

# **LECTURE 4**

# **MINING WEB**

# **CONTENT I**

**LEK HSIANG HUI**

# **OUTLINE**

**Introduction to Web Content Mining**

**Web Basics**

**Techniques for performing Web Content Mining**

**Extracting Content from HTML Source**

# INTRODUCTION TO WEB MINING



Introduction to  
Web Content  
Mining

Web Basics

Techniques for  
performing  
Web Content  
Mining

Extracting  
Content from  
HTML Source

# WEB MINING

**Web mining is concerned with mining data from the web which can then be transformed into knowledge**

**2 main aspects we will be focus on**

- Mining Web Content
- Data mining techniques which are more specific to web content (e.g. recommender systems)

# MINING WEB CONTENT

Data is usually not available for us to download as a spreadsheet

We might also want to aggregate multiple data sources together

- E.g. Suppose we have the location information (address) from food review sites (yelp/dianping)
- Might want to augment this information with Google Maps/Baidu Map (which contains other info like amenities) to create a dataset suitable for doing predictive modeling

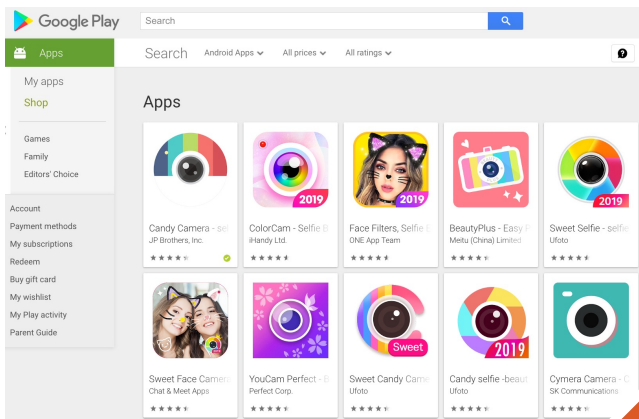


# MINING WEB CONTENT

Knowledge on how to acquire data in a **programmatic** manner is very important

- Not realistic to manually copy and paste all the time
- Not always easy to acquire data off the web
- ... but there are some techniques to do this

# EXAMPLE



## Mining Google Play store Apps Statistics

|   | App   | Category       | Rating | Reviews | Size | Installs    | Type | Price | Content Rating | Genres                    | Last Updated | Current Ver        | Android Ver  |
|---|---|----------------|--------|---------|------|-------------|------|-------|----------------|---------------------------|--------------|--------------------|--------------|
| 0 | Photo Editor & Candy Camera & Grid & ScrapBook    | ART_AND_DESIGN | 4.1    | 159     | 19M  | 10,000+     | Free | 0     | Everyone       | Art & Design              | 7-Jan-18     | 1.0.0              | 4.0.3 and up |
| 1 | Coloring book moana                               | ART_AND_DESIGN | 3.9    | 967     | 14M  | 500,000+    | Free | 0     | Everyone       | Art & Design;Pretend Play | 15-Jan-18    | 2.0.0              | 4.0.3 and up |
| 2 | U Launcher Lite – FREE Live Cool Themes, Hide ... | ART_AND_DESIGN | 4.7    | 87510   | 8.7M | 5,000,000+  | Free | 0     | Everyone       | Art & Design              | 1-Aug-18     | 1.2.4              | 4.0.3 and up |
| 3 | Sketch - Draw & Paint                             | ART_AND_DESIGN | 4.5    | 215644  | 25M  | 50,000,000+ | Free | 0     | Teen           | Art & Design              | 8-Jun-18     | Varies with device | 4.2 and up   |
| 4 | Pixel Draw - Number Art Coloring Book             | ART_AND_DESIGN | 4.3    | 967     | 2.8M | 100,000+    | Free | 0     | Everyone       | Art & Design;Creativity   | 20-Jun-18    | 1.1                | 4.4 and up   |

# APPLICATIONS

|   | App   | Category       | Rating | Reviews | Size | Installs    | Type | Price | Content Rating | Genres                    | Last Updated | Current Ver        | Android Ver  |
|---|---|----------------|--------|---------|------|-------------|------|-------|----------------|---------------------------|--------------|--------------------|--------------|
| 0 | Photo Editor & Candy Camera & Grid & ScrapBook    | ART_AND_DESIGN | 4.1    | 159     | 19M  | 10,000+     | Free | 0     | Everyone       | Art & Design              | 7-Jan-18     | 1.0.0              | 4.0.3 and up |
| 1 | Coloring book moana                               | ART_AND_DESIGN | 3.9    | 967     | 14M  | 500,000+    | Free | 0     | Everyone       | Art & Design;Pretend Play | 15-Jan-18    | 2.0.0              | 4.0.3 and up |
| 2 | U Launcher Lite – FREE Live Cool Themes, Hide ... | ART_AND_DESIGN | 4.7    | 87510   | 8.7M | 5,000,000+  | Free | 0     | Everyone       | Art & Design              | 1-Aug-18     | 1.2.4              | 4.0.3 and up |
| 3 | Sketch - Draw & Paint                             | ART_AND_DESIGN | 4.5    | 215644  | 25M  | 50,000,000+ | Free | 0     | Teen           | Art & Design              | 8-Jun-18     | Varies with device | 4.2 and up   |
| 4 | Pixel Draw - Number Art Coloring Book             | ART_AND_DESIGN | 4.3    | 967     | 2.8M | 100,000+    | Free | 0     | Everyone       | Art & Design;Creativity   | 20-Jun-18    | 1.1                | 4.4 and up   |

