



CHRIST
(DEEMED TO BE UNIVERSITY)
BANGALORE · INDIA

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

COMPUTER SOCIETY OF INDIA

PRESENTS

WORKSHOP ON

WEB SCRAPING

Date & Time: 4th February 2020 (4.30 - 6.00 PM)



Coordinator : Dr. Jayapandian N
njayapandian@gmail.com

Venue: Room No. 218,
First floor, second block,
Kengeri Campus

SCAN TO REGISTER

All the Participants get E-Certificate

Vision: To Fortify Ethical Computational Excellence





CHRIST
(DEEMED TO BE UNIVERSITY)
B A N G A L O R E · I N D I A

**Department of Computer Science and Engineering
Computer Society of India**

**Event Report
Workshop
On
WEB SCRAPING**

CONDUCTED BY : COMPUTER SOCIETY OF INDIA
DATE & TIME : 4TH FEB, 2020. 4:30-6:00 PM
VENUE : #218, 2ND BLOCK, CHRIST - KENGERI
CAMPUS
RESOURCE PERSONS: Puneeth C
EVENT COORDINATOR: Dr JAYAPANDIAN N.
Total No. of Participants: 26

Web Scraping (also termed Screen Scraping, Web Data Extraction, Web Harvesting etc.) is a technique employed to extract large amounts of data from websites whereby the data is extracted and saved to a local file in your computer or to a database in table (spreadsheet) format.

Data displayed by most websites can only be viewed using a web browser. They do not offer the functionality to save a copy of this data for personal use. The only option then is to manually copy and paste the data - a very tedious job which can take many hours or sometimes days to complete. Web Scraping is the technique of automating this process, so that instead of

manually copying the data from websites, the Web Scraping software will perform the same task within a fraction of the time.

A web scraping software will automatically load and extract data from multiple pages of websites based on your requirement. It is either custom built for a specific website or is one which can be configured to work with any website. With the click of a button you can easily save the data available in the website to a file in your computer.

The problem with most generic web scraping software is that they are very difficult to setup and use. There is a steep learning curve involved. WebHarvy was designed to solve this problem. With a very intuitive, point and click interface, using WebHarvy you can start extracting data within minutes from any website.

Please watch the following demonstration which shows how easy it is to configure and use WebHarvy for your data extraction needs.

In this section, you will learn

- about how to store scraped data in databases
- how to process HTML documents and HTTP requests
- techniques for selecting and extracting data from websites
- about writing web spiders that crawl and scrape large portions of the web

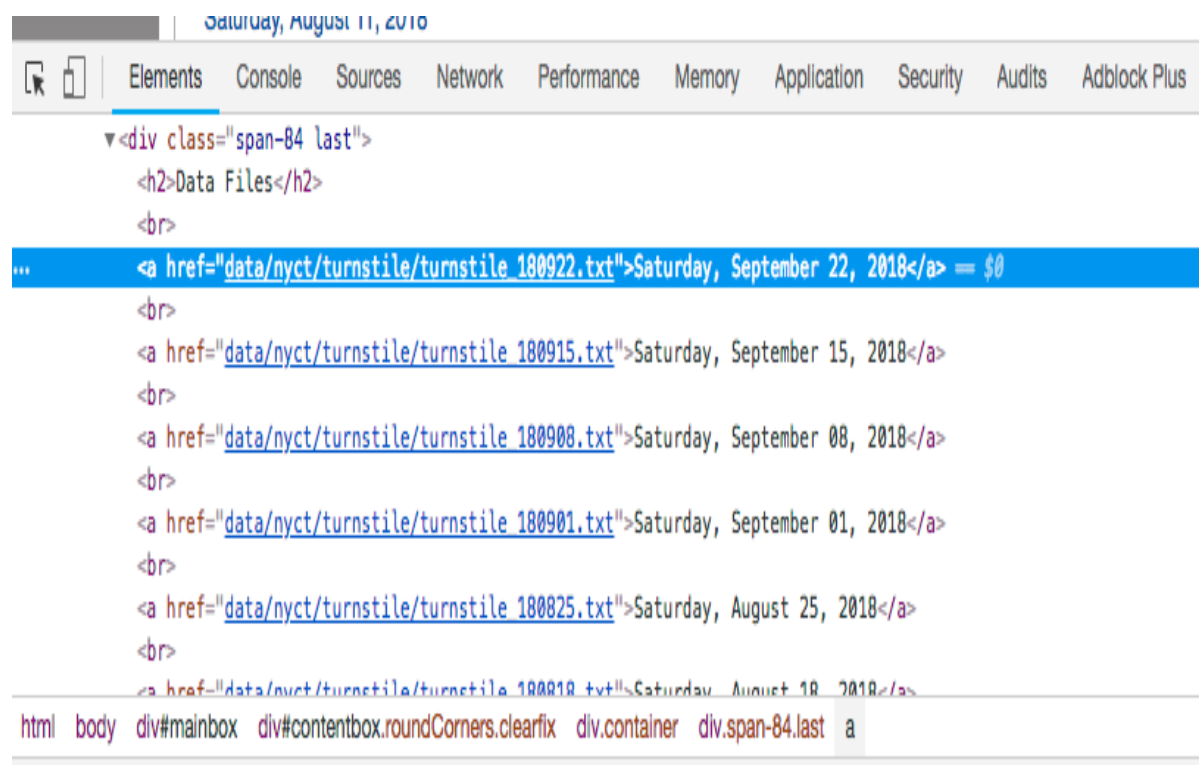
Web scraping is a technique to automatically access and extract large amounts of information from a website, which can save a huge amount of time and effort. In this article, we will go through an easy example of how to automate downloading hundreds of files from the New York MTA. This is a great exercise for web scraping beginners who are looking to understand how to

