Workshop 2

Web Scraping
Part 1

Slides courtesy of Skooldio
Modified and used with permission

Data Collection

Websites

There are over 1 billion websites on the world wide web today!

Wikipedia
5 million articles in the English Wikipeda

Amazon
400M products sold on amazon.com

TripAdvisor
6.8 million business and properties

Data Collection

Social media

Tons of user-generated content

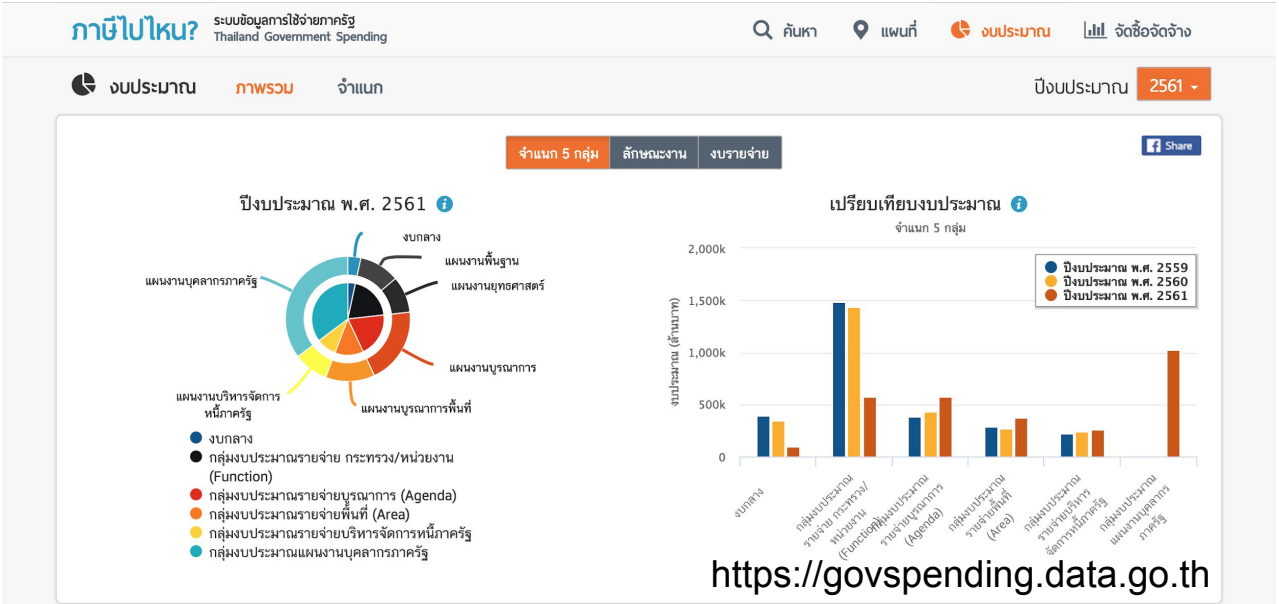
Facebook
more than 60 million active business Pages

Twitter
500 million tweets per day

Instagram
80 million photos uploaded every day

Data Collection

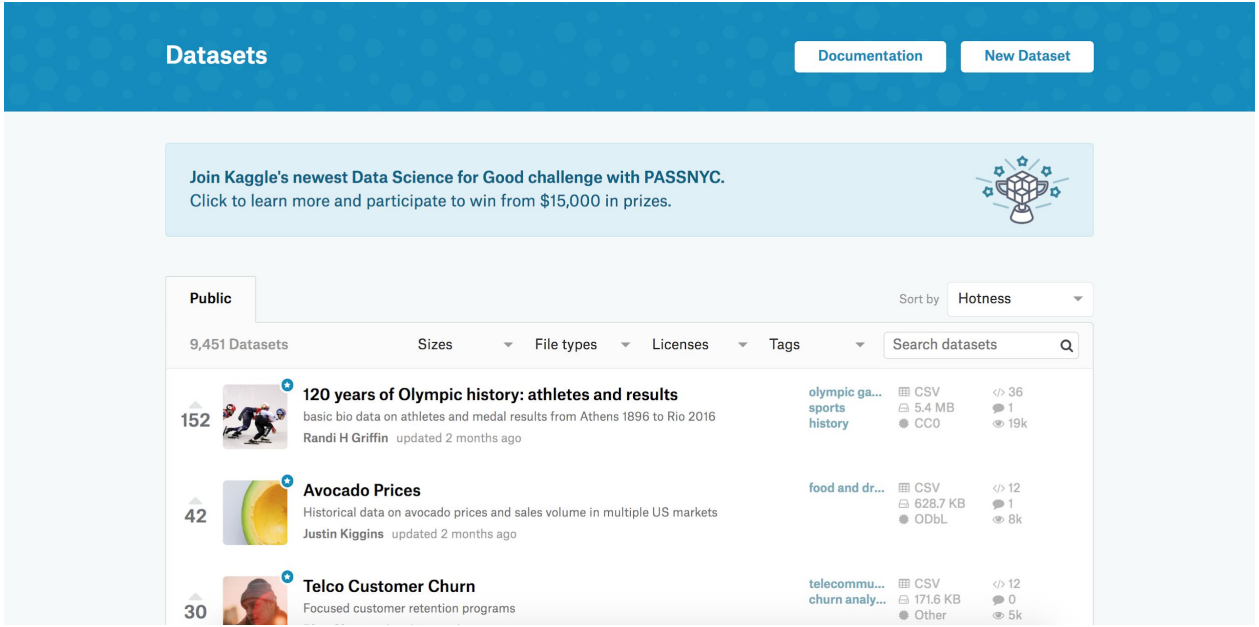
Open data



Governmental Open Data https://govspending.data.go.th/

Data Collection

Kaggle



TOPICS

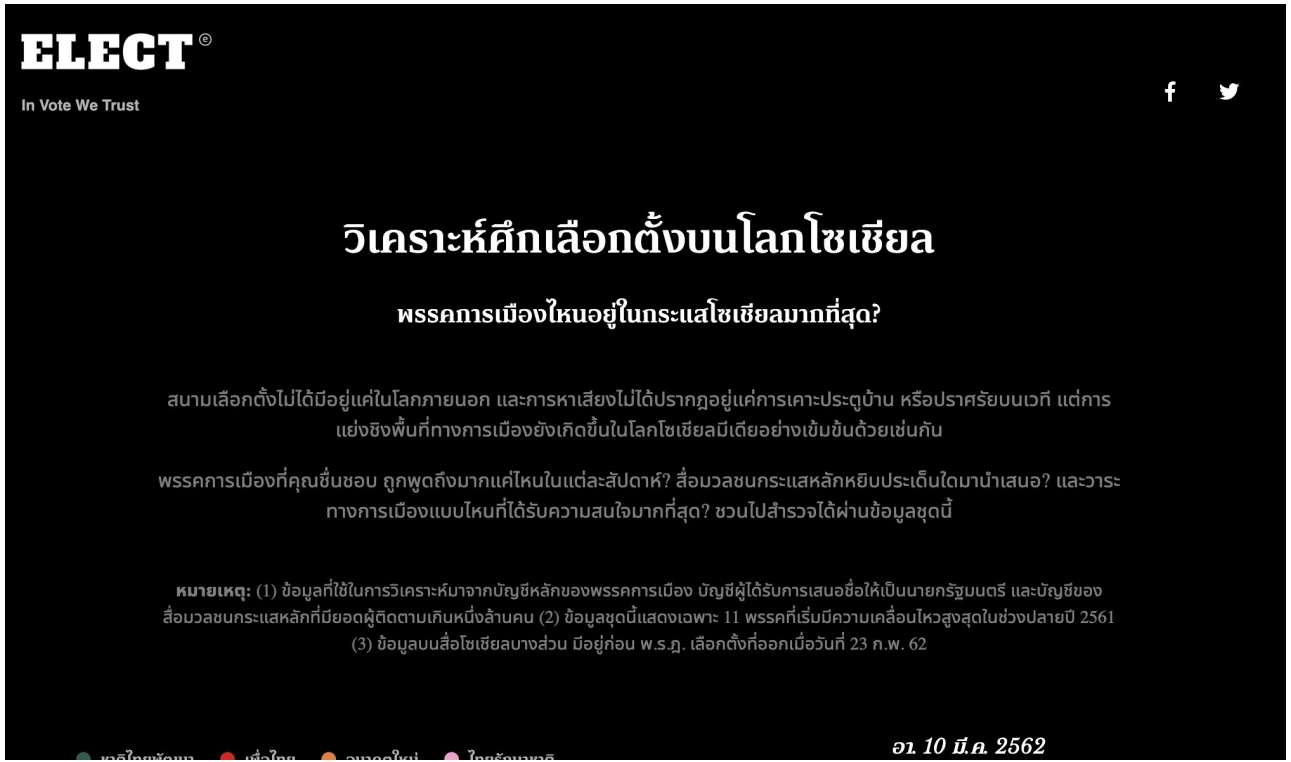
- Web Scraping
- APIs (next week)
- Regular Expressions (next week)