Further manipulate the data to gain insights of the mobile app market



| Type                | Reviews     |          |
|---------------------|-------------|----------|
|                     | Free        | Paid     |
| Category            |             |          |
| ART_AND_DESIGN      | 1712274.0   | 2166.0   |
| AUTO_AND_VEHICLES   | 1159503.0   | 4163.0   |
| BEAUTY              | 396240.0    | NaN      |
| BOOKS_AND_REFERENCE | 21957273.0  | 1796.0   |
| BUSINESS            | 13907683.0  | 46869.0  |
| COMICS              | 3383276.0   | NaN      |
| COMMUNICATION       | 815378046.0 | 84214.0  |
| DATING              | 7288124.0   | 3154.0   |
| EDUCATION           | 39561141.0  | 34645.0  |
| ENTERTAINMENT       | 59168145.0  | 10009.0  |
| ...                 | ...         | ...      |
| PERSONALIZATION     | 88896483.0  | 449657.0 |
| PHOTOGRAPHY         | 213285996.0 | 230654.0 |
| PRODUCTIVITY        | 113945254.0 | 171721.0 |
| SHOPPING            | 115040738.0 | 484.0    |
| SOCIAL              | 621241180.0 | 242.0    |
| SPORTS              | 70679534.0  | 150635.0 |
| TOOLS               | 273013107.0 | 171937.0 |

Number of reviews between Free and Paid Apps for different category

|                     |          | Rating       | Reviews |
|---------------------|----------|--------------|---------|
| Category            |          |              |         |
| ART_AND_DESIGN      | 4.358065 | 2.637600e+04 |         |
| AUTO_AND_VEHICLES   | 4.190411 | 1.369019e+04 |         |
| BEAUTY              | 4.278571 | 7.476226e+03 |         |
| BOOKS_AND_REFERENCE | 4.346067 | 9.506090e+04 |         |
| BUSINESS            | 4.121452 | 3.033598e+04 |         |
| COMICS              | 4.155172 | 5.638793e+04 |         |
| COMMUNICATION       | 4.158537 | 2.107138e+06 |         |
| DATING              | 3.970769 | 3.115931e+04 |         |
| EDUCATION           | 4.389032 | 2.538191e+05 |         |
| ENTERTAINMENT       | 4.126174 | 3.971688e+05 |         |
| ...                 | ...      | ...          |         |
| PERSONALIZATION     | 4.335987 | 2.279238e+05 |         |
| PHOTOGRAPHY         | 4.192114 | 6.373631e+05 |         |
| PRODUCTIVITY        | 4.211396 | 2.691438e+05 |         |
| SHOPPING            | 4.259664 | 4.424662e+05 |         |
| SOCIAL              | 4.255598 | 2.105903e+06 |         |
| SPORTS              | 4.223511 | 1.844536e+05 |         |
| TOOLS               | 4.047411 | 3.240629e+05 |         |

Average rating and number of reviews for different category



# WEB CRAWLER

A **web crawler** (or **web spider**) is a computer program that systematically access the World Wide Web (WWW) to download web data

They are typically used for the purpose of **web indexing** (i.e. to power a search engine)

- Bingbot
- Googlebot
- etc

# WEB SCRAPER

A **web scraper** is also a computer program that can be used for downloading web data

- However, they are more targeted at extracting specific data
- E.g. Web scraper for:  
Google Play Store reviews, Amazon Products, etc

# WEB CRAWLER VS WEB SCRAPER

**Web crawlers tend to extract the entire content of a page as a whole** (e.g. entire page's html)

- Lesser work needed as compared to web scrapers
- However, the data is quite raw and requires more processing (for the purpose of analytics)


**Web scrapers are more specialized**

- They are specially designed to extract only certain parts of the page
- Usually, multiple parts in a single page

# WEB CRAWLER EXAMPLE

Input:  
`https://sws.comp.nus.edu.sg/`

Output:  
Text and HTML of  
each page



|   | url  | text  | html  |
|---|--|---|---|
| 0 | <code>https://sws.comp.nus.edu.sg/</code>                  | NUS School of Computing Summer Workshop\n\n\n...  | b'<!DOCTYPE html>\n<html style="font-size: 16p... |
| 1 | <code>https://sws.comp.nus.edu.sg/FAQ.html</code>          | FAQ\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n...            | b'<!DOCTYPE html>\n<html style="font-size: 16p... |
| 2 | <code>https://sws.comp.nus.edu.sg/Home.html</code>         | NUS School of Computing Summer Workshop\n\n\n...  | b'<!DOCTYPE html>\n<html style="font-size: 16p... |
| 3 | <code>https://sws.comp.nus.edu.sg/Application.html</code>  | Application   NUS School of Computing Summer W... | b'<!DOCTYPE html>\n<html style="font-size: 16p... |
| 4 | <code>https://sws.comp.nus.edu.sg/Contact.html</code>      | Contact   NUS School of Computing Summer Works... | b'<!DOCTYPE html>\n<html style="font-size: 16p... |
| 5 | <code>https://sws.comp.nus.edu.sg/Facilities.html</code>   | Facilities   NUS School of Computing Summer Wo... | b'<!DOCTYPE html>\n<html style="font-size: 16p... |
| 6 | <code>https://sws.comp.nus.edu.sg/Testimonials.html</code> | Testimonials   NUS School of Computing Summer ... | b'<!DOCTYPE html>\n<html style="font-size: 16p... |
| 7 | <code>https://sws.comp.nus.edu.sg/Courses.html</code>      | Clusters and Courses   NUS School of Computing... | b'<!DOCTYPE html>\n<html style="font-size: 16p... |
| 8 | <code>https://sws.comp.nus.edu.sg/Structure.html</code>    | Structure   NUS School of Computing Summer Wor... | b'<!DOCTYPE html>\n<html style="font-size: 16p... |

# WEB SCRAPER EXAMPLE

Input:  
**IMDB Scraper**  
(created specially for  
scrapping IMDB content)



Output:  
**Each column is  
one piece of useful  
information**

Scraper has to be  
crafted to extract  
these information

|   | title                   | rating |
|---|-------------------------|--------|
| 0 | Avengers: Endgame       | 8.4    |
| 1 | Captain Marvel          | 6.8    |
| 2 | Avengers: Infinity War  | 8.4    |
| 3 | Guardians of the Galaxy | 8.0    |
| 4 | The Avengers            | 8.0    |
| 5 | Avengers: Age of Ultron | 7.3    |

# CHALLENGES IN WEB CONTENT MINING



SINGAPORE

200 people evacuated from Tangs Plaza in Orchard Road after fire breaks out

15 Jun 2019 10:18PM

(Updated: 15 Jun 2019 10:29PM)



Bookmark



Singapore

## 200 people evacuated from Tangs Plaza in Orchard Road after fire breaks out



Employees clearing water outside Tangs Plaza on Jun 15, 2019. (Photo: Johannes Tjendro)

SINGAPORE: Around 200 people were evacuated from Tangs Plaza in Orchard Road on Saturday (Jun 15) evening after a fire broke out in the department store.

