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

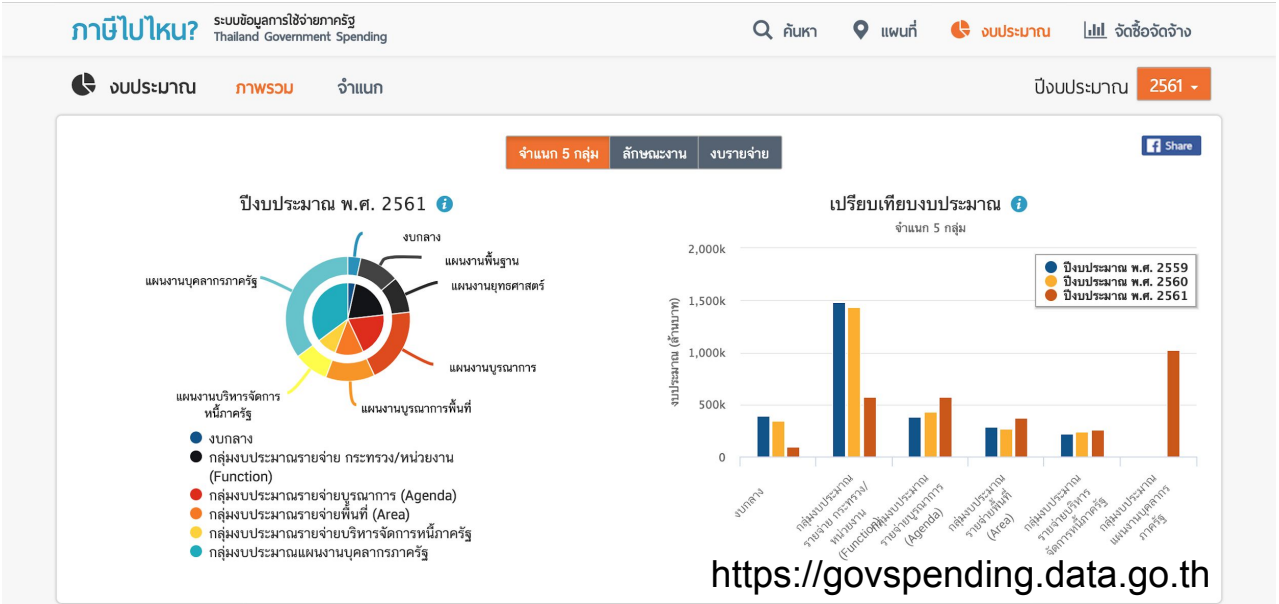
Facebook
more than 60 million active business Pages