Web Scraping

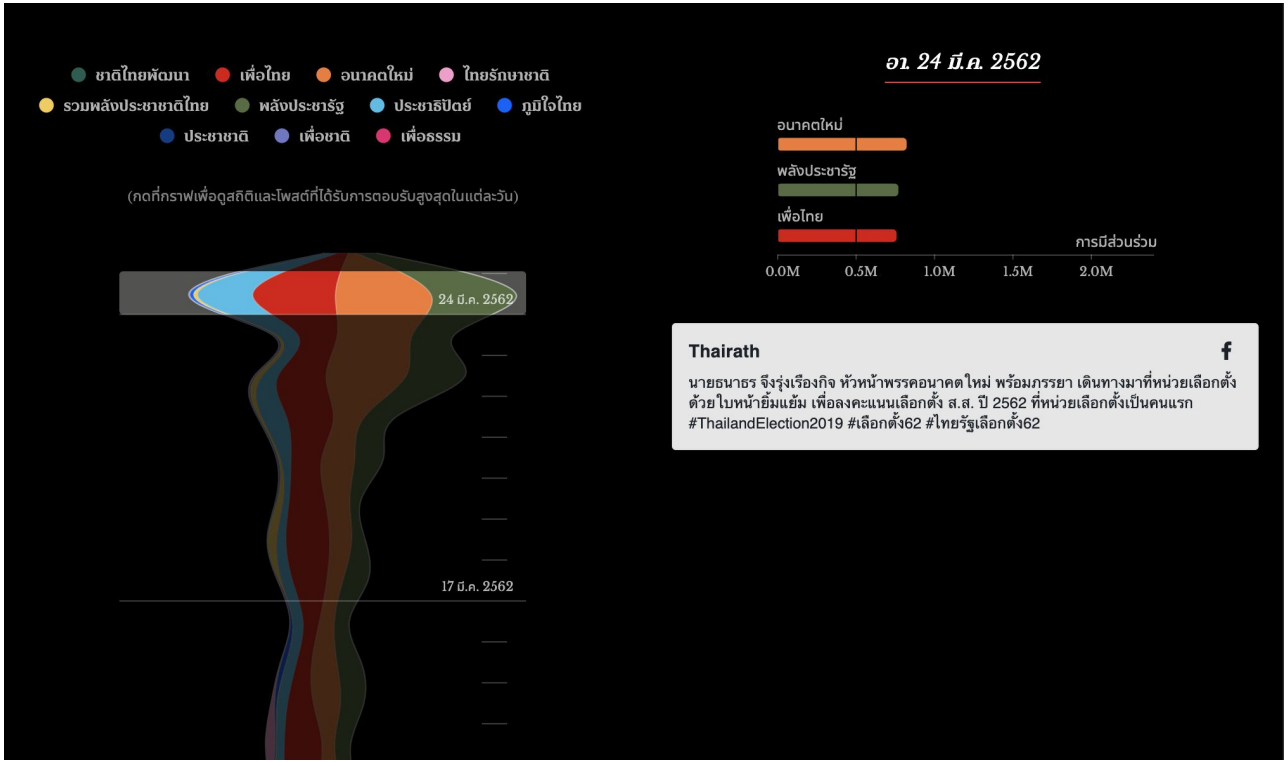
What is web scraping ?

- A process of extracting information from websites
- It usually refers to an automated program that simulates a person viewing a website
- The process involves downloading a web page, parsing and extracting information from it, and store the target information in a proper format

Use case


ELECT ©
In Vote We Trust
วิเคราะห์ศึกเลือกตั้งบนโลกโซเชียล
พรรคการเมืองไหนอยู่ในกระแสโซเชียลมากที่สุด?
สนามเลือกตั้งไม่ได้มีอยู่แค่ในโลกภายนอก และการหาเสียงไม่ได้ปรากฏอยู่แค่การเคาะประตูบ้าน หรือปราศรัยบนเวที แต่การแย่งชิงพื้นที่ทางการเมืองยังเกิดขึ้นในโลกโซเชียลมีเดียอย่างเข้มข้นด้วยเช่นกัน
พรรคการเมืองที่คุณชื่นชอบ ถูกพูดถึงมากแค่ไหนในแต่ละสัปดาห์? สื่อมวลชนกระแสหลักหยิบประเด็นใดมานำเสนอ? และวาระทางการเมืองแบบไหนที่ได้รับความสนใจมากที่สุด? ชวนไปสำรวจได้ผ่านข้อมูลชุดนี้
หมายเหตุ: (1) ข้อมูลที่ใช้ในการวิเคราะห์มาจากบัญชีหลักของพรรคการเมือง บัญชีผู้ได้รับการเสนอชื่อให้เป็นนายกรัฐมนตรี และบัญชีของสื่อมวลชนกระแสหลักที่มียอดผู้ติดตามเกินหนึ่งล้านคน (2) ข้อมูลชุดนี้แสดงเฉพาะ 11 พรรคที่เริ่มมีความเคลื่อนไหวสูงสุดในช่วงปลายปี 2561 (3) ข้อมูลบนสื่อโซเชียลบางส่วน มีอยู่ก่อน พ.ร.ฎ. เลือกตั้งที่ออกเมื่อวันที่ 23 ก.พ. 62
อา. 10 มี.ค. 2562

Use case


ชาติไทยพัฒนา
เพื่อไทย
อนาคตใหม่
ไทยรักษาชาติ
รวมพลังประชาชาติไทย
พลังประชารัฐ
ประชาธิปัตย์
ภูมิใจไทย
ประชาชาติ
เพื่อชาติ
เพื่อธรรม
(กดที่กราฟเพื่อดูสถิติและโพสต์ที่ได้รับการตอบรับสูงสุดในแต่ละวัน)
24 มี.ค. 2562
17 มี.ค. 2562
อา. 24 มี.ค. 2562
อนาคตใหม่
พลังประชารัฐ
เพื่อไทย
การมีส่วนร่วม
0.0M
0.5M
1.0M
1.5M
2.0M
Thairath
นายธนาธร จึงรุ่งเรืองกิจ หัวหน้าพรรคอนาคตใหม่ พร้อมภรรยา เดินทางมาที่หน่วยเลือกตั้ง ด้วยใบหน้ายิ้มแย้ม เพื่อลงคะแนนเลือกตั้ง ส.ส. ปี 2562 ที่หน่วยเลือกตั้งเป็นคนแรก #ThailandElection2019 #เลือกตั้ง62 #ไทยรัฐเลือกตั้ง62

Web Scraping

Ethics

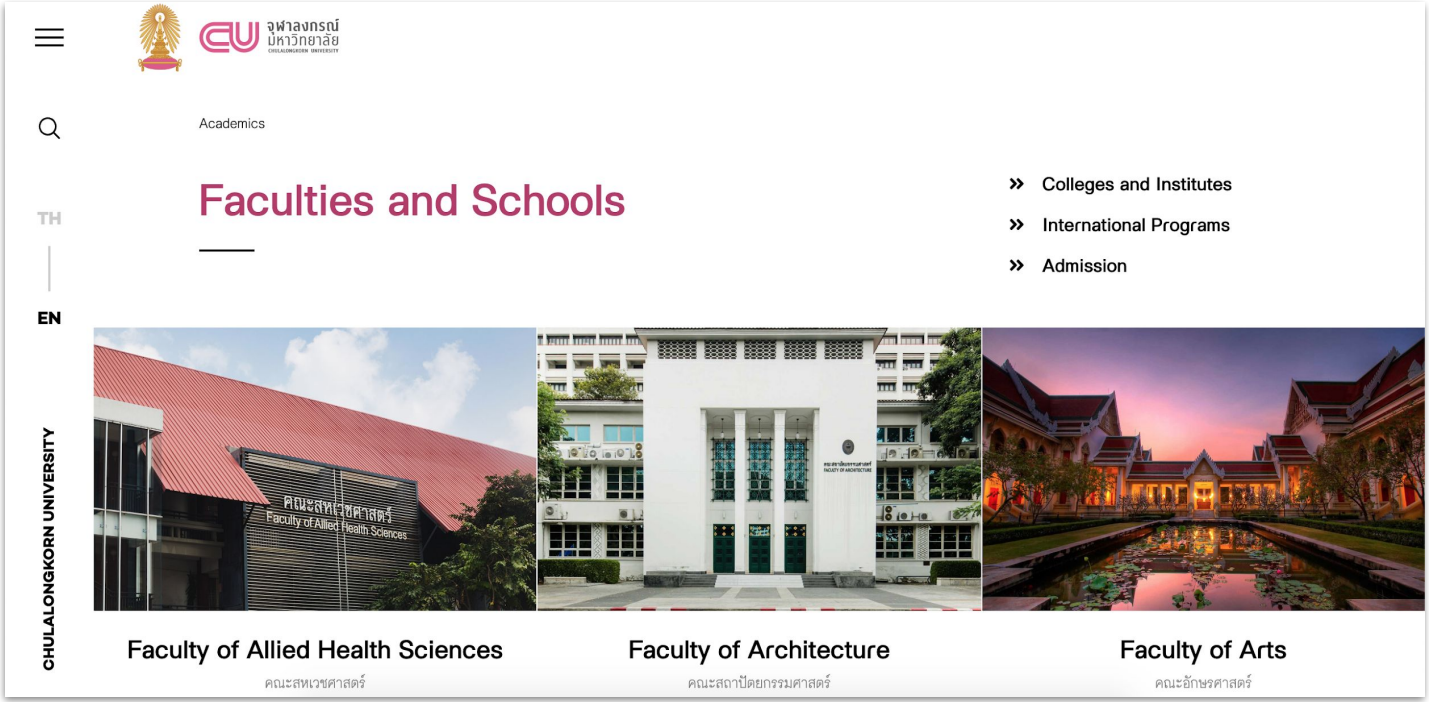
- Always check a website's **Terms and Conditions**
- Publishing the scraped data might violate copyright laws
- Act like a human - make requests at a reasonable rate
- Check the robots.txt file

robot.txt

- The file tells robots which pages on the site they should not visit
- The file is located in the top-level directory of websites
  - https://en.wikipedia.org/robots.txt
- Robots may simply ignore your instructions!

