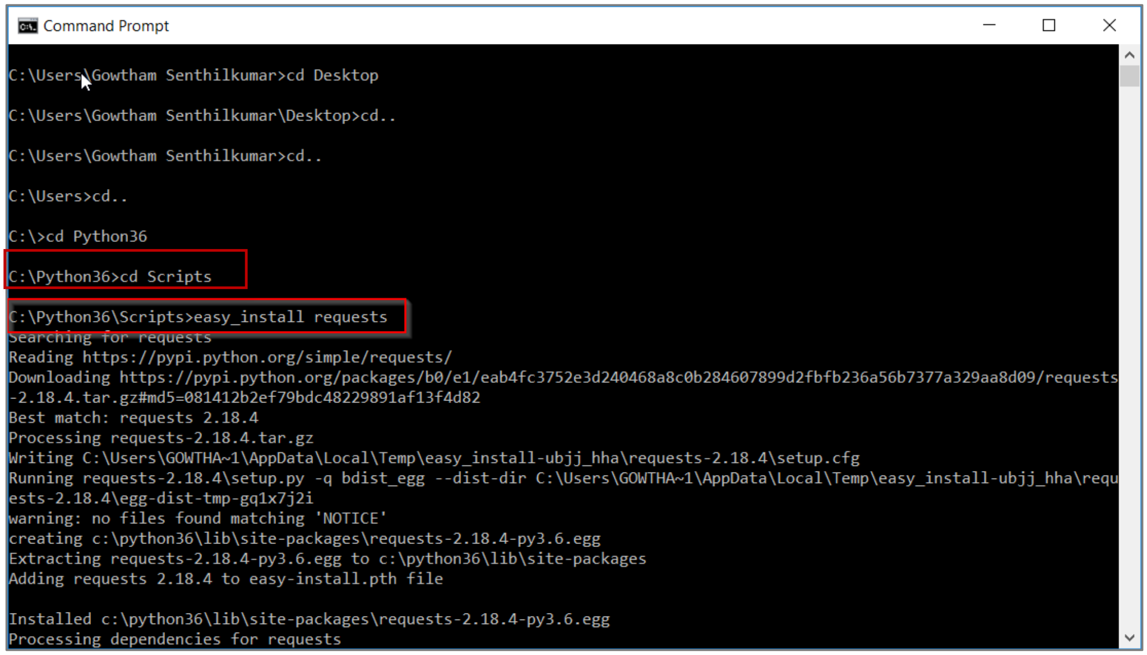
```
C:\Users\Gowtham Senthilkumar>cd Desktop

C:\Users\Gowtham Senthilkumar\Desktop>cd..

C:\Users\Gowtham Senthilkumar>cd..

C:\Users>cd..

C:\>cd Python36

C:\Python36>
```

- To install easy_install requests, move to the Scripts folder in Python36.