Twitter
500 million tweets per day

Instagram
80 million photos uploaded every day

Data Collection

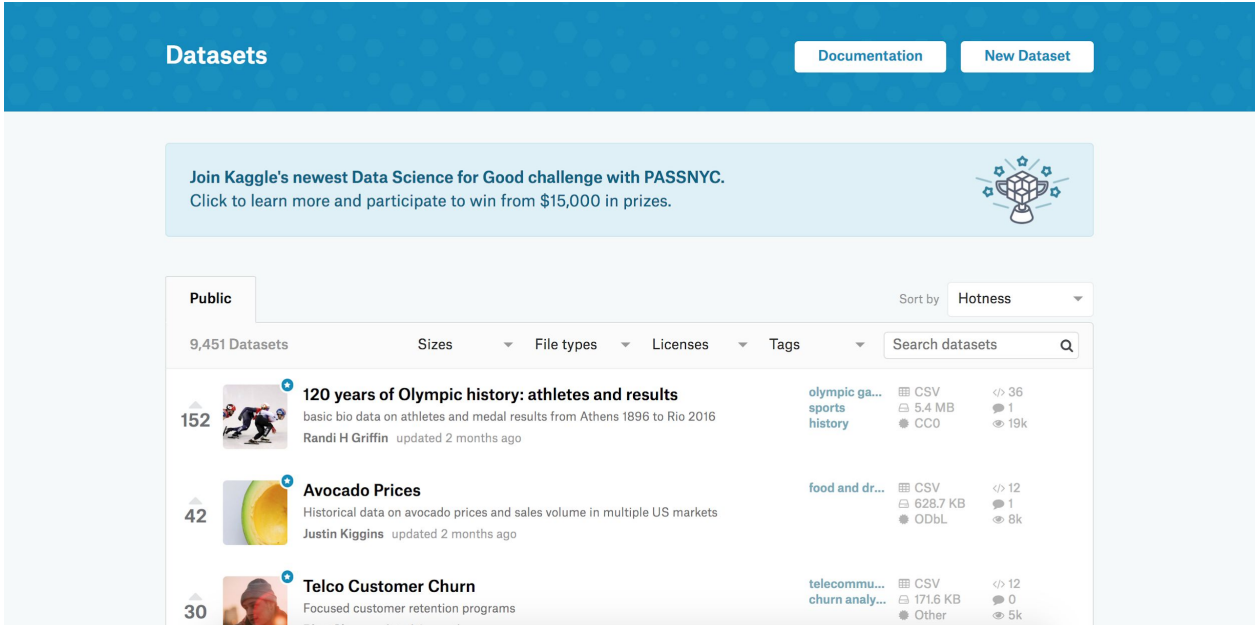
Open data



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

Data Collection

Kaggle



TOPICS

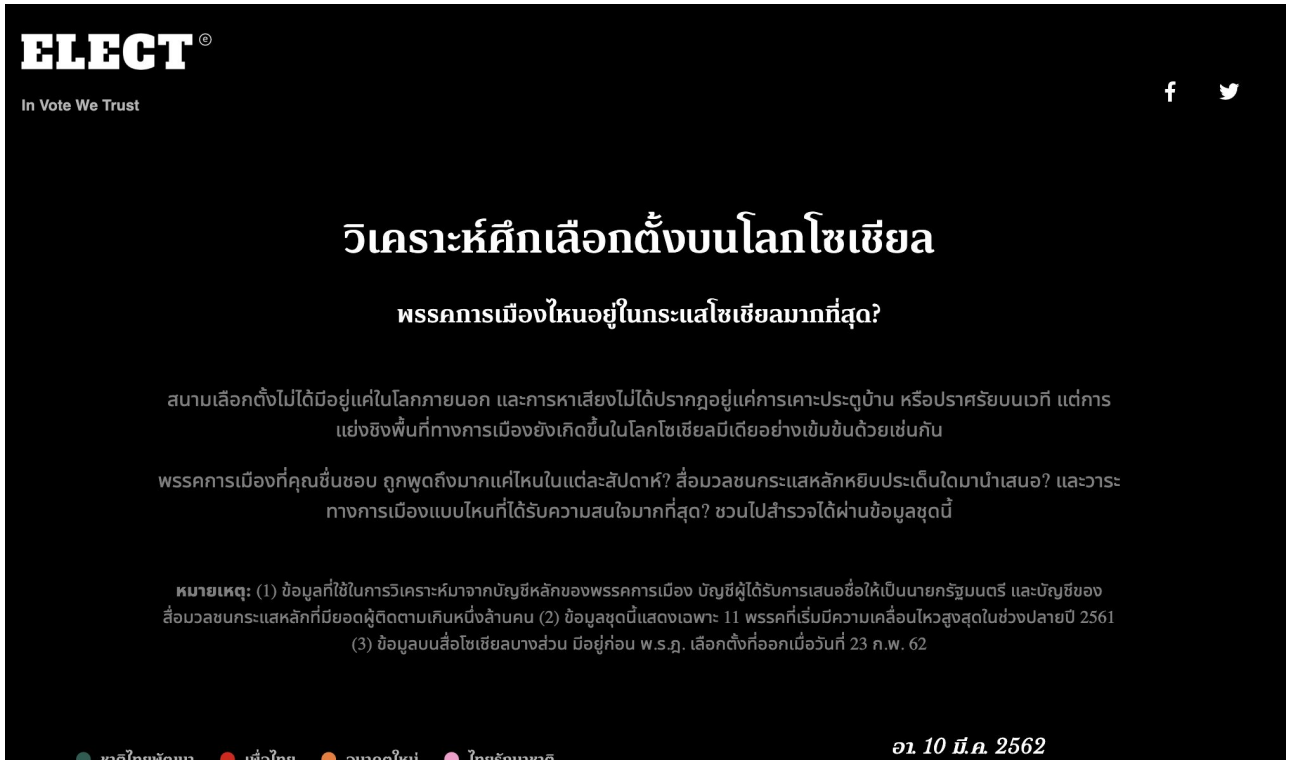
- Web Scraping
- Beautiful soup (next week)