Scraping Workshop:

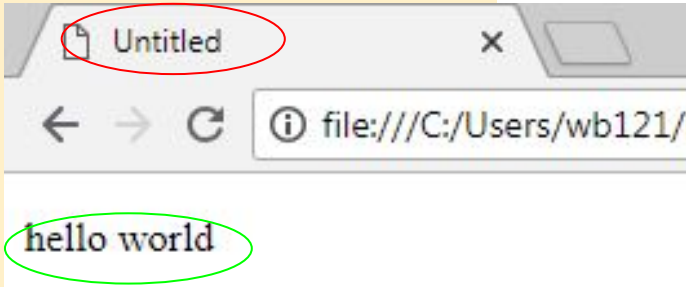
Web Scraping



HTML Essentials

HTML (Hypertext Markup Language)

```
<!DOCTYPE html>
<html>
    <head>
        <title>Untitled</title>
    </head>
    <body>
        <p>hello world</p>
    <body>
</html>
```

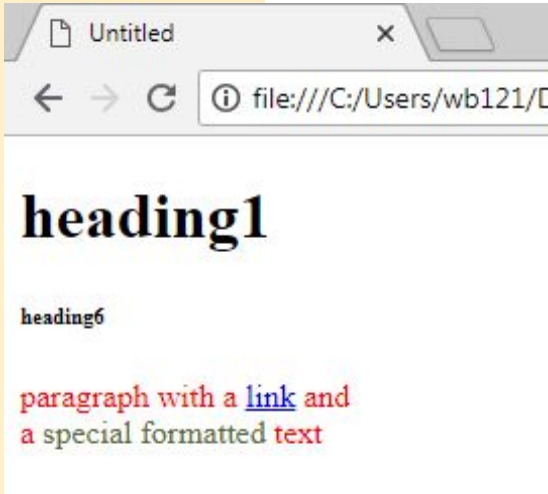


Every statement has opening < > and ending </ >

HTML Tags

 defines a section

```
<! - -  This is a comment  - - >
<div id="group1" class="footnote">
<h1>heading1</h1>
<h6>heading6</h6>
<p style="color:red;">
paragraph with a <a href="google.com">link</a>
and <br>
a <span style="color:darkolivegreen">special
formatted</span> text
</p>
</div>
```



Html ignores “Enter” in the code

 create a new line
Headings <h> and paragraphs <p> automatically enters a new line

HTML Lists

Unordered list

- item
- item
- item

Ordered list

1. first
2. second
3. third

```
<ul>
    <li>item</li>
    <li>item</li>
    <li>item</li>
</ul>

<ol>
    <li>first</li>
    <li>second</li>
    <li>third</li>
</ol>
```

HTML Tables

A	B
A1	B1
A2	B2

<tr> starts a row
<td> starts a cell

```
<table>
    <tr>
        <th>A</th>
        <th>B</th>
    </tr>
    <tr>
        <td>A1</td>
        <td>B1</td>
    </tr>
    <tr>
        <td>A2</td>
        <td>B2</td>
    </tr>
</table>
```

HTML Tables

A	B
A1	B1
A2	B2

<thead> <tbody> <tfoot>

Specifies which part is the header or body.
Can assign special tricks to each part.

```
<table>
    <thead>
        <tr>
            <th>A</th>  <th>B</th>
        </tr>
    </thead>
    <tbody>
        <tr>
            <td>A1</td>  <td>B1</td>
        </tr>
        <tr>
            <td>A2</td>  <td>B2</td>
        </tr>
    </tbody>
</table>
```

HTML Tables

<table>
<tr><td rowspan="2">A</td><td>B1</td></tr>
<tr><td>B2</td></tr>
</table>

```
<table>
        <tr>
                <td rowspan=”2”>A</td>
                <td>B1</td>
        </tr>
        <tr>
                <td>B2</td>
        </tr>
</table>
```

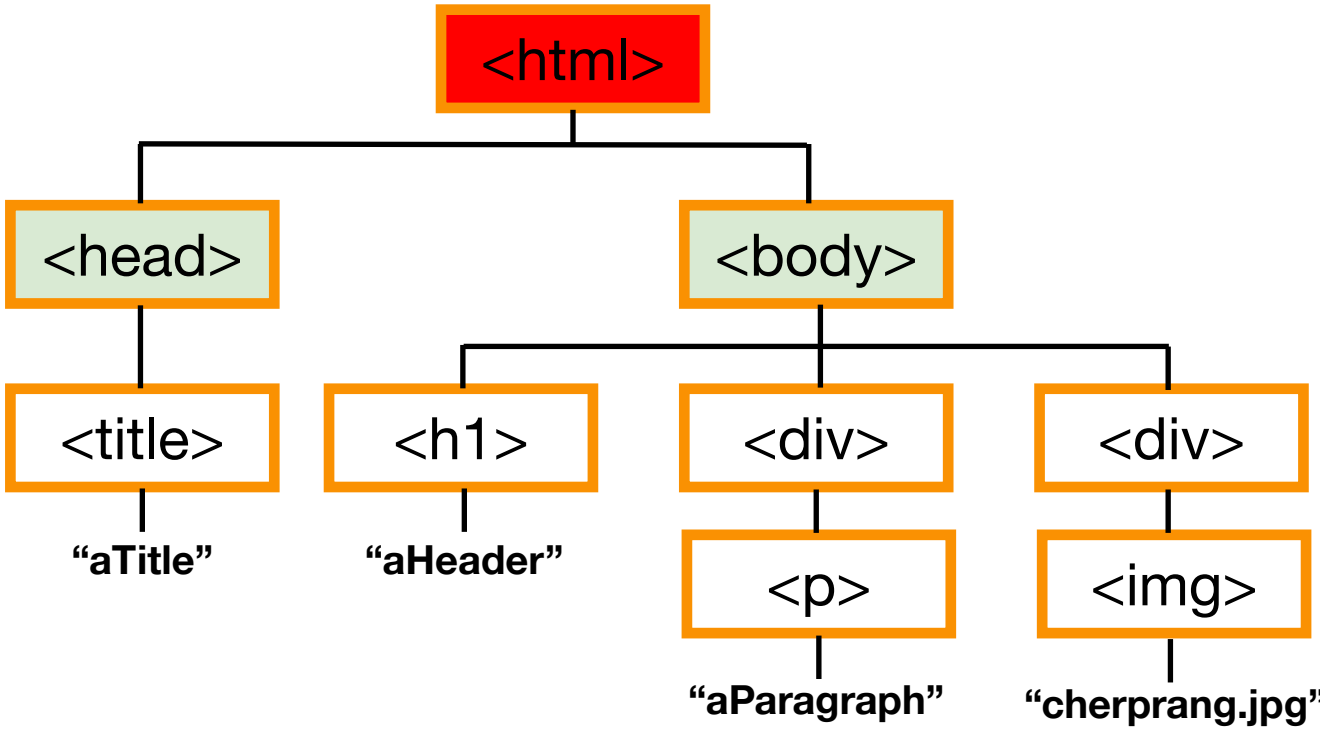
HTML Attributes

- **id** provides a document-wide unique identifier for an element

- **class** specifics one or more classes for an element

```
<div class=“content”></div>
<div class=“content highlight”></div>
```

DOM Tree (Document Object Model)




```
1   <!DOCTYPE html>
2   <html>
3   <head>
4   <title>aTitle</title>
5   </head>
6   <body>
7   <h1>aHeader</h1>
8   <div id="group1" class=
9   <p>aParagraph</p>
10  </div>
11  <div id="group2">
12  <img src="cherprang.jpg
13  </div>
14  </body>
15  </html>
```

Lab 0 : Inspect a web page

- ทดลอง inspect เว็บไซต์ โดยเข้าไปที่
  https://www.chula.ac.th/en/academics/faculties-and-schools/

- เปิด Developer Tools ใน web browser (แนะนำให้ใช้ Chrome)

- Google Chrome:
  View -> Developer -> Developer Tools

Lab 1 : Crawl a web page

- Part I: แสดงชื่อคณะทั้งหมดของจุฬา บรรทัดละชื่อ
- Part II: โหลดรูปคณะต่างๆ
- Part III: หาเบอร์โทร