Singapore Civil Defence Force (SCDF) said it was alerted to the incident at 320 Orchard Road at about 8.40pm.

Advertisement



The all-new Audi Q8.  
Discover the 8th dimension.

Learn more

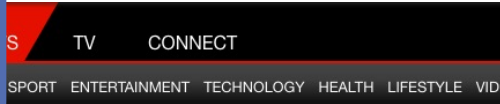


The fire, which involved electrical wiring, was extinguished by sprinklers before firefighters arrived, SCDF added.

Handling dynamic content (e.g. ads) which might affect extraction accuracy

# CHALLENGES IN WEB CONTENT MINING

Jan 2017



## TOP STORIES



Daughter of South Korea's 'Rasputin' arrested in Denmark  
4 hours ago in ASIA PACIFIC



Myanmar detains police over Rohingya abuse video  
4 hours ago in ASIA PACIFIC



Islamic State claims responsibility for Istanbul attack, gunman remains at large  
6 hours ago in WORLD



Inmates beheaded in Brazil jail riot, 60 killed  
2 hours ago in WORLD

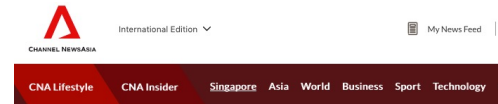


In Iraq, Hollande says IS battle prevents attacks at home  
45 minutes ago in WORLD



Obama to deliver farewell address in Chicago on Jan 10  
9 hours ago in WORLD

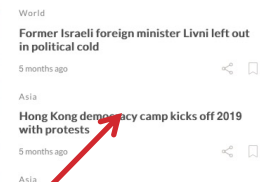
Jan 2019



## TOP stories



## LATEST news

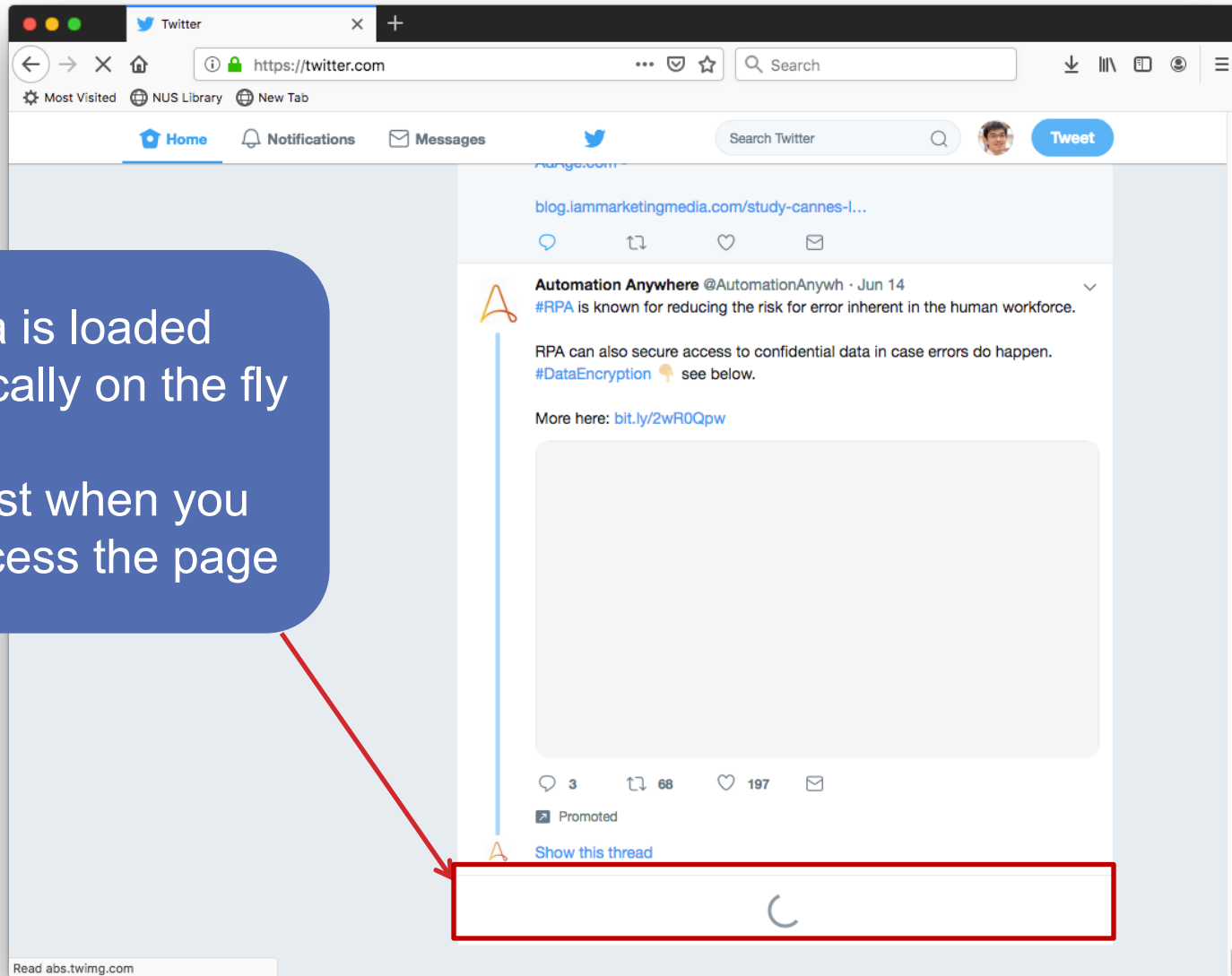


Each time when the site layout changes, potentially, the web scrapers might break

