

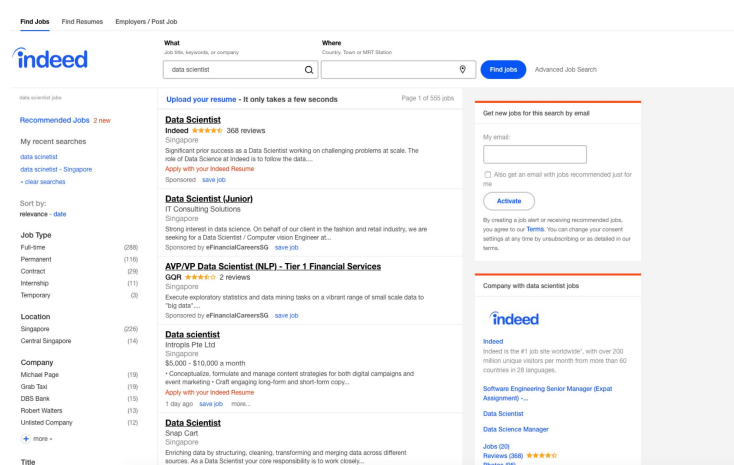
# Web Data Crawling

# Agenda

- What is HTML
- Hands-On

# What is HTML?

- **HTML: HyperText Markup Language**
  - a computer language that is used to create documents on the World Wide Web
  - simple and logical
  - a mark-up language that uses **<Tags>** instead of programming language
- All websites over the internet are plain text files that consist of HTML Tags.



```
<!DOCTYPE html>
<html lang="en" dir="ltr">
<head>
<meta http-equiv="content-type" content="text/html; charset=UTF-8">
<script type="text/javascript" src="//d3fw5vhllyvee.cloudfront.net/s/af4998f/en_5G.js"></script>
<link href="//d3fw5vhllyvee.cloudfront.net/s/978d98c/jobsearch_all.css" rel="stylesheet" type="text/css">
<link rel="alternate" type="application/rss+xml" title="Data Scientist Jobs, careers" href="http://www.indeed.com.sg/rss?q=data-scientist">
<link rel="alternate" media="only screen and (max-width: 640px)" href="/m/jobs?q=data-scientist">
<script type="text/javascript">

    if (typeof window['closureReadyCallbacks'] == 'undefined') {
        window['closureReadyCallbacks'] = [];
    }


    function call_when_jsall_loaded(cb) {
        if (window['closureReady']) {
            cb();
        } else {
            window['closureReadyCallbacks'].push(cb);
        }
    }
</script>
<meta name="ppstripist" content="">
<script type="text/javascript" src="//d3fw5vhllyvee.cloudfront.net/s/4c9f4c8/jobsearch-all-compiled_en_5G.js"></script>

var searchUID = '1ctkqc5ku7g6m888';
var tk = '1ctkqc5ku7g6m888';

var loggedIn = false;
var dcnpayload = 'jobse0;jobal8;viewj0;savej0;8232381';
var myindeed = true;
var userEmail = '';
var tellFriendEmail = '';
var globalLoginURL = 'https://www.indeed.com.sg/account/login?dest=k2fjobs%3Fq%3Ddata%2Bscientist%26l%3D';
var globalRegisterURL = 'https://www.indeed.com.sg/account/register?dest=k2fjobs%3Fq%3Ddata%2Bscientist%26l%3D';
var searchKey = 'f5281a4aeef1eal';
var searchState = 'q=data-scientist&pl=';
var searchQS = 'q=data-scientist';
var eventType = 'jobsearch';
var locale = 'en_5G';
function clickId (var a = document.getElementById(id); var hr = a.href; var si = a.href.indexOf('Gjsa='); if (si > 0) return;
function sjondId (var a = document.getElementById(id); var hr = a.href; var ocs = hr.indexOf('&oc='); if (ocs < 0) return;
function etatId (var a = document.getElementById(id); var hr = a.href; var l = a.href.indexOf('l='); if (l < 0) return;
```

# What is HTML?

[Find Jobs](#) [Find Resumes](#) [Employers / Post Job](#)



What

Job title, keywords, or company

data scientist

Where

Country, Town or MRT Station

Find jobs

Advanced Job Search

data scientist jobs

Recommended Jobs 2 new

My recent searches

data scientist

data scientist - Singapore

clear searches

Sort by:

relevance - date

Job Type

Full-time (288)

Permanent (116)

Contract (29)

Internship (11)

Temporary (3)

Location

Singapore (226)

Central Singapore (14)

Company

Michael Page (19)

Grab Taxi (19)

DBS Bank (15)

Robert Walters (13)

Unlisted Company (12)

more »

Title

Upload your resume - It only takes a few seconds

Page 1 of 555 jobs

Data Scientist

Indeed ★★★★★ 368 reviews

Singapore

Significant prior success as a Data Scientist working on challenging problems at scale. The role of Data Science at Indeed is to follow the data....

Apply with your Indeed Resume

Sponsored save job

Data Scientist (Junior)

IT Consulting Solutions

Singapore

Strong interest in data science. On behalf of our client in the fashion and retail industry, we are seeking for a Data Scientist / Computer vision Engineer at...

Sponsored by eFinancialCareersSG save job

AVP/VP Data Scientist (NLP) - Tier 1 Financial Services

GQR ★★★★★ 2 reviews

Singapore

Execute exploratory statistics and data mining tasks on a vibrant range of small scale data to "big data"....