Next week : beautifulsoup, API, regex

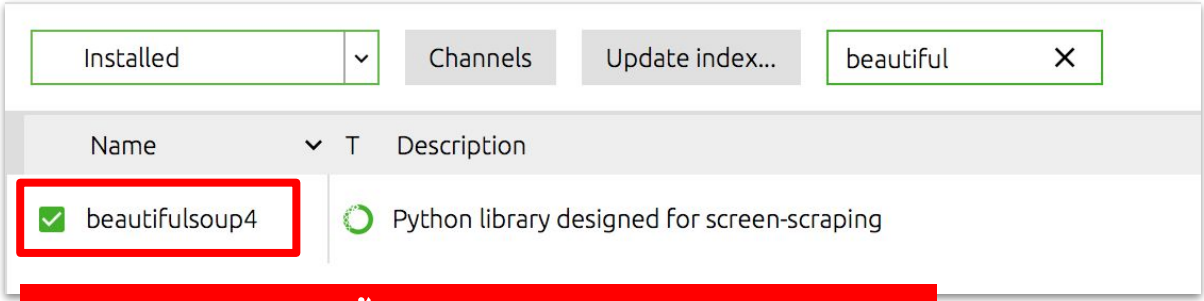
Web Scraping with Python

Setup prerequisite

1. Anaconda (Python 3)

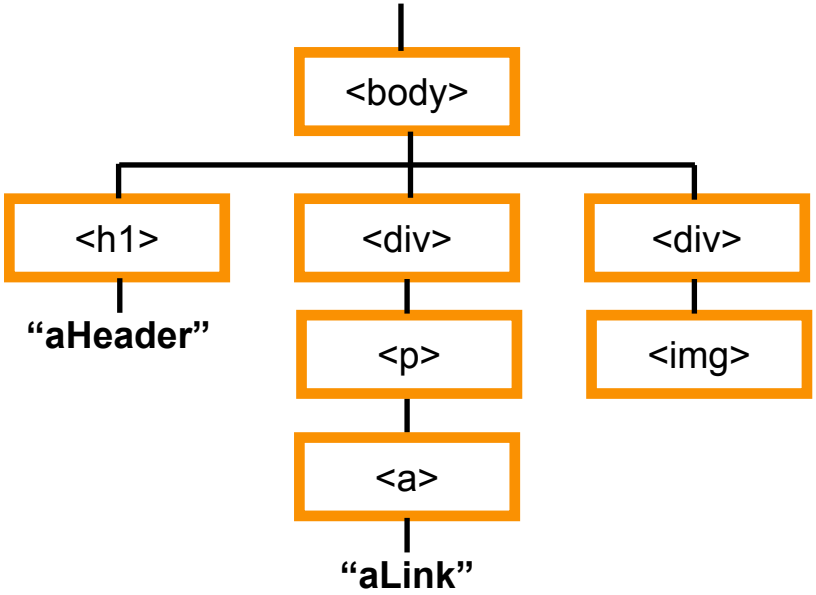
2. Jupyter Notebook or Jupyter Lab

3. BeautifulSoup Library
   -> Run from bs4 import BeautifulSoup



ตรวจสอบว่าได้ติดตั้ง BeautifulSoup4 เรียบร้อยแล้ว

BeautifulSoup Primer

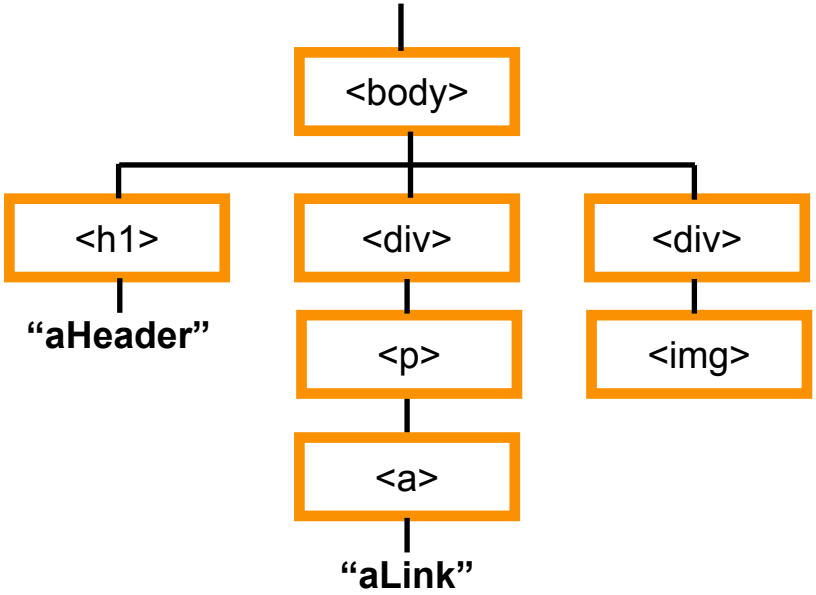



```
<body>
        <h1>aHeader</h1>
        <div class=”section1”>
                <p>
                        <a href=”#”>aLink</a>
                </p>
        </div>
        <div class=”section2”>
                <img src=”img.jpg/>
        </div>
</body>
```

Workshop 2.1 : BeautifulSoup

01-basic_beautifulsoup.ipynb

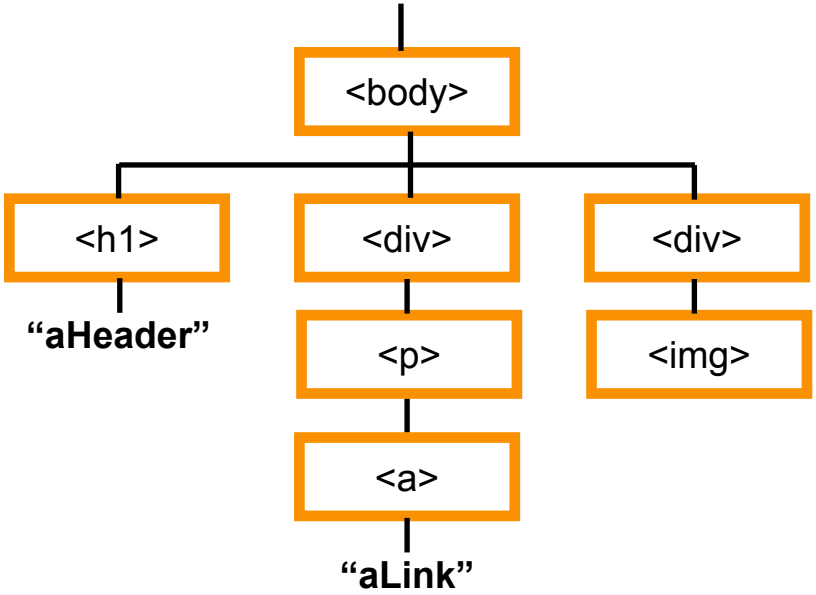
BeautifulSoup Primer : Find the target element




```
s = BeautifulSoup(html,’html.parser’)

s.body
```

BeautifulSoup Primer : Find the target element

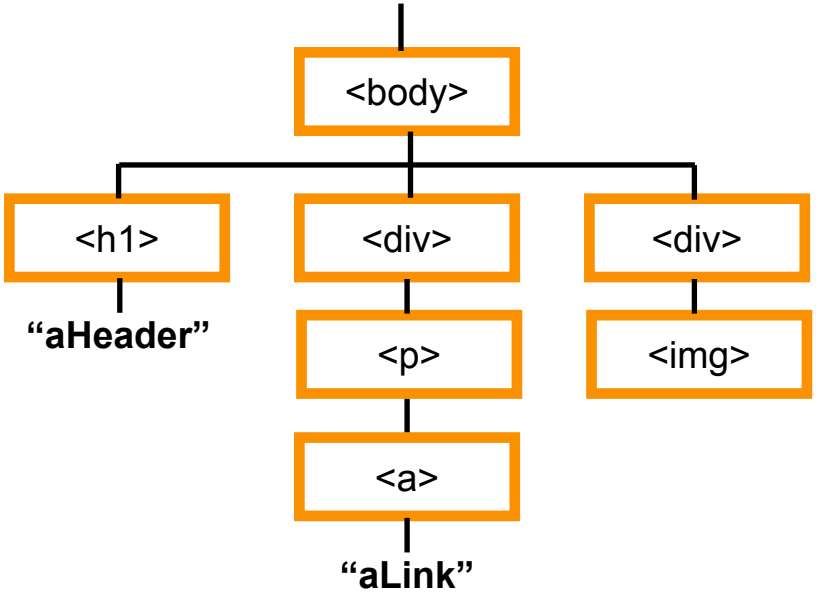


s.h1

or

s.find('h1')

BeautifulSoup Primer : Find the target element

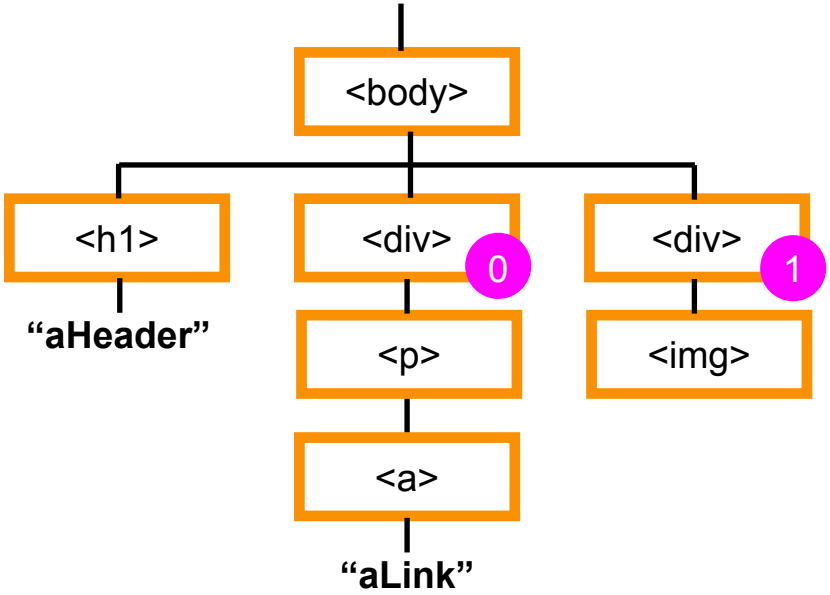


s.div

or

s.find('div')