# CHALLENGES IN WEB CONTENT MINING

Data is loaded  
dynamically on the fly

Not just when you  
first access the page





# WEB BASICS

Introduction to  
Web Content  
Mining

Web Basics

Techniques for  
performing  
Web Content  
Mining

Extracting  
Content from  
HTML Source

# HTML

The web is powered by **HTML**

- **HTML** = **H**yper**T**ext **M**arkup **L**anguage
- Scripting language used to create webpages
- Defines the format, layout, resources in a webpage
- Used together with:
  - **JavaScript**: Additional features/logic
  - **Cascading Style Sheets (CSS)**: Styling

# HTML

```
<html>

  <head>

    <title>SWS3023</title>

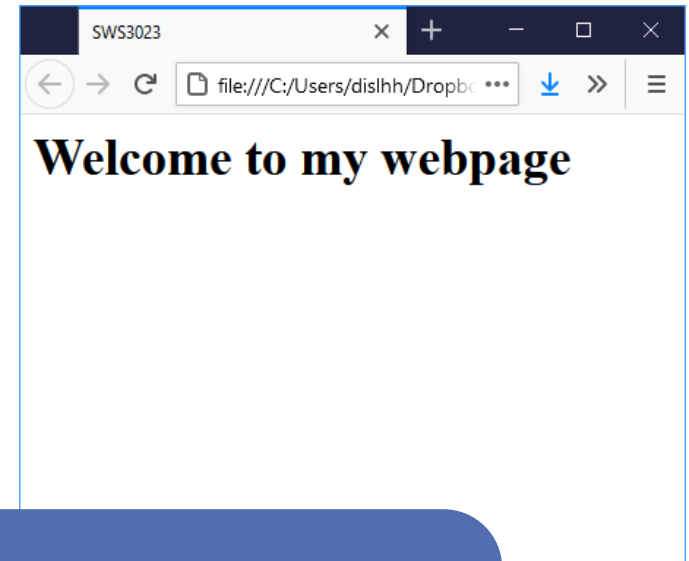
  </head>

  <body>

    <h1>Welcome to my webpage</h1>

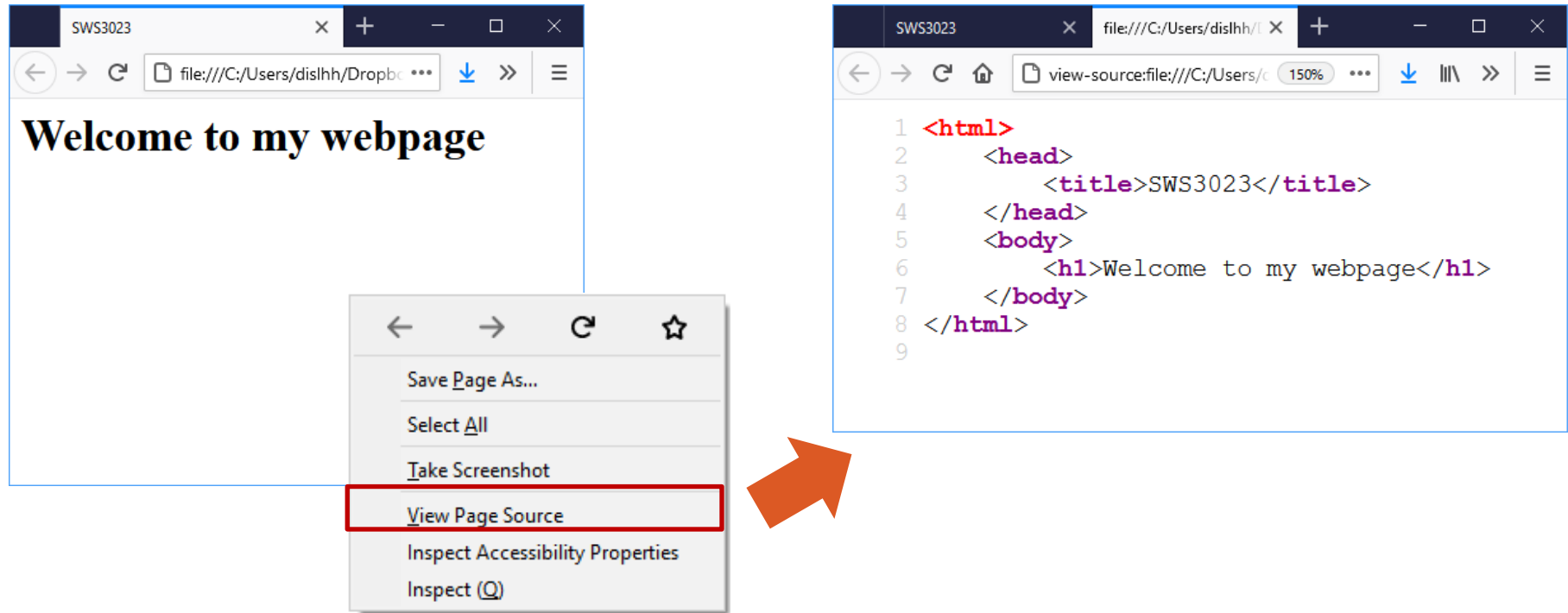
  </body>

</html>
```



HTML is just a text document  
that defines both the content  
and the layout of the page