Web Scraping

What is web scraping ?

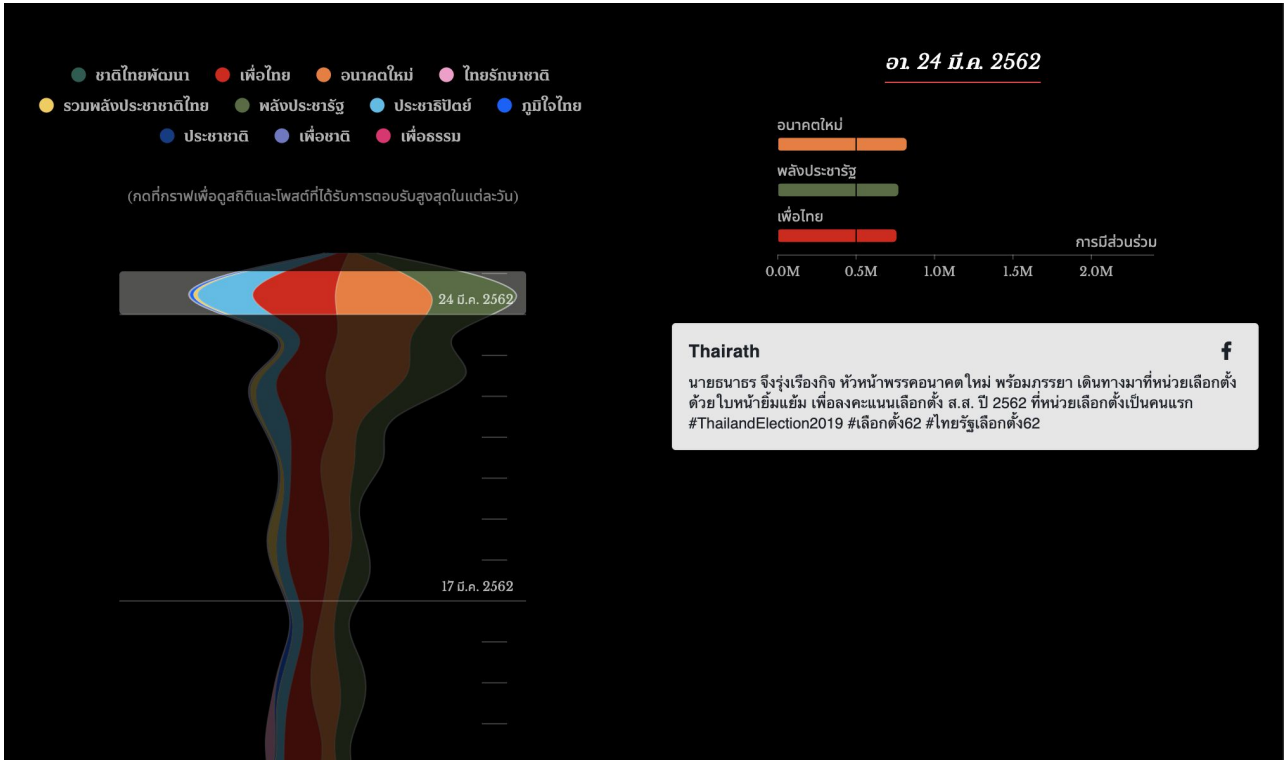
- A process of extracting information from websites
- It usually refers to an automated program that simulates a person viewing a website
- The process involves downloading a web page, parsing and extracting information from it, and store the target information in a proper format