BeautifulSoup Primer : Find the target element

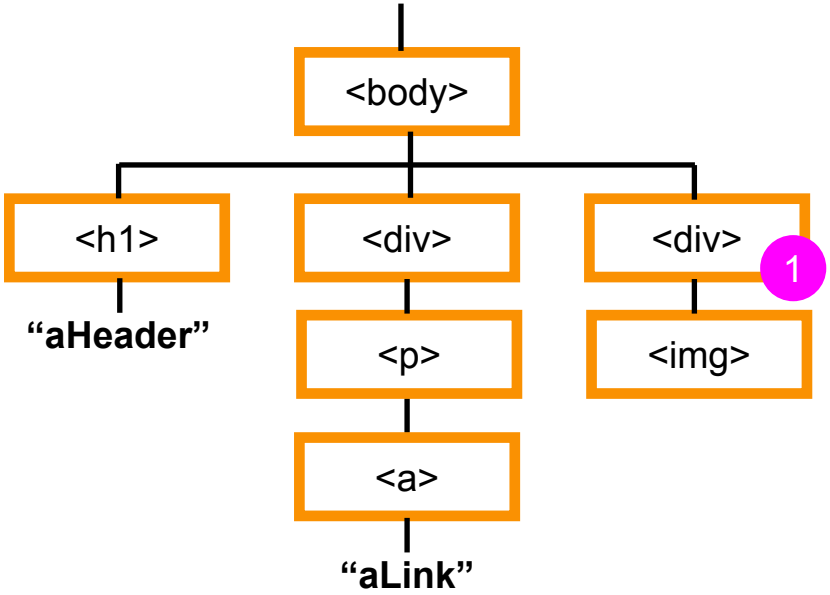


s.('div')

or

s.find_all('div')

BeautifulSoup Primer : Find the target element

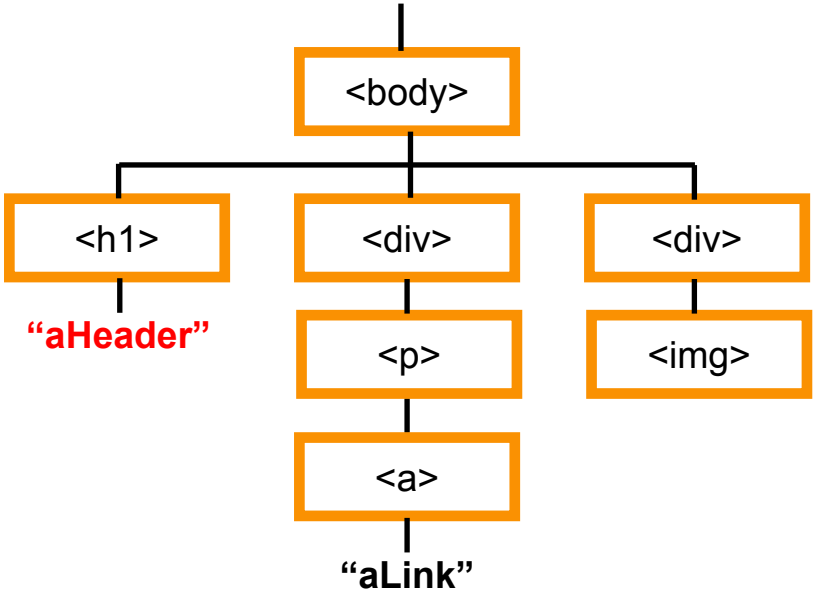


s.('div')[1]

or

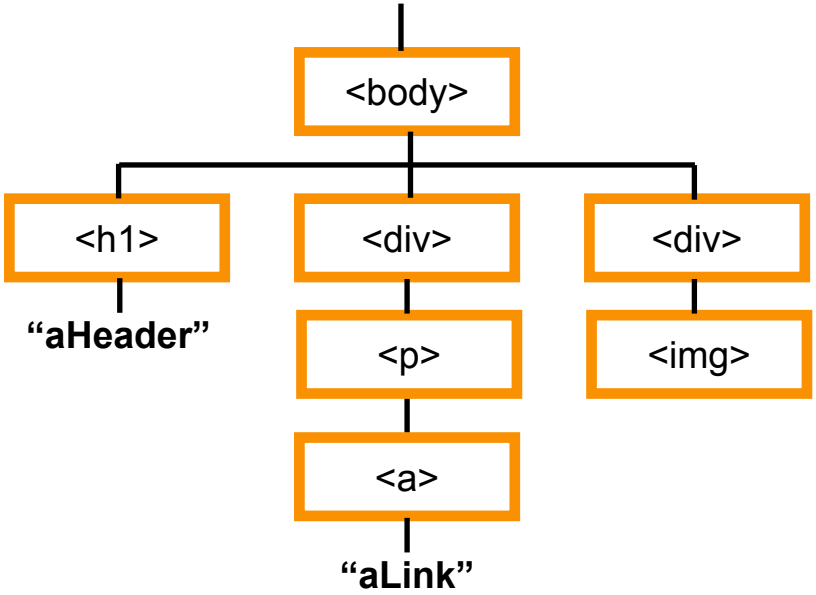
s.find_all('div')[1]

BeautifulSoup Primer : Find the target element



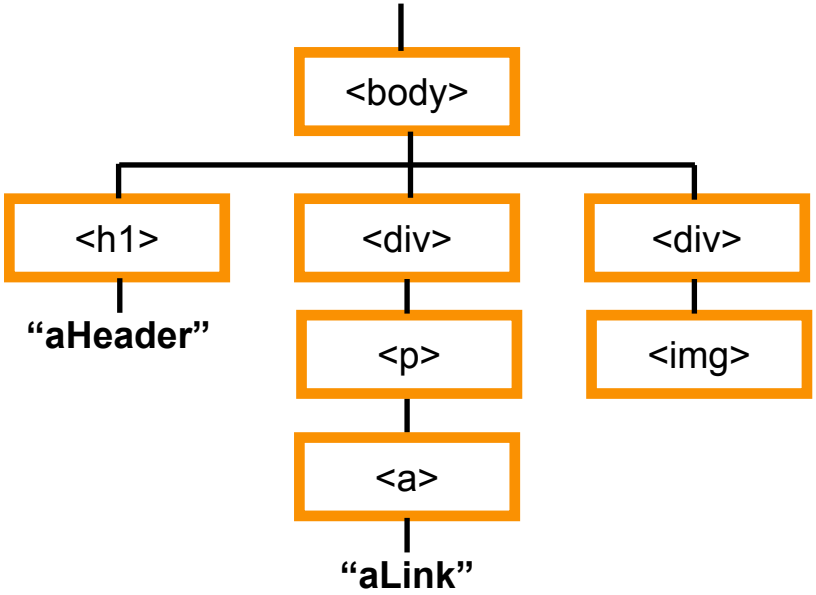

```
s.find(text='aHeader')
```

BeautifulSoup Primer : Find the target element



s.find(‘h1’ , string=’aHeader’)

BeautifulSoup Primer : Find the target element




```
s.find( attrs={ 'class' : 'section1' } )
```

or

```
s.find(class='section1')
```

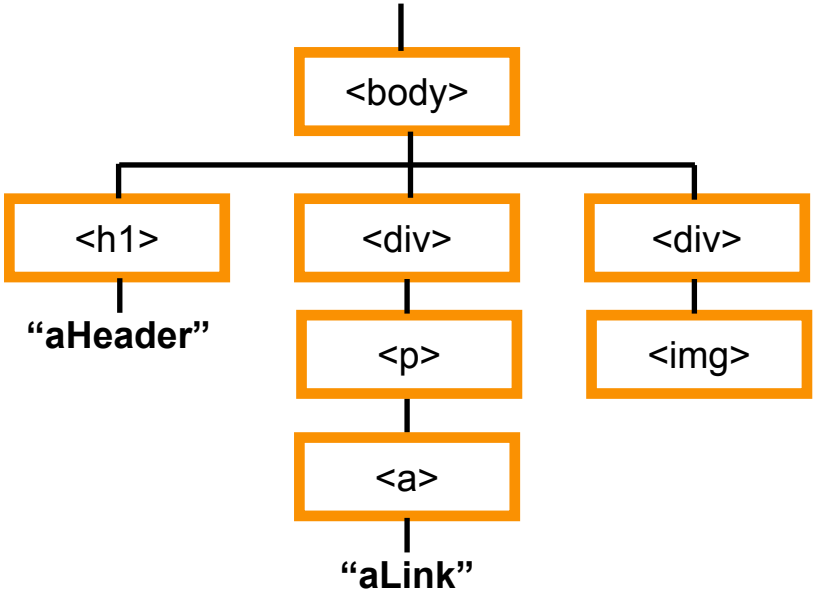
find( )

Using : find(name, attrs, recursive, string, **kwargs)