# HOW TO GET THE HTML SOURCE OF A PAGE?



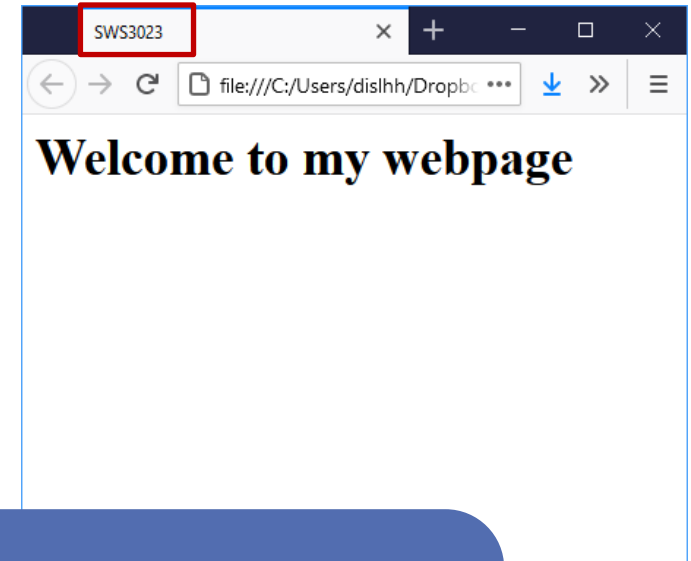
Right click >>  
View Page Source

# HTML

```
<html>

<head>
  <title>SWS3023</title>
</head>

<body>
  <h1>Welcome to my webpage</h1>
</body>
</html>
```



`<title>...</title>`  
defines the title of the page

# HTML

```
<html>
```

```
<head>
```

```
<title>SWS3023</title>
```

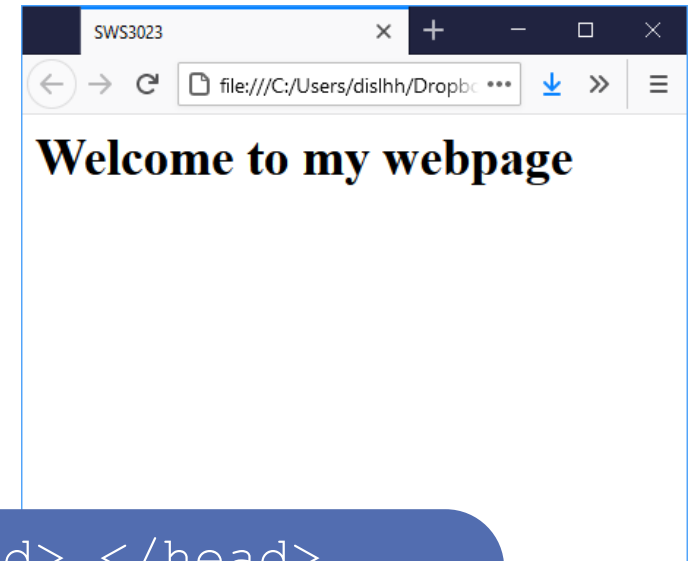
```
</head>
```

```
<body>
```

```
<h1>Welcome to my webpage</h1>
```

```
</body>
```

```
</html>
```



`<head>...</head>`  
section contains information  
about the page that is usually  
not visible (with the exception  
of `<title>...</title>`)

# HTML

```
<html>

  <head>

    <title>SWS3023</title>

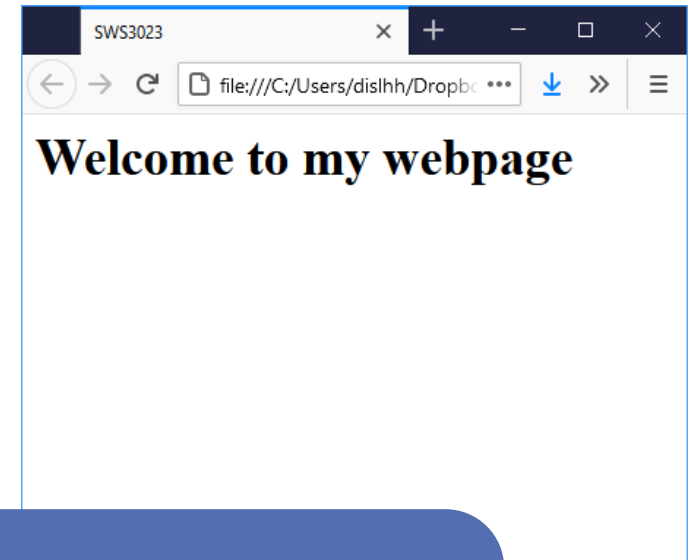
  </head>

  <body>

    <h1>Welcome to my webpage</h1>

  </body>

</html>
```



`<body>...</body>`  
section contains presentational  
content for the page

# HTML

```
<html>

<head>

  <title>SWS3023</title>

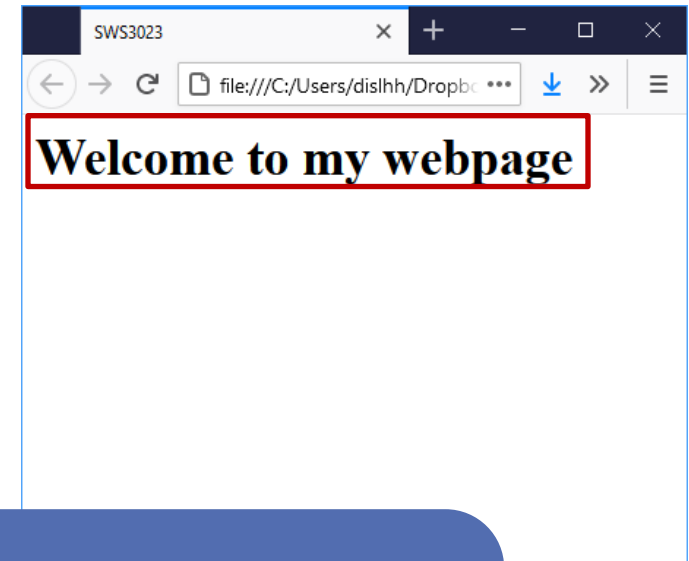
</head>

<body>

  <h1>Welcome to my webpage</h1>

</body>

</html>
```