  The following commands will help to install the easy_install Python module

cd Scripts

easy_install requests



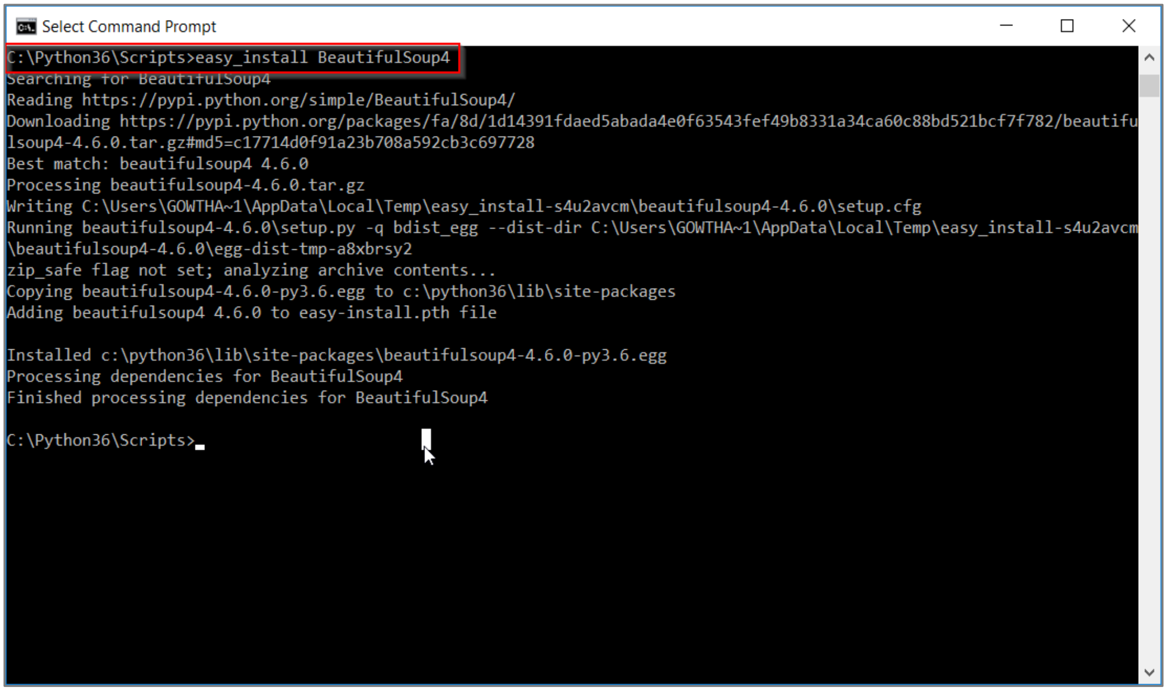
```
C:\Users\Gowtham Senthilkumar>cd Desktop

C:\Users\Gowtham Senthilkumar\Desktop>cd..

C:\Users\Gowtham Senthilkumar>cd..

C:\Users>cd..

C:\>cd Python36

C:\Python36>cd Scripts

C:\Python36\Scripts>easy_install requests
Searching for requests
Reading https://pypi.python.org/simple/requests/
Downloading https://pypi.python.org/packages/b0/e1/eab4fc3752e3d240468a8c0b284607899d2fbfb236a56b7377a329aa8d09/requests
-2.18.4.tar.gz#md5=081412b2ef79bdc48229891af13f4d82
Best match: requests 2.18.4
Processing requests-2.18.4.tar.gz
Writing C:\Users\GOWTHA~1\AppData\Local\Temp\easy_install-ubjj_hha\requests-2.18.4\setup.cfg
Running requests-2.18.4\setup.py -q bdist_egg --dist-dir C:\Users\GOWTHA~1\AppData\Local\Temp\easy_install-ubjj_hha\requ
ests-2.18.4\egg-dist-tmp-gq1x7j2i
warning: no files found matching 'NOTICE'
creating c:\python36\lib\site-packages\requests-2.18.4-py3.6.egg
Extracting requests-2.18.4-py3.6.egg to c:\python36\lib\site-packages
Adding requests 2.18.4 to easy-install.pth file

Installed c:\python36\lib\site-packages\requests-2.18.4-py3.6.egg
Processing dependencies for requests
```

easy_install is a python module bundled with setuptools that lets you automatically

download, build, install, and manage Python packages.