ใช้สำหรับกรองข้อมูลโดยใช้ tag name , attribute และ ข้อความใน string

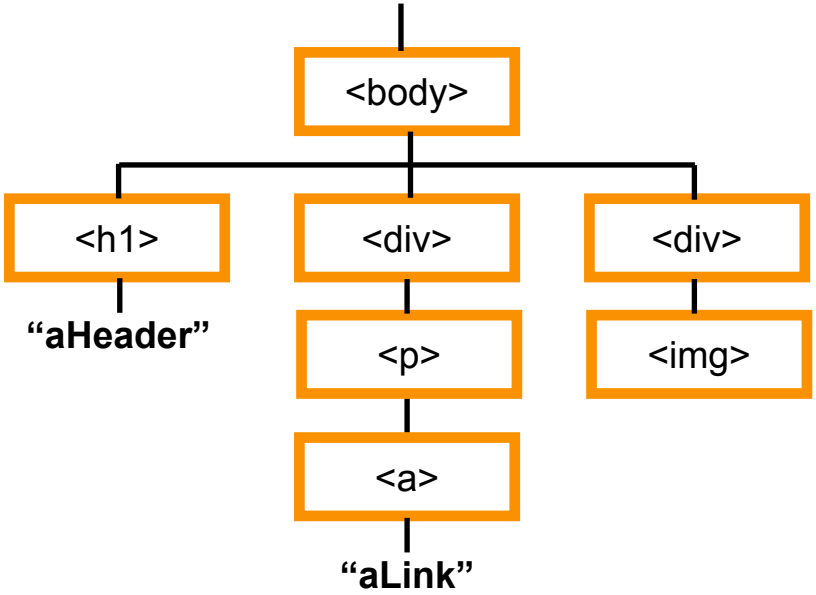
1. string “b”
2. regular expression re.compile(“^b”)
3. list [“a”,”b”]
4. True
5. Function ที่มีการคืนค่า True หรือ False

BeautifulSoup Primer : Traverse the DOM tree



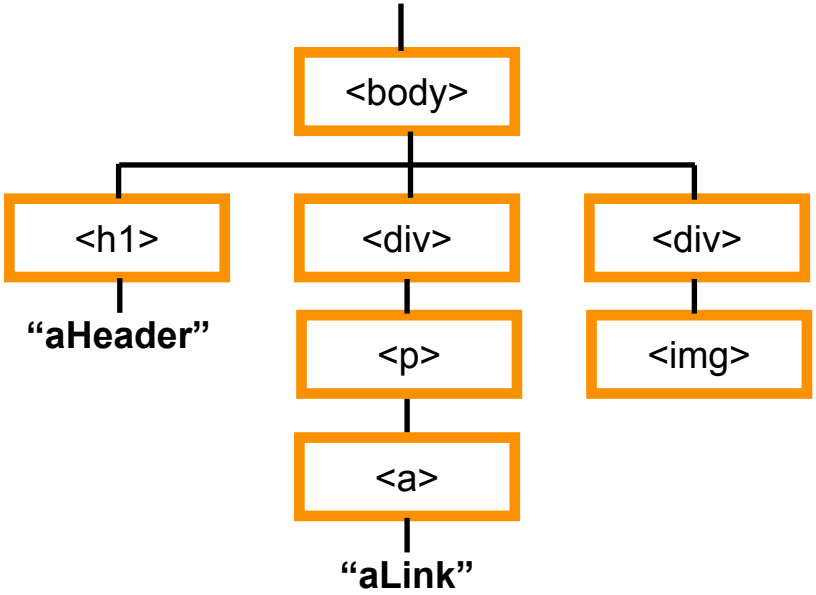
s.div.parent

BeautifulSoup Primer : Traverse the DOM tree



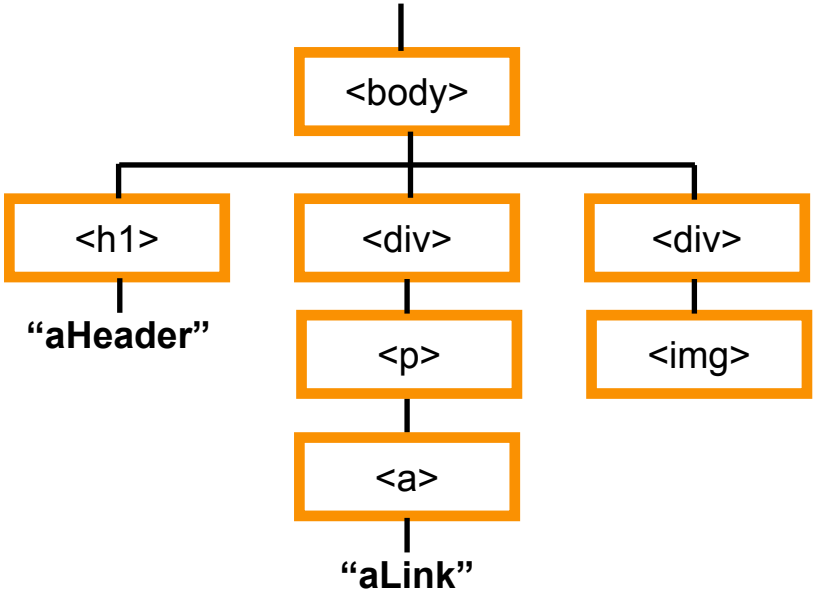
s.div.previous_sibling

BeautifulSoup Primer : Traverse the DOM tree



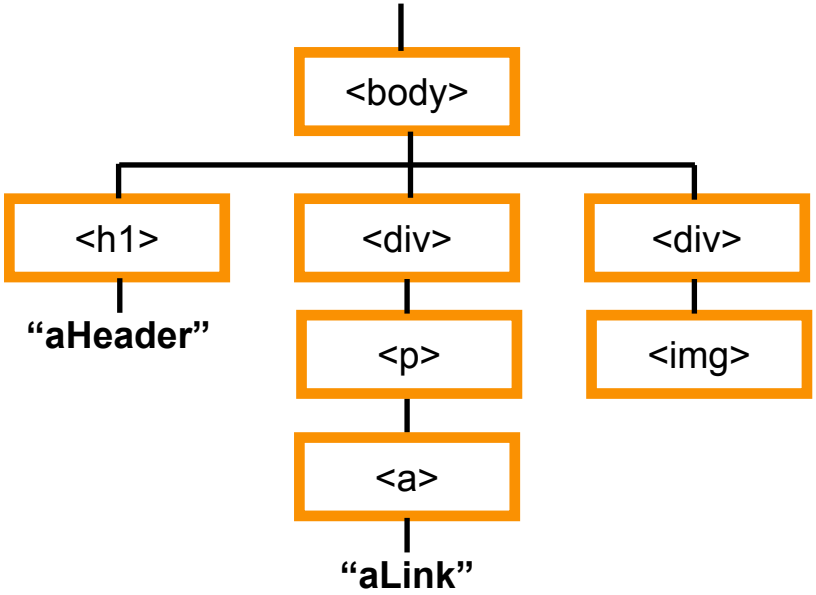

```
s.div.next_sibling
```

BeautifulSoup Primer : Traverse the DOM tree



s.div.next_element

BeautifulSoup Primer : Traverse the DOM tree

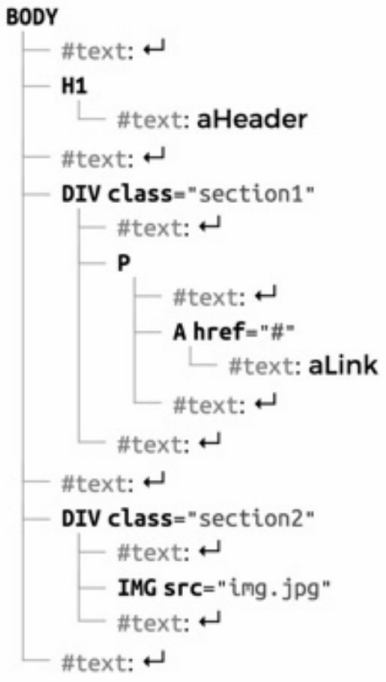


s.div.parent

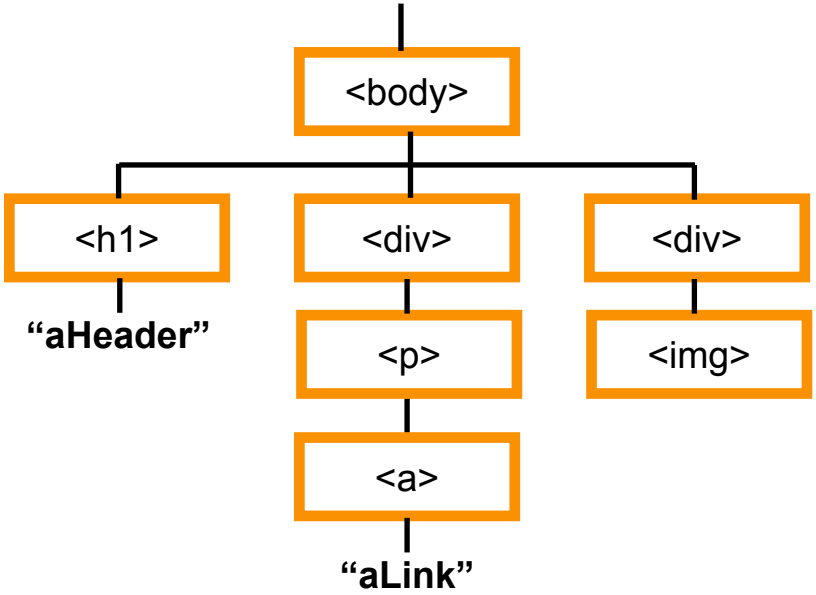
Warning : for white spaces !!!