`<h1>...</h1>`  
allows you to add a heading  
text (similar to Microsoft Word)



# HTML

```
<html>
```

```
<head>
```

```
<title>SWS3023</title>
```

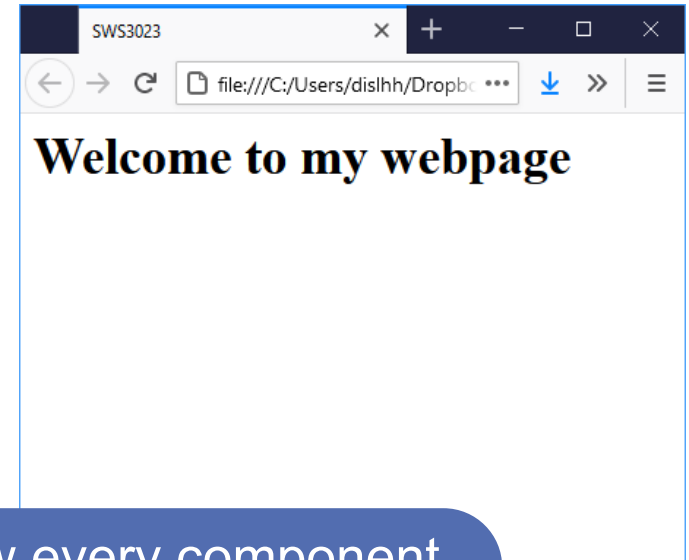
```
</head>
```

```
<body>
```

```
<h1>Welcome to my webpage</h1>
```

```
</body>
```

```
</html>
```



Notice how every component of a webpage is defined with a starting tag (e.g. `<title>`) and an ending tag (e.g. `</title>`)

# HTML

```
<html>

<head>

  <title>SWS3023</title>

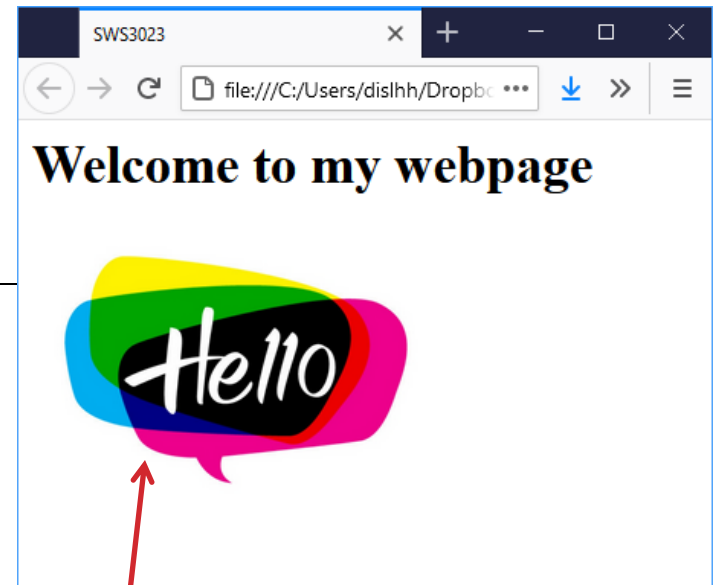
</head>

<body>

  <h1>Welcome to my webpage</h1>

  
```

src and width are attributes of <img> tag  
Can use " or ' to enclose attribute values (as long as they are matching quotes)



Some elements have a single self-closing tag instead (notice the /> at the end)

# BODY SECTION

Many other types of tags can be added to the body section:

- Paragraph (`<p>...</p>`)
- Heading (`<h1>...</h1>`, `h1`, `h2`, ... `h6`)
- Formatting tags
  - Bold (`<b>...</b>`)
  - Italics (`<i>...</i>`)
  - Underline (`<u>...</u>`)
- Links
- Images
- List
- Table
- Div, Span

# CHARACTERISTICS OF HTML

**HTML is case insensitive:**

- `<HTML>` is the same as `<html>`

**Web browsers do not recognize line breaks (or multiple empty spaces)**

- i.e.

I want to have  
another line

will be seen on the browser as a single line:

I want to have another line

**Line break tags (`<br>`) must be used to denote line breaks**

# TECHNIQUES FOR PERFORMING WEB CONTENT MINING

Introduction to  
Web Content  
Mining

Web Basics

Techniques for  
performing  
Web Content  
Mining

Extracting  
Content from  
HTML Source

# HOW DOES A WEB CRAWLER WORK?

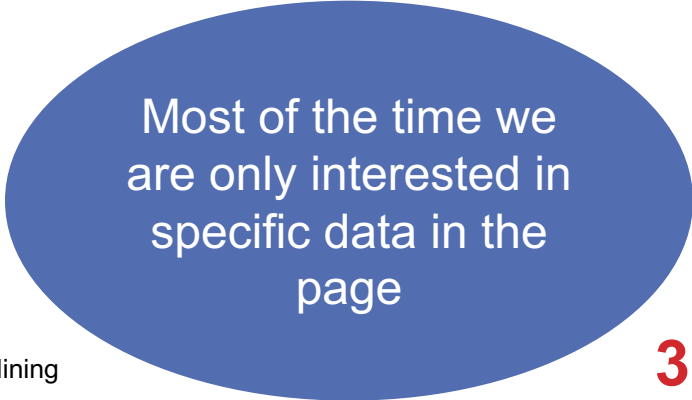
## General Approach:

- Maintain a queue (or stack)
- The queue is first populated with a list of URLs (seed) to start crawling from
- while (there are items in the queue)
  - Remove the first item (URL)
  - Extract a list of hyperlinks on the page and add these URLs to the queue
  - Download the page (if the page is “needed”)