Use case



https://elect.in.th/social-listening/

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 เดินทางมาที่หน่วยเลือกตั้งด้วยใบหน้ายิ้มแย้ม เพื่อลงคะแนนเลือกตั้ง ส.ส. ปี 2562 ที่หน่วยเลือกตั้งเป็นคนแรก #ThailandElection2019 #เลือกตั้ง62 #ไทยรัฐเลือกตั้ง62

Web Scraping

Ethics

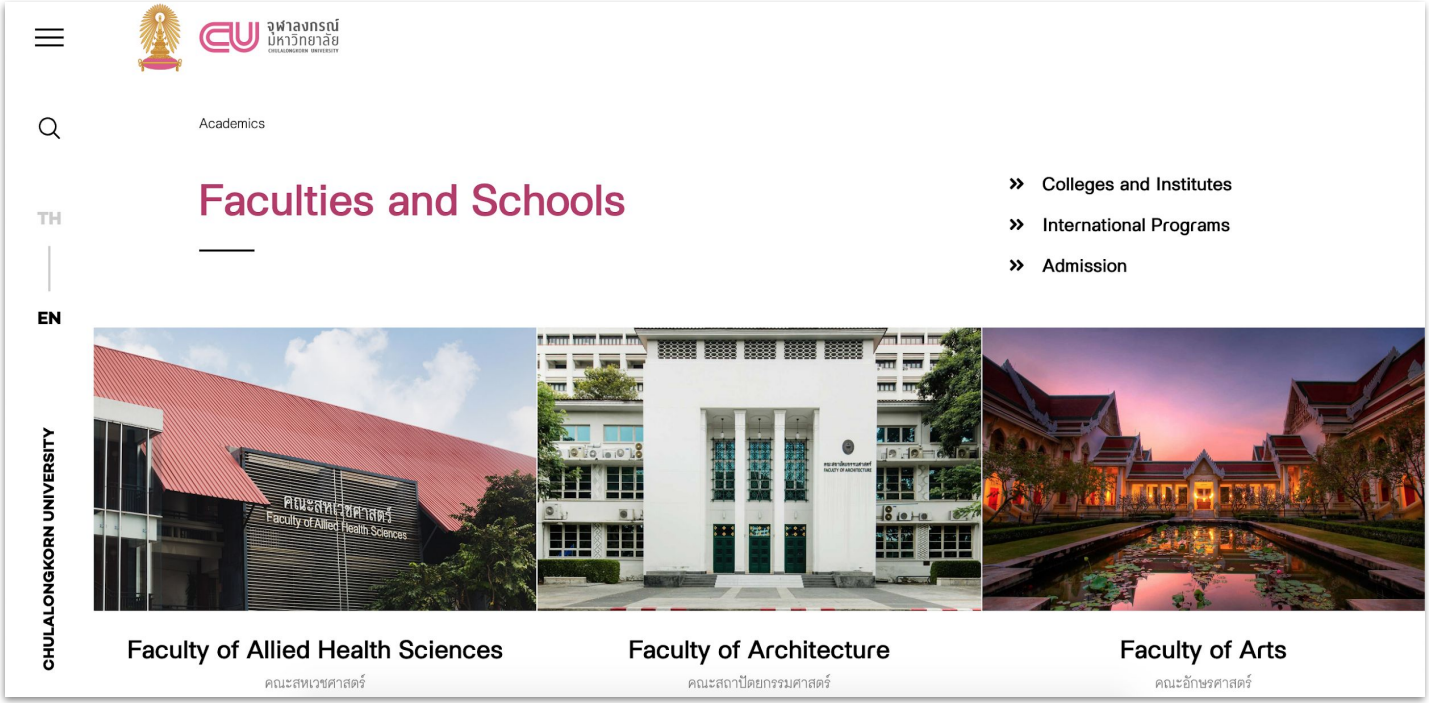
- Always check a website's **Terms and Conditions**
- Publishing the scraped data might violate copyright laws
- Act like a human - make requests at a reasonable rate
- Check the robots.txt file.