จะเกิด ช่องว่าง (space) และ ขึ้นบรรทัดใหม่ (new lines) ระหว่าง tag

```
<body>
        <h1>aHeader</h1>
        <div class=”section1”>
                <p>
                        <a href=”#”>aLink</a>
                </p>
        </div>
        <div class=”section2”>
                <img src=”img.jpg/>
        </div>
</body>
```

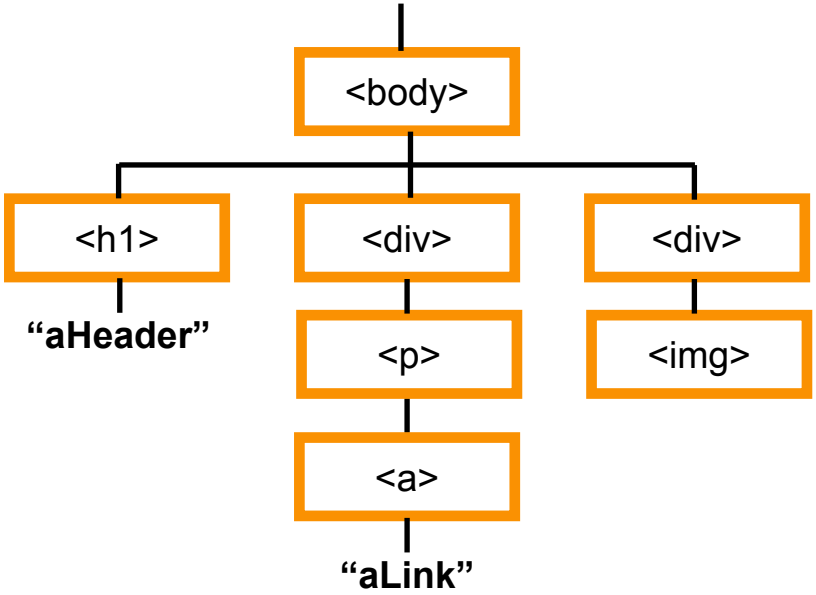


BeautifulSoup Primer : Traverse the DOM tree



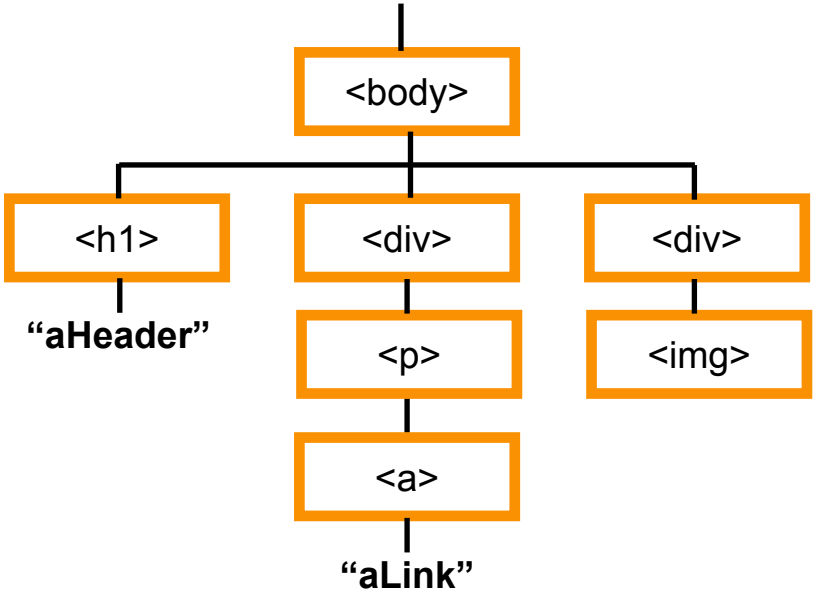

```
s.a.find_next()
```

BeautifulSoup Primer : Traverse the DOM tree



s.a.find_next(‘img’)

BeautifulSoup Primer : Traverse the DOM tree




```
s.find('div', class_='section2') \
    .find_previous_sibling('h1')
```

Workshop 2.2 : Data Scraping

02-web_scraping.ipynb

- ดึงข้อมูลรายชื่อบริษัทในเครือปัจจุบันในสังกัด GDH
- ดึงข้อมูลรายชื่อผู้กำกับภาพยนตร์ในสังกัด GDH
- ดึงข้อมูลรายชื่อนักแสดงในสังกัดนาดาวบางกอก
- ดึงข้อมูลรายชื่อภาพยนตร์ในเครือ GDH พร้อมทั้ง วันเปิดตัว , รายได้ และ ผู้กำกับ

APIs

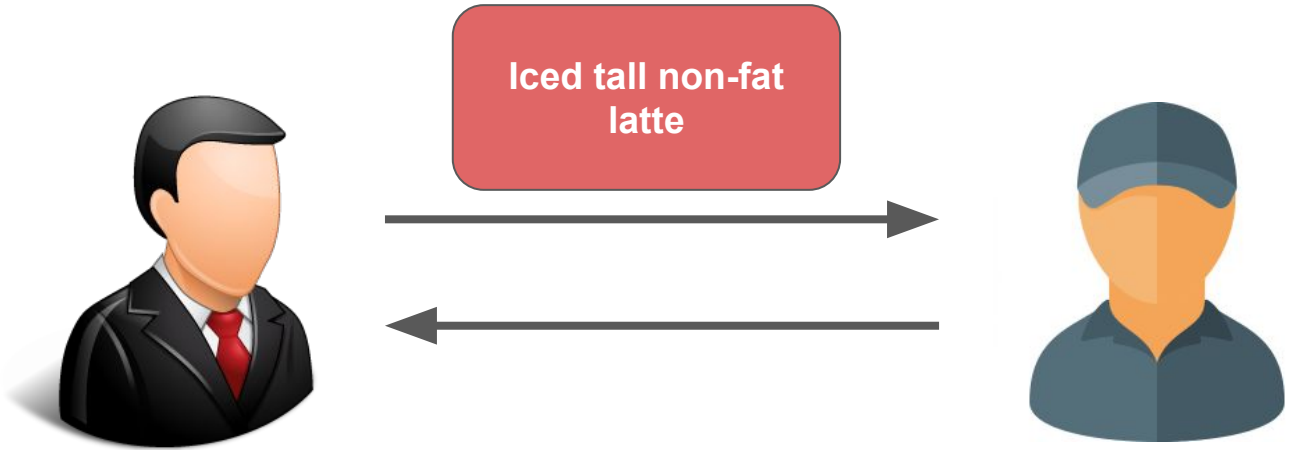
What is an API ?

- Application Programming Interface (API)

- It’s like a coding contract provided by computer software to another describing the way they can interact:
  - the expected input (request)
  - the expected output (response)

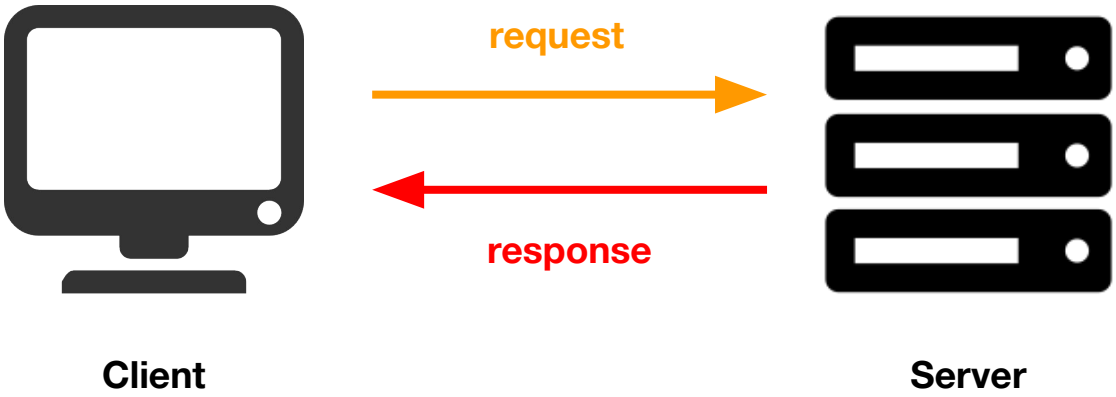
What is an API ?

What is an API ?

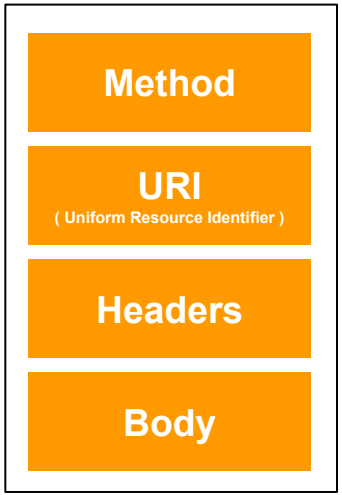


HTTP : Hypertext Transfer Protocol

- A request–response protocol
- Foundation of data communication for the WWW



HTTP Requests



GET / POST / PUT / DELETE

eg. www.google.co.th

eg. User-agent, Content-type

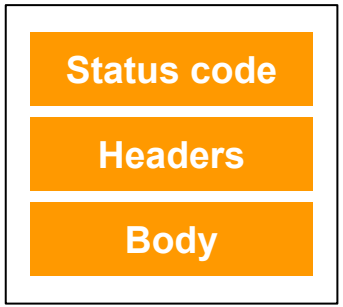
Additional data

Query strings

- Data can be included as part of a URL instead of inside the request body
- A query string comes after the path and is indicated by ?