- Enter this command to install BeautifulSoup4 and press enter

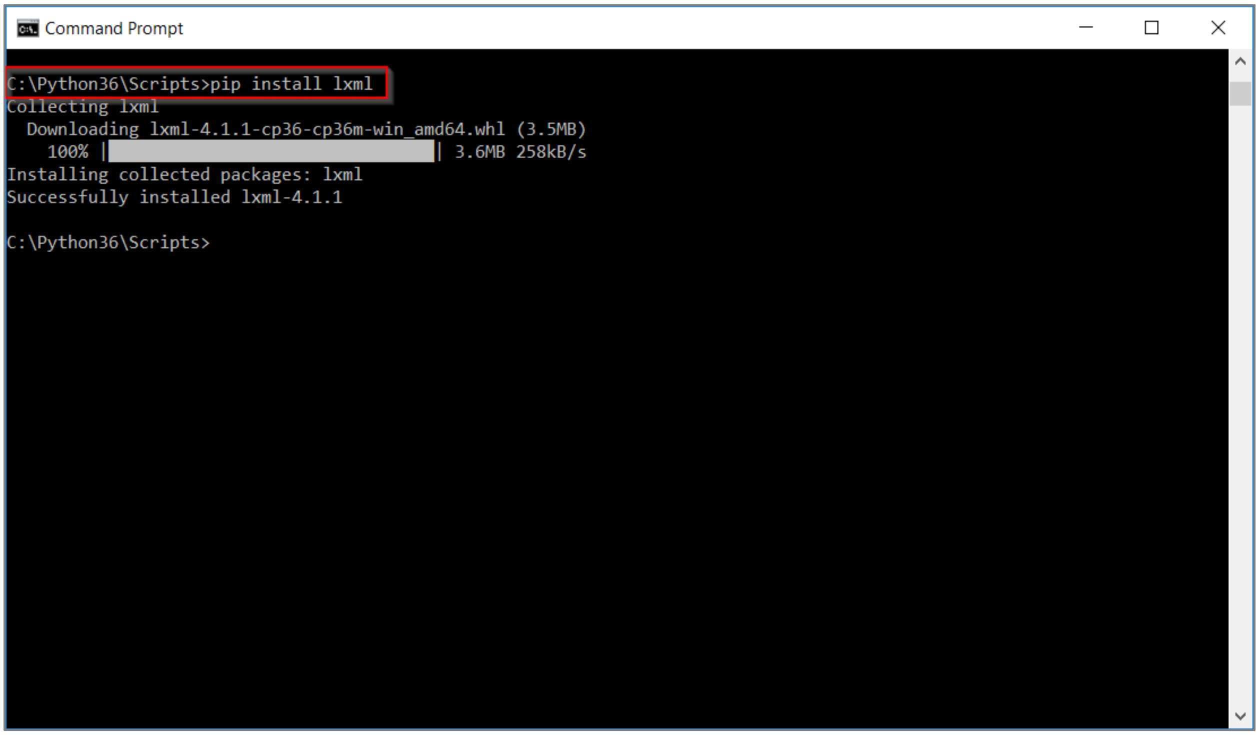
easy_install BeautifulSoup4



```
C:\Python36\Scripts>easy_install BeautifulSoup4
Searching for BeautifulSoup4
Reading https://pypi.python.org/simple/BeautifulSoup4/
Downloading https://pypi.python.org/packages/fa/8d/1d14391fdaed5abada4e0f63543fef49b8331a34ca60c88bd521bcf7f782/beautifu
lsoup4-4.6.0.tar.gz#md5=c17714d0f91a23b708a592cb3c697728
Best match: beautifulsoup4 4.6.0
Processing beautifulsoup4-4.6.0.tar.gz
Writing C:\Users\GOWTHA~1\AppData\Local\Temp\easy_install-s4u2avcm\beautifulsoup4-4.6.0\setup.cfg
Running beautifulsoup4-4.6.0\setup.py -q bdist_egg --dist-dir C:\Users\GOWTHA~1\AppData\Local\Temp\easy_install-s4u2avcm
\beautifulsoup4-4.6.0\egg-dist-tmp-a8xbrsy2
zip_safe flag not set; analyzing archive contents...
Copying beautifulsoup4-4.6.0-py3.6.egg to c:\python36\lib\site-packages
Adding beautifulsoup4 4.6.0 to easy-install.pth file

Installed c:\python36\lib\site-packages\beautifulsoup4-4.6.0-py3.6.egg
Processing dependencies for BeautifulSoup4
Finished processing dependencies for BeautifulSoup4

C:\Python36\Scripts>
```