# PERFORMING WEB SCRAPING

## General Approach:

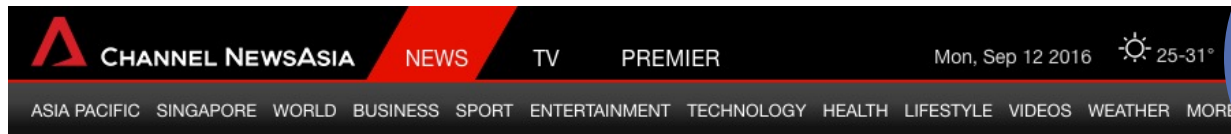
- Maintain a queue (or stack)
- The queue is first populated with a list of URLs (seed) to start crawling from
- while (there are items in the queue)
  - ...
  - Extract data into a database (“web scraping”)



Most of the time we are only interested in specific data in the page

# PERFORMING WEB SCRAPING

Queue contains the front page of a website as a seed URL



**SGSECURE**  
STAY ALERT • STAY UNITED • STAY STRONG

Be part of the SGSecure movement,  
visit [www.sgsecure.sg](http://www.sgsecure.sg)  
#SGSecure #StayAlert #StayUnited #StayStrong

## TOP STORIES



China, Russia to stage military drills in South China Sea

29 minutes ago in **ASIA PACIFIC**



Paralympics: Singapore's Theresa Goh wins bronze in 100m breaststroke SB4 final

1 hour ago in **SPORT**



PM Lee wishes Muslims Selamat Hari Raya Haji

1 hour ago in **SINGAPORE**



11 new cases of Zika confirmed Sunday; 10 have no known links to existing clusters

12 hours ago in **SINGAPORE**



Clinton has pneumonia, was dehydrated at 9/11 event

5 hours ago in **WORLD**



North Korea ready to conduct another nuclear test: Report

2 hours ago in **ASIA PACIFIC**



## HAPPENING NOW

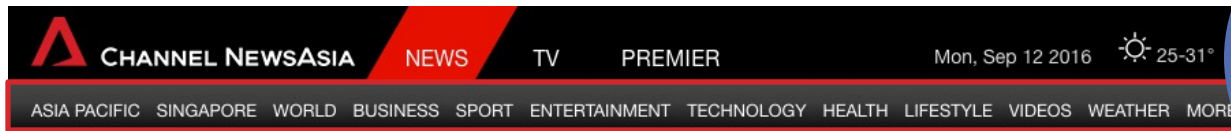
Latest News refreshed just now

- 12:10 **BUSINESS** Fed's Kashkari sees cure for slow growth in immigration, tax reform
- 12:05 **ENTERTAINMENT** Miss Arkansas crowned Miss America 2017 in pageant's 96th year
- 11:45 **ENTERTAINMENT** Box Office: 'Sully' flies high to US\$35.5 million US opening
- 11:44 **ASIA PACIFIC** China, Russia to stage military drills in South China Sea
- <SPONSORED> Knight Rider alert: How Audi light technology saves lives
- <ADV> One Day in Taiwan? Don't Miss the Night Markets
- 11:06 **SINGAPORE** StarHub releases iPhone 7 prices



# PERFORMING WEB SCRAPING

Extract the list of hyperlinks and add them back to the queue



**SGSECURE**  
STAY ALERT • STAY UNITED • STAY STRONG

Be part of the SGSecure movement,  
visit [www.sgsecure.sg](http://www.sgsecure.sg)  
#SGSecure #StayAlert #StayUnited #StayStrong

## TOP STORIES



China, Russia to stage military drills in South China Sea

29 minutes ago in ASIA PACIFIC



Paralympics: Singapore's Theresa Goh wins bronze in 100m breaststroke SB4 final

1 hour ago in SPORT



PM Lee wishes Muslims Selamat Hari Raya Haji

1 hour ago in SINGAPORE



11 new cases of Zika confirmed Sunday; 10 have no known links to existing clusters

12 hours ago in SINGAPORE



Clinton has pneumonia, was dehydrated at 9/11 event

5 hours ago in WORLD



North Korea ready to conduct another nuclear test: Report

2 hours ago in ASIA PACIFIC



## HAPPENING NOW

Latest News refreshed just now

12:10  
BUSINESS

Fed's Kashkari sees cure for slow growth in immigration, tax reform

12:05  
ENTERTAINMENT

Miss Arkansas crowned Miss America 2017 in pageant's 96th year

11:45  
ENTERTAINMENT

Box Office: 'Sully' flies high to US\$35.5 million US opening

11:44  
ASIA PACIFIC

China, Russia to stage military drills in South China Sea



<SPONSORED> Knight Rider alert: How Audi light technology saves lives



<ADV> One Day in Taiwan? Don't Miss the Night Markets

11:06  
SINGAPORE

StarHub releases iPhone 7 prices

# PERFORMING WEB SCRAPING

Extract useful content from it such as text or images

SPORT

## Paralympics: Singapore's Theresa Goh wins bronze in 100m breaststroke SB4 final

By Philip Goh, Sports Editor Posted 12 Sep 2016 07:21 Updated 12 Sep 2016 10:28



816 Email More

SINGAPORE: After more than a decade of trying, Singapore's Theresa Goh has finally made it to the podium at the Paralympics.

On Monday (Sep 12) morning, Singapore time, the 29-year-old Goh touched home in third place in the women's SB4 100m breaststroke final in Rio de Janeiro.

Her time of 1min 55.55s was a second off the Asian record she had set in the morning heats when she qualified second fastest behind world record holder Sarah Louise Rung of Norway.