➡ https://twitter.com/search?q=data+science&lang=th

HTTP Responses



eg. 200, 404, 503, ...

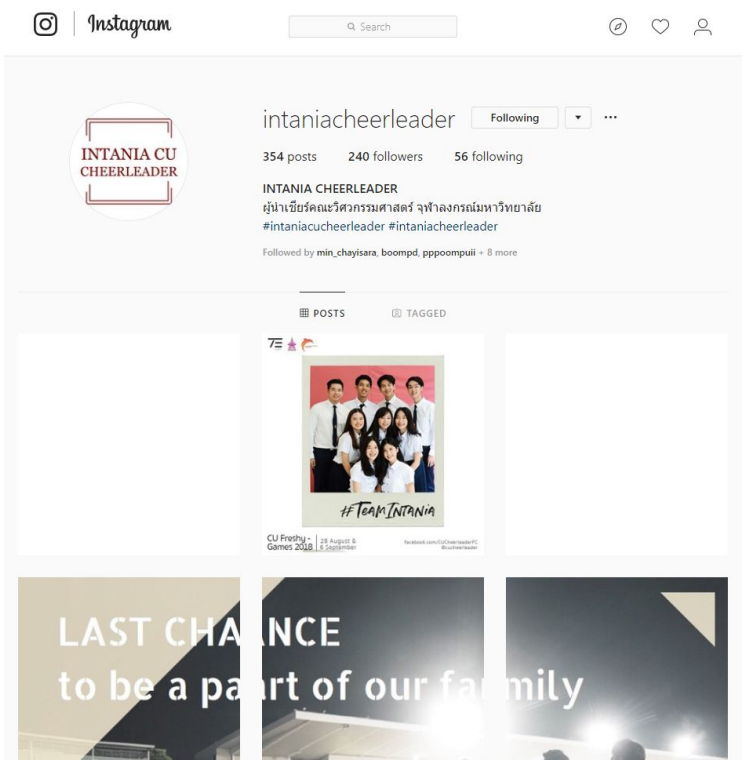
eg. Set-cookie, Last-modified

Additional data

HTTP in action!

- Open a new tab in your web browser
- Go to Developer Tools and select Network tab
- Enter https://www.instagram.com/intaniacheerleader/

HTTP in action!



▾ General

Request URL: https://instagram.fbkk2-4.fna.fbcdn.net/vp/f67cc7600a9b6421c4fa6dc76c8f11f4/5C35FF19/t51.2885-15/e35c180.0.719.719/s320x320/35353464_219510902023358_6962022676919484416_n.jpg

Request Method: GET

Status Code: 200 (from disk cache)

Remote Address: 27.123.18.160:443

Referrer Policy: no-referrer-when-downgrade

▾ Response Headers

access-control-allow-origin: *

cache-control: max-age=1209600, no-transform

content-length: 14156

content-type: image/jpeg

date: Fri, 31 Aug 2018 16:08:11 GMT

expires: Sat, 01 Sep 2018 14:58:34 GMT

last-modified: Thu, 28 Jun 2018 08:24:28 GMT

status: 200

timing-allow-origin: *

x-fb-config-version-elb-prod: 364

x-fb-config-version-flb-prod: 216

x-fb-config-version-olb-prod: 357

▾ Request Headers

⚠ **Provisional headers are shown**

Referer: https://www.instagram.com/intaniacheerleader/?hl=en

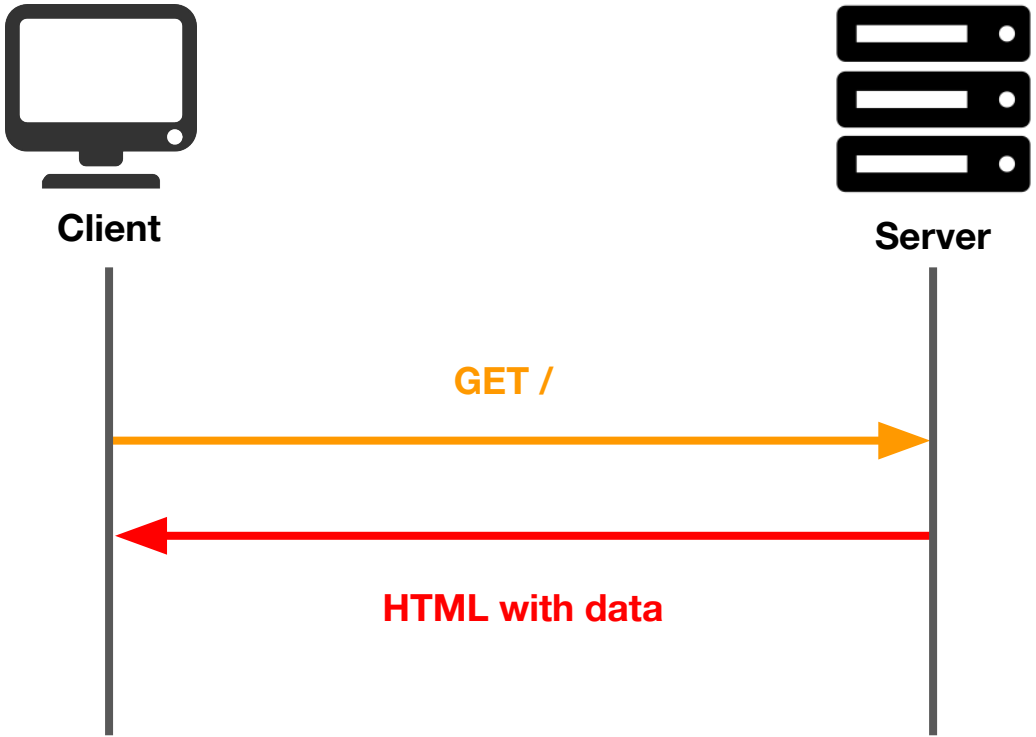
User-Agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/68.0.3440.106 Safari/537.36

AJAX Websites

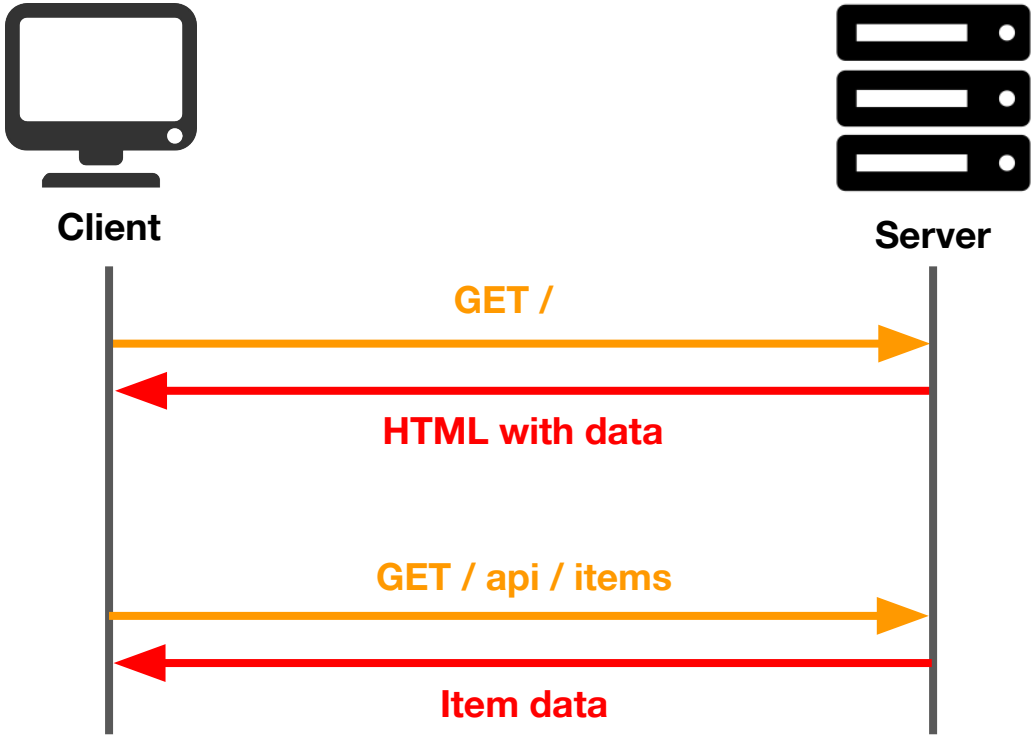
AJAX = Asynchronous JavaScript and XML

- AJAX enables web pages to be updated asynchronously
- Data are typically requested through APIs

Static Websites



AJAX Websites



Beyond collectiong data

- Google Directions API
  Get directions + estimated travel time
- IBM Watson Translation API
  Translate text into another language
- FacePlusPlus API
  Detect and locate detects human faces within an image
- And many more!

Workshop 2.3 : APIs

03-api.ipynb

Assignment 1 : Genie records and BNK48

Assignment_1.ipynb