- Enter this command to install lxml library and press enter

  pip install lxml



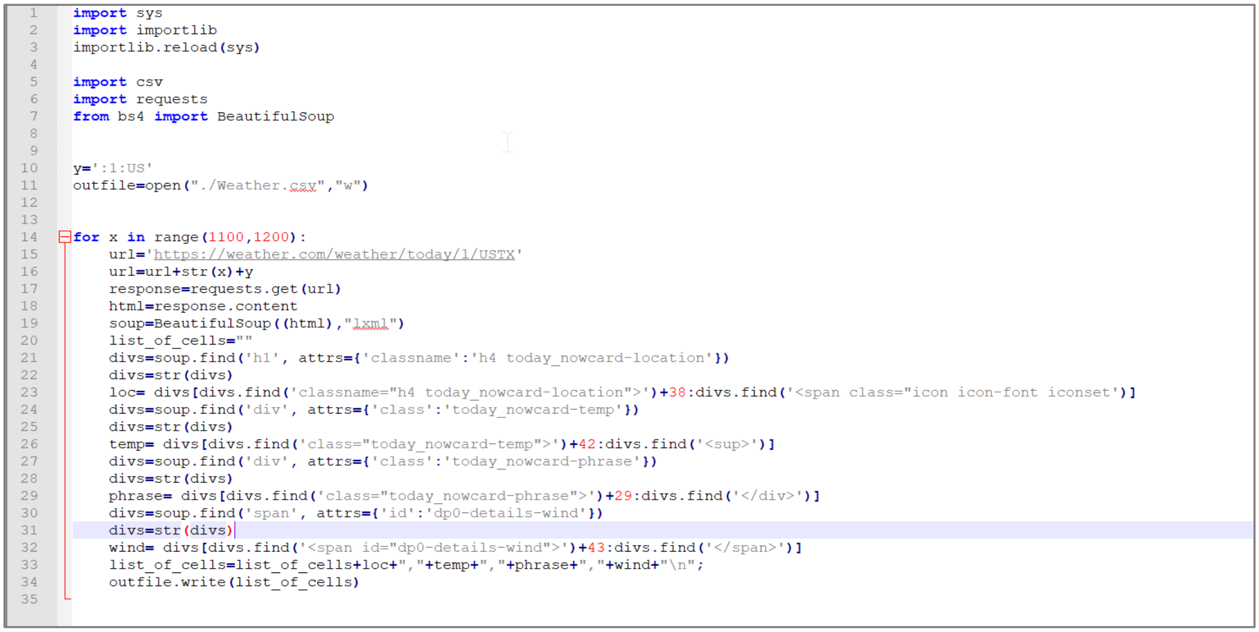
```
C:\Python36\Scripts>pip install lxml
Collecting lxml
  Downloading lxml-4.1.1-cp36-cp36m-win_amd64.whl (3.5MB)
    100% |████████████████████████████████| 3.6MB 258kB/s
Installing collected packages: lxml
Successfully installed lxml-4.1.1

C:\Python36\Scripts>
```

Our libraries are now installed and it is time to start writing our data scraping code.

Step 4: Extracting the entire list of Rankings table

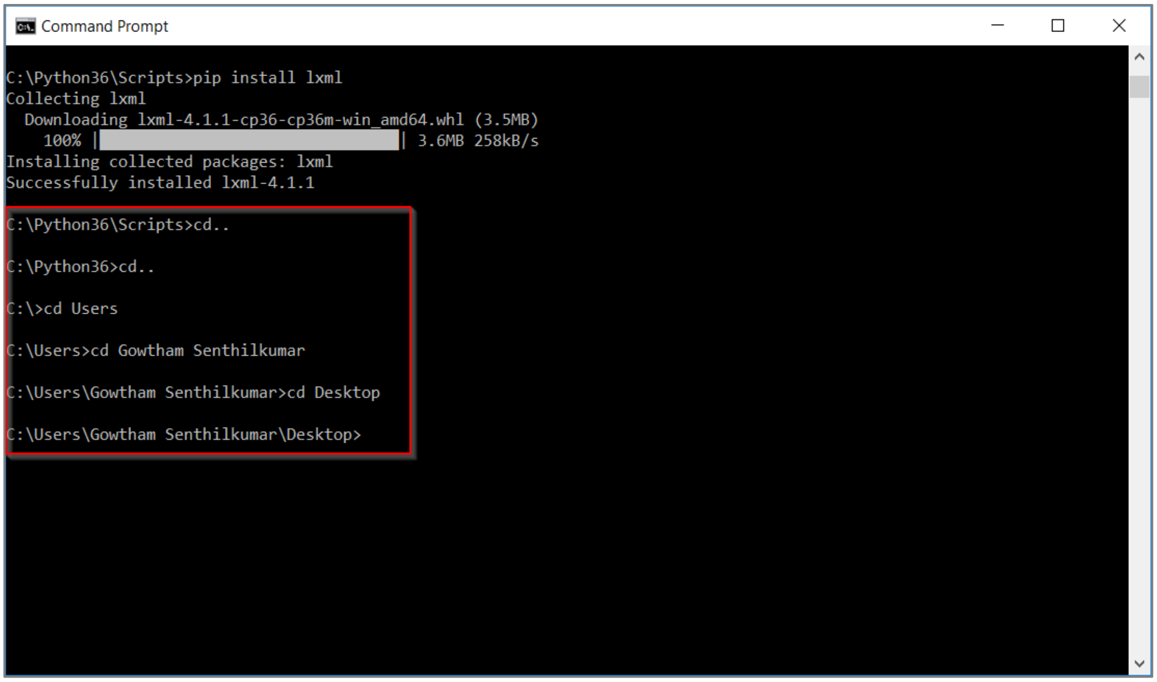
a. Go to the URL - https://weather.com/weather/today/l/USTX1110:1:US . Here you can view the entire list of Rankings in a Table. (Look at the definitions of the table headings at the end of this document.)
   So our next objective is to scrape the entire table.