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:

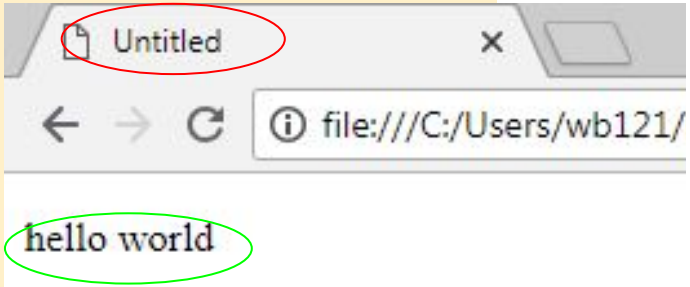
Web Scraping



HTML Essentials

HTML (Hypertext Markup Language)

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

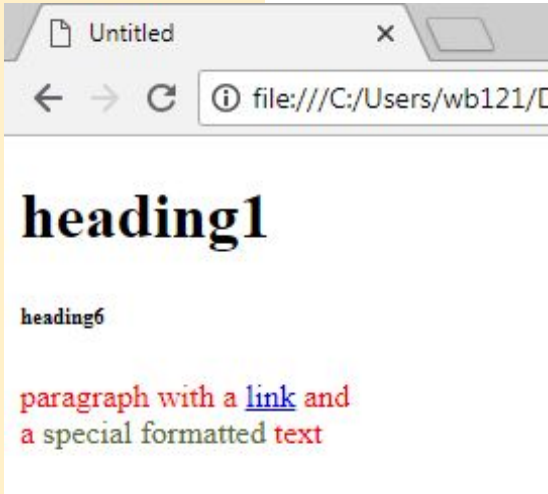


Every statement has opening < > and ending </ >

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>
```

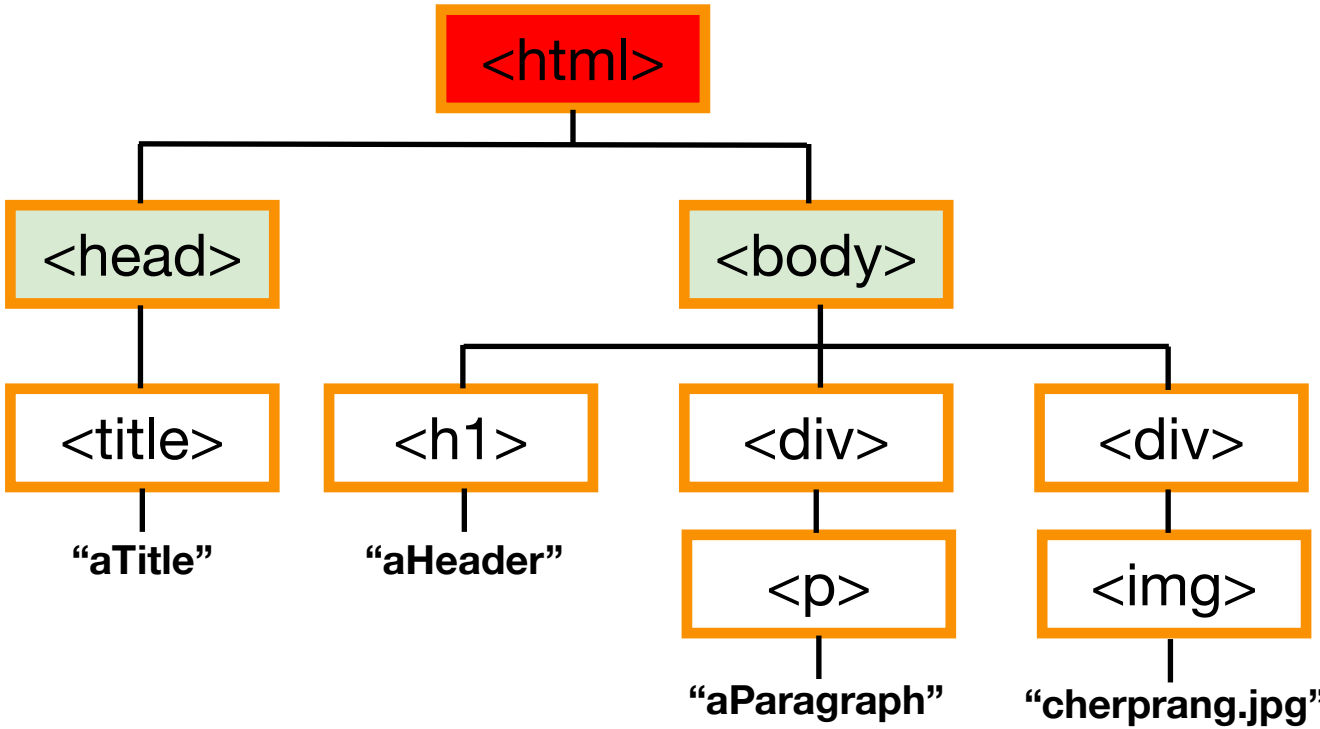
HTML Attributes

- **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://comprogchula.github.io/

- 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

Useful things to know

str.find

Find location of text

```
txt = “This is a pen. That is a pencil.”
ind = “01234567890123456789012345678901”

x = txt.find(“pen”)
print(x)
>10

x = txt.find(“pen”,11)
print(x)
>25

x = txt.find(“pens”)
print(x)
>-1
```

str.replace

Find location of text

```
txt = “This is a pen. That is a pencil.”

x = txt.replace(“pen”,“ben”)
print(x)
>This is a ben. That is a bencil.

x = txt.replace(“pen”,”ben”,1)