HAPPENING

Latest News

12:10  
BUSINESS

12:05  
ENTERTAINMENT

11:45  
ENTERTAINMENT

11:44  
ASIA PACIFIC

11:06  
SINGAPORE

20:00

million US open

China, Russia to stage military drills in South China Sea

<SPONSORED> Knight Rider alert: How Audi light technology saves lives

<ADV> One Day in Taiwan? Don't Miss the Night Markets

StarHub releases iPhone 7 prices



Manulife

11:05  
ASIA PACIFIC

10:40  
ASIA PACIFIC

10:34  
SINGAPORE

South Korea's Yonhap says US delays sending B-1 bomber to Korean peninsula due to weather

Japan's Abe says North Korea nuclear tests "absolutely unacceptable"

PM Lee wishes Muslims Selamat Hari Raya Haji

# TECHNIQUES FOR WEB SCRAPING

**The following are some of the techniques for doing web scraping:**

- Extracting content from HTML source
- Extracting content using a HTML parser
- Web Scraping using APIs
- Scraping using an actual browser/headless browser

# EXTRACTING CONTENT FROM HTML SOURCE

Introduction to  
Web Content  
Mining

Web Basics

Techniques for  
performing  
Web Content  
Mining

Extracting  
Content from  
HTML Source

# STRING MATCHING

One possible way to do web scraping is to download the HTML source and perform string-based matching

- This is often achieved through the use of **regular expressions**
- A regular expression is a sequence of characters which defines a search pattern for matching string or for extracting string

# STRING MATCHING WEB SCRAPER

```
import re

html = "<html><body>Web Mining<br/>Web mining is concerned  
with mining data from the web the web which can then be  
transformed into knowledge</body></html>"

pattern = "^<html><body>(.*)<br/>(.*)</body></html>$"

#group 1
title = re.search(pattern, html).group(1)
title

#group 2
description = re.search(pattern, html).group(2)
description
```

# STRING MATCHING WEB SCRAPER

```
import re
```

Import the regular expression library (re)

```
html = "<html><body>Web Mining<br/>Web mining is concerned  
with mining data from the web the web which can then be  
transformed into knowledge</body></html>"
```

```
pattern = "^<html><body>(.*)<br/>(.*)</body></html>$"
```

```
#group 1
```

```
title = re.search(pattern, html).group(1)
```

```
title
```

```
#group 2
```

```
description = re.search(pattern, html).group(2)
```

```
description
```

# STRING MATCHING WEB SCRAPER

```
import re
```

```
html = "<html><body>Web Mining  
with mining data from the web  
transformed into knowledge</body></html>"
```

```
pattern = "^<html><body>(.*?)<br/>(.*?)</body></html>$"
```

```
#group 1
```

```
title = re.search(pattern, html).group(1)  
title
```

```
#group 2
```

```
description = re.search(pattern, html).group(2)  
description
```

search() takes 2 parameters :  
pattern, string





pattern is a regular expression (regex)  
The pattern is satisfied if it matches this regex

# STRING MATCHING WEB SCRAPER

```
import re
```

```
html = "<html><body>Web Mining<br/>Web mining is concerned  
with mining data from the web the web which can then be  
transformed into knowledge</body></html>"
```

```
pattern = "^<html><body>(.*)<br/>(.*)</body></html>$"
```

Notice some special characters: ^ (.\* ) \$

^ indicates the start of the string

( ) defines a new capturing group

. indicates any character

\* indicates 0 or more times

.\* indicates a sequence of any character occurring 0 or more times

\$ indicates the end of the string

pattern is a regular expression (regex)  
The pattern is satisfied if it matches this regex

# STRING MATCHING WEB SCRAPER

```
import re

html = "<html><body>Web Mining<br/>Web mining is concerned  
with mining data from the web the web which can then be  
transformed into knowledge</body></html>"

pattern = "^<html><body>(.*)<br/>(.*)</body></html>$"

#group 1
title = re.search(pattern, html).group(1)
title
```

```
#group 2
des
des
```

^ indicates the start of the string  
\$ indicates the end of the string

By using ^ and \$, there will either be one match or no match (there won't be multiple matches)

First ( ) is the **first** capturing group

Second ( ) is the **second** capturing group

# STRING MATCHING WEB SCRAPER

```
import re

html = "<html><body>Web Mining<br/>Web mining is concerned  
with mining data from the web the web which can then be  
transformed into knowledge</body></html>"

pattern = "^<html><body>(.*)<br/>(.*)</body></html>$"

#group 1
title = re.search(pattern, html).group(1)
title

#group 2
description = re.search(pattern, html).group(2)
description
```

# STRING MATCHING WEB

group(1) will retrieve the contents of the first capturing group

```
import re

html = "<html><body>Web Mining<br/>Web mining is concerned  
with mining data from the web the web which can then be  
transformed into knowledge</body></html>"

pattern = "^<html><body>(.*?)<br/>(.*?)</body></html>$"

#group 1
title = re.search(pattern, html).group(1)
title

#group 2
description = re.search(pattern, html).group(2)
description
```

# STRING MATCHING WEB

group(2) will retrieve the contents of the second capturing group

```
import re

html = "<html><body>Web Mining<br/>Web mining is concerned  
with mining data from the web the web which can then be  
transformed into knowledge</body></html>"

pattern = "^<html><body>(.*?)<br/>(.*?)</body></html>$"

#group 1
title = re.search(pattern, html).group(1)
title

#group 2
description = re.search(pattern, html).group(2)
description
```

# STRING MATCHING WEB

group() will match the entire string that matches the pattern

```
import re
```

```
html = "<html><body>Web Mining<br/>Web mining is concerned  
with mining data from the web the web which can then be  
transformed into knowledge</body></html>"
```

```
pattern = "^<html><body>(.*?)<br/>(.*?)</body></html>$"
```

```
#group 1
```

```
title = re.search(pattern, html).group(1)
```

```
title
```

```
#group 2
```

```
description = re.search(pattern, html).group(2)
```

```
description
```

# REGULAR EXPRESSION (REGEX) REFERENCES

<https://www.regular-expressions.info/python.html>

<https://www.dataquest.io/wp-content/uploads/2019/03/python-regular-expressions-cheat-sheet.pdf>

# EXTRACTING MATCHING PATTERNS

```
import re

html2 = "<html><body>
<ul><li>johndoe@gmail.com</li><li>janedoe@gmail.com</li>
<li>Robin</li></ul></body></html>"

pattern2 = "<li>(.*?)</li>"

list_values = re.findall(pattern2, html2)
list_values
```



Pattern to match any string that starts with `<li>` followed by (zero or more characters