b. Open your text editor (Notepad++) and type the code given in the below screen shot.



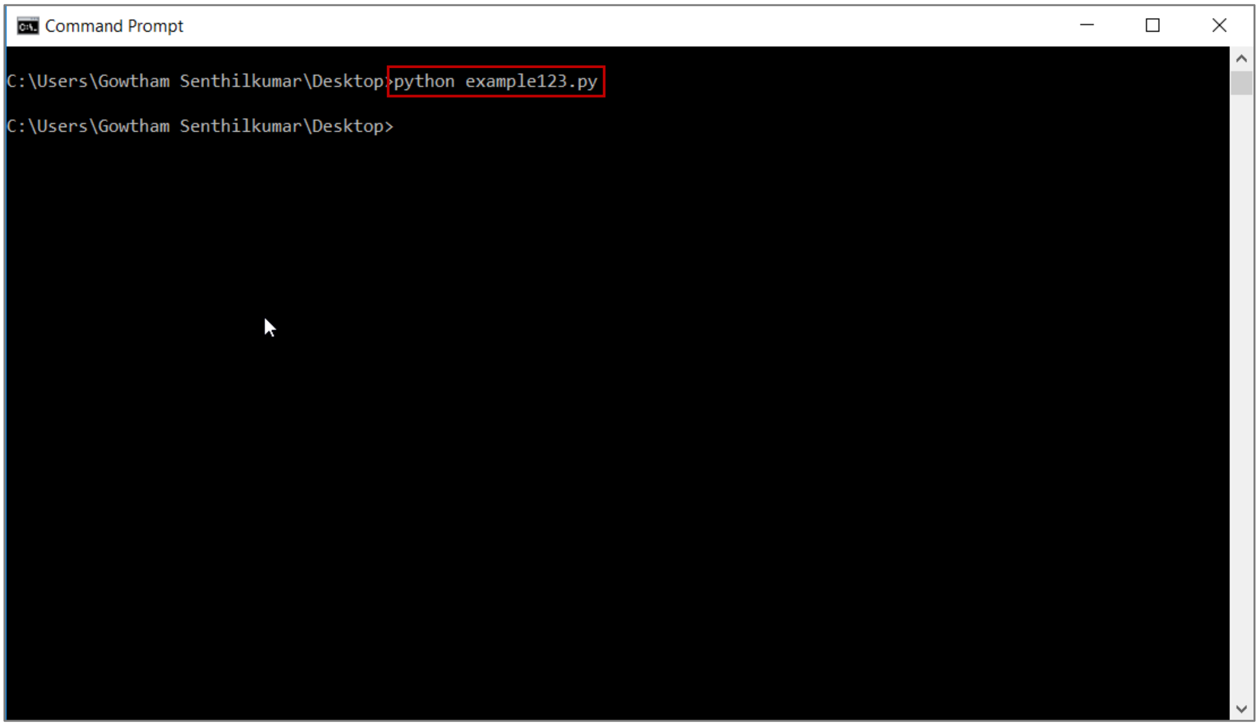
```
import sys
import importlib
importlib.reload(sys)

import csv
import requests
from bs4 import BeautifulSoup


y=':1:US'
outfile=open("./Weather.csv","w")


for x in range(1100,1200):
    url='https://weather.com/weather/today/l/USTX'
    url=url+str(x)+y
    response=requests.get(url)
    html=response.content
    soup=BeautifulSoup((html),"lxml")
    list_of_cells=""
    divs=soup.find('h1', attrs={'classname':'h4 today_nowcard-location'})
    divs=str(divs)
    loc= divs[divs.find('classname="h4 today_nowcard-location">')+38:divs.find('<span class="icon icon-font iconset')]
    divs=soup.find('div', attrs={'class':'today_nowcard-temp'})
    divs=str(divs)
    temp= divs[divs.find('class="today_nowcard-temp">')+42:divs.find('<sup>')]
    divs=soup.find('div', attrs={'class':'today_nowcard-phrase'})
    divs=str(divs)
    phrase= divs[divs.find('class="today_nowcard-phrase">')+29:divs.find('</div>')]
    divs=soup.find('span', attrs={'id':'dp0-details-wind'})
    divs=str(divs)
    wind= divs[divs.find('<span id="dp0-details-wind">')+43:divs.find('</span>')]
    list_of_cells=list_of_cells+loc+","+temp+","+phrase+","+wind+"\n";
    outfile.write(list_of_cells)