Sponsored by eFinancialCareersSG save job

Data scientist

Intropolis Pte Ltd

Singapore

\$5,000 - \$10,000 a month

• Conceptualize, formulate and manage content strategies for both digital campaigns and event marketing • Craft engaging long-form and short-form copy...

Apply with your Indeed Resume

1 day ago save job more...

Data Scientist

Snap Cart

Singapore

Enriching data by structuring, cleaning, transforming and merging data across different sources. As a Data Scientist your core responsibility is to work closely...

Get new jobs for this search by email

My email:

☐ Also get an email with jobs recommended just for me

Activate

By creating a job alert or receiving recommended jobs, you agree to our Terms. You can change your consent settings at any time by unsubscribing or as detailed in our terms.

Company with data scientist jobs

Indeed

Indeed is the #1 job site worldwide\*, with over 200 million unique visitors per month from more than 60 countries in 28 languages.

Software Engineering Senior Manager (Expat Assignment) ~...

Data Scientist

Data Science Manager

Jobs (20)

Reviews (368) ★★★★★

Photos (95)

```
<!DOCTYPE html>
<html lang="en" dir="ltr">
<head>
<meta http-equiv="content-type" content="text/html; charset=UTF-8">
<script type="text/javascript" src="//d3fw5vlhllyvee.cloudfront.net/s/af4998f/en_SG.js"></script>
<link href="//d3fw5vlhllyvee.cloudfront.net/s/970d98c/jobsearch_all.css" rel="stylesheet" type="text/css">
<link rel="alternate" type="application/rss+xml" title="Data Scientist Jobs, careers" href="http://www.indeed.com.sg/rss?q=dat
<link rel="alternate" media="only screen and (max-width: 640px)" href="/m/jobs?q=data+scientist">
<link rel="alternate" media="handheld" href="/m/jobs?q=data+scientist">
<script type="text/javascript">

    if (typeof window['closureReadyCallbacks'] == 'undefined') {
        window['closureReadyCallbacks'] = [];
    }

    function call_when_jsall_loaded(cb) {
        if (window['closureReady']) {
            cb();
        } else {
            window['closureReadyCallbacks'].push(cb);
        }
    }
</script>
<meta name="ppstriptst" content="1">

<script type="text/javascript" src="//d3fw5vlhllyvee.cloudfront.net/s/4c9f4c0/jobsearch-all-compiled_en_SG.js"></script>
<script type="text/javascript">

var searchUID = '1ctkqc5ku7g6m800';
var tk = '1ctkqc5ku7g6m800';

var loggedIn = false;
var dcmPayload = 'jobse0;jobal0;viewj0;savej0;8232301';
var myindeed = true;
var userEmail = '';
var tellFriendEmail = '';
var globalLoginURL = 'https://\www.indeed.com.sg/account/login?dest=%2Fjobs%3Fq%3Ddata%2Bscientist%26l%3D';
var globalRegisterURL = 'https://\www.indeed.com.sg/account/register?dest=%2Fjobs%3Fq%3Ddata%2Bscientist%26l%3D';
var searchKey = 'f5281a4aee71eca1';
var searchState = 'q=data+scientist&ml=';
var searchQS = 'q=data+scientist';
var eventType = 'jobsearch';
var locale = 'en_SG';
function clk(id) { var a = document.getElementById(id); var hr = a.href; var si = a.href.indexOf('&jsa='); if (si > 0) return;
function sjomd(id) { var a = document.getElementById(id); var hr = a.href; var ocs = hr.indexOf('&oc=1'); if (ocs < 0) return;
function sjocl(id, cal) { var a = document.getElementById(id); a href = a href + '&oc=1&cal=' + cal; }
```

# Tags

- Tags are instructions to markup the text shown on your Web browser.
- All tags are in the format **<Tags>**
- Each tag must be accompanied by a closing tag **</Tags>**
- Elements are made up of two tags (start one and end one) and the element content.

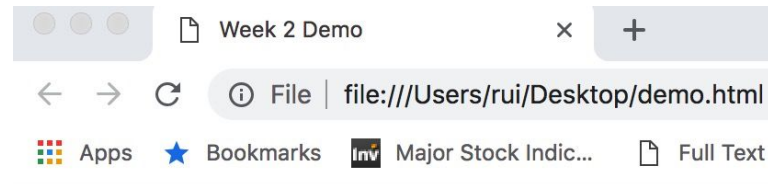
`<title>Business Analytics</title>`

# Toy Example

```
<!DOCTYPE html>
<html>
<head>
<title>Week 2 Demo</title>
</head>
<body>

<h1>My first heading.</h1>
<p>My first pargarph.</p>

</body>
</html>
```



**My first heading.**

My first pargarph.

- Browser use HTML tags to decide how to display the document.
  - **<html>** root element of an HTML page
  - **<head>** contains elements that are about the document which are not displayed in the page itself. **<title>** is one of such element
  - **<body>** is the web page itself
  - **<h1>** defines a large heading and **<p>** defines a paragraph

# Beautiful Soup

- Beautiful Soup is a Python library for parsing HTML documents (including having malformed markup), whose name is derived more from the unrelated “tag soup”.
- Help you pull data out of HTML and XML files.