print(x)
>This is a ben. That is a pencil.
```

open

Prepares a file for reading/writing

```
fin = open( “asdf.txt”, “r”)  # open for reading
line = fin.readline() # read a line
for line in fin: # read until end of file
  ...
fin.close()
```

open

Prepares a file for reading/writing

```
fout = open( “asdf.txt”, “w”)  # open for writing
fout.write(“something”)
fout.close()
```

When writing binary files,
add “b”
เช่น “rb” “wb”

urllib

Read a url

```
import urllib
import urllib.request as urq
url = 'https://www.chula.ac.th/en/academics/faculties-and-schools'
html = str(urq.urlopen(url).read().decode('utf-8'))
```

เข้าเว
บไซต์

อ่าน
html

แปลงภาษาไทย

read and write images

- ขั้นตอนการอ่านและบันทึกไฟล์ภาพ

1. อ่านภาพจากลิงค์

- d = url.urlopen([ลิงค์ของภาพ])

---

2. สร้างไฟล์พร้อมระบุตำแหน่งที่จะเก็บไฟล์ภาพ

- l = open([ระบุตำแหน่งที่จะเก็บภาพ])

---

3. บันทึกข้อมูลภาพไปยังตำแหน่งที่เก็บไฟล์ตามที่ระบุไว้ในข้อ (2.)

- l.write(d.read())

---

4. ปิดไฟล์

- l.close()

Next week : beautifulsoup