```

Save the file with name ‘example123.py’ on the Desktop and navigate your way to the Desktop on Command Prompt.



```
C:\Python36\Scripts>pip install lxml
Collecting lxml
  Downloading lxml-4.1.1-cp36-cp36m-win_amd64.whl (3.5MB)
    100% |████████████████████████████████| 3.6MB 258kB/s
Installing collected packages: lxml
Successfully installed lxml-4.1.1

C:\Python36\Scripts>cd..

C:\Python36>cd..

C:\>cd Users

C:\Users>cd Gowtham Senthilkumar

C:\Users\Gowtham Senthilkumar>cd Desktop

C:\Users\Gowtham Senthilkumar\Desktop>
```

Type the command below to run the python file.

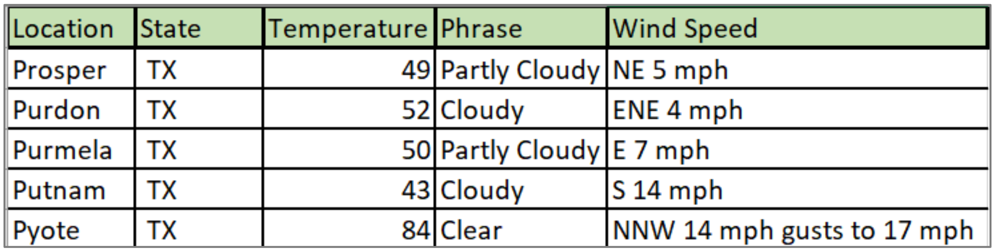


```
C:\Users\Gowtham Senthilkumar\Desktop>python example123.py

C:\Users\Gowtham Senthilkumar\Desktop>
```

Since you had listed the outfile name as Ranking.csv in the program code, after you execute the above command, a new csv file will be created on the Desktop.

Question: Paste the screenhsot of the CSV file

Please format the csv file as shown below and save file as .xlsx.



Location	State	Temperature	Phrase	Wind Speed
Prosper	TX	49	Partly Cloudy	NE 5 mph
Purdon	TX	52	Cloudy	ENE 4 mph
Purmela	TX	50	Partly Cloudy	E 7 mph
Putnam	TX	43	Cloudy	S 14 mph
Pyote	TX	84	Clear	NNW 14 mph gusts to 17 mph

Questions:

Provide complete screen uncropped screenshots for the questions below.

1. What is the windspeed in Quitaque?

2. What is the average temperature in Texas?

3. List the top 5 Locations with highest temperatures.

4. Plot a line chart with two data series:
   - Average temperature
   - Temperatures of the locations mentioned in the Excel file.

. **Note**: Title of the graph should be your name and don't forget to add AXIS titles.

Instructions:

1. Submit the assignment document in Microsoft word
2. Submit excel file on eLearning
3. Submit .py file created