web scrape. Web scraping can be slightly intimidating, so this tutorial will break down the process of how to go about the process.

Important notes about web scraping:

Read through the website's Terms and Conditions to understand how you can legally use the data. Most sites prohibit you from using the data for commercial purposes.

Make sure you are not downloading data at too rapid a rate because this may break the website. You may potentially be blocked from the site as well.



```

Saturday, August 11, 2018
Elements Console Sources Network Performance Memory Application Security Audits Adblock Plus
▼ <div class="span-84 last">
  <h2>Data Files</h2>
  <br>
  ... <a href="data/nyct/turnstile/turnstile_180922.txt">Saturday, September 22, 2018</a> = $0
  <br>
  <a href="data/nyct/turnstile/turnstile_180915.txt">Saturday, September 15, 2018</a>
  <br>
  <a href="data/nyct/turnstile/turnstile_180908.txt">Saturday, September 08, 2018</a>
  <br>
  <a href="data/nyct/turnstile/turnstile_180901.txt">Saturday, September 01, 2018</a>
  <br>
  <a href="data/nyct/turnstile/turnstile_180825.txt">Saturday, August 25, 2018</a>
  <br>
  <a href="data/nyct/turnstile/turnstile_180818.txt">Saturday, August 18, 2018</a>
html body div#mainbox div#contentbox.roundCorners.clearfix div.container div.span-84.last a

```

If you click on this arrow and then click on an area of the site itself, the code for that particular item will be highlighted in the console. I've clicked on the very first data file, Saturday, September 22, 2018 and the console has highlighted in blue the link to that particular file.

This code saves ‘data/nyct/turnstile/turnstile_180922.txt’ to our variable link. The full url to download the data is actually ‘http://web.mta.info/developers/data/nyct/turnstile/turnstile_180922.txt’ which I discovered by clicking on the first data file on the website as a test. We can use our urllib.request library to download this file path to our computer. We provide request.urlretrieve with two parameters: file url and the filename. For my files, I named them “turnstile_180922.txt”, “turnstile_180901”, etc.

```
<a href="http://web.mta.info/accountability">Main Page</a>,
<a href="http://web.mta.info/mta/boardmaterials.html">Board Materials</a>,
<a href="http://web.mta.info/mta/budget/">Budget Info</a>,
<a href="http://web.mta.info/capital">Capital Program Info</a>,
<a href="http://web.mta.info/capitaldashboard/CPDHome.html">Capital Program Dashboard</a>,
<a href="http://web.mta.info/mta/investor/">Investor Information</a>,
<a href="http://web.mta.info/mta/leadership/">MTA Leadership</a>,
<a href="http://web.mta.info/persdashboard/performance14.html">Performance Indicators</a>,
<a href="http://www.mta.info/mta-news">Press Releases and News</a>,
<a href="http://web.mta.info/mta/news/hearings">Public Hearings</a>,
<a class="last" href="http://web.mta.info/mta/news/hearings/index-reinvention.html">Transportation Reinvention Commission</a>,
<a name="main-content"> </a>,
<a href="resources/nyct/turnstile/ts_Field_Description_pre-10-18-2014.txt">Prior to 10/18/14</a>,
<a href="resources/nyct/turnstile/ts_Field_Description.txt">Current</a>,
<a href="resources/nyct/turnstile/Remote-Booth-Station.xls">Remote Unit/Control Area/Station Name Key</a>,
<a href="data/nyct/turnstile/turnstile_180922.txt">Saturday, September 22, 2018</a>,
<a href="data/nyct/turnstile/turnstile_180915.txt">Saturday, September 15, 2018</a>,
<a href="data/nyct/turnstile/turnstile_180908.txt">Saturday, September 08, 2018</a>,
<a href="data/nyct/turnstile/turnstile_180901.txt">Saturday, September 01, 2018</a>,
<a href="data/nyct/turnstile/turnstile_180825.txt">Saturday, August 25, 2018</a>,
<a href="data/nyct/turnstile/turnstile_180818.txt">Saturday, August 18, 2018</a>,
<a href="data/nyct/turnstile/turnstile_180811.txt">Saturday, August 11, 2018</a>,
<a href="data/nyct/turnstile/turnstile_180804.txt">Saturday, August 04, 2018</a>,
<a href="data/nyct/turnstile/turnstile_180728.txt">Saturday, July 28, 2018</a>,
<a href="data/nyct/turnstile/turnstile_180721.txt">Saturday, July 21, 2018</a>,
<a href="data/nyct/turnstile/turnstile_180714.txt">Saturday, July 14, 2018</a>,
<a href="data/nyct/turnstile/turnstile_180707.txt">Saturday, July 07, 2018</a>,
<a href="data/nyct/turnstile/turnstile_180630.txt">Saturday, June 30, 2018</a>,
<a href="data/nyct/turnstile/turnstile_180623.txt">Saturday, June 23, 2018</a>,
<a href="data/nyct/turnstile/turnstile_180616.txt">Saturday, June 16, 2018</a>,
<a href="data/nyct/turnstile/turnstile_180609.txt">Saturday, June 09, 2018</a>,
<a href="data/nyct/turnstile/turnstile_180602.txt">Saturday, June 02, 2018</a>,
<a href="data/nyct/turnstile/turnstile_180526.txt">Saturday, May 26, 2018</a>,
<a href="data/nyct/turnstile/turnstile_180519.txt">Saturday, May 19, 2018</a>,
<a href="data/nyct/turnstile/turnstile_180512.txt">Saturday, May 12, 2018</a>,
<a href="data/nyct/turnstile/turnstile_180505.txt">Saturday, May 05, 2018</a>,
<a href="data/nyct/turnstile/turnstile_180428.txt">Saturday, April 28, 2018</a>,
<a href="data/nyct/turnstile/turnstile_180421.txt">Saturday, April 21, 2018</a>,
<a href="data/nyct/turnstile/turnstile_180414.txt">Saturday, April 14, 2018</a>,</pre>

```



Workshop on
Web Scraping



CHRIST
(DEEMED TO BE UNIVERSITY)
BANGALORE · INDIA

Computer Society of India & Department of CSE
CHRIST (Deemed to be University), Bangalore-560074, India
4th Feb 2020

S.NO	Name of the Participants	Signature
1	S yoshita manavi	
2	Sushma S	
3	Nikitha bonthala	
4	Chakradhar	
5	Jagadesh	
6	Vasukivasan	
7	Rahul Raj Dixit	
8	Syed Saeed Ahmed	
9	Ritik Sahu	
10	SHIVAM Kumar Sahu	
11	T.vishnu vardhan	
12	Sarvesh Patel	
13	Tanisha Manoj	
14	Anagha D Ananth	
15	Saharsh Pamecha	
16	Ria Shrivastava	
17	Ahamed shafi usman therambil	
18	Shubham	
19	Chirag Kumar prajapat	
20	Ujjwal Kuikel	



Workshop on
Web Scraping



CHRIST
(DEEMED TO BE UNIVERSITY)
BANGALORE · INDIA

Computer Society of India & Department of CSE
CHRIST (Deemed to be University), Bangalore-560074, India
4th Feb 2020

S.NO	Name of the Participants	Signature
21	Jennifer Gracias	<i>Jennifer</i>
22	Prathyusha vaigandla venkata Sai	<i>Prathyusha</i>
23	Shubhkar yadav	— AB —
24	Mohammed Saadh Numaan	— AB —
25	K s yugesh	— AB —
26	Shruthi R	— AB —
27	ANKIT TIRKEY	<i>Ankit</i>
28	Sonam Pal	<i>Sonam</i>
29	Yash Sharma	— AB —
30	Shrishti Patidar	— AB —
31	Bhavya	— AB —
32	T.vishnu vardhan	— AB —
33	Ruchitha Reddy G	<i>Ruchitha</i>
34	Varuni M	<i>Varuni</i>
35	Sohan bachuwar	— AB —
36	Sharan Kumar M	— AB —
37	Niharika dsouza	<i>Niharika</i>
38	Ayush Choubey	<i>Ayush</i>
39	Aman kumar	<i>Aman</i>
40	Vinuthna Nekkanti	<i>Vinuthna</i>



Workshop on
Web Scraping



CHRIST
(DEEMED TO BE UNIVERSITY)
BANGALORE · INDIA

Computer Society of India & Department of CSE
CHRIST (Deemed to be University), Bangalore-560074, India
4th Feb 2020

S.NO	Name of the Participants	Signature
41	Sreelakshmi V A	— AB —
42	Hrishikesh Ajith	— AB —
43	Allam Vineeth Reddy	— AB —
44	Tania v	— AB —
45	Deepika b k	— AB —
46	Sonalika Gupta	— AB —
47	Sreelakshmi V A	— AB —
48	Yashesh Mankad	— AB —
49	Sreelakshmi K Nair	— AB —
50	Alex Biju	— AB —
51	A.Sushanth Reddy	— AB —
52	SHARLET MATHEWS	— AB —
53	Sona Ale Wilson	— AB —
54	Chinthala Sneha Reddy	— AB —
55	Anusha - P.V	— AB —
56	Govind Menon	— AB —
57		— AB —
58		— AB —
59		— AB —
60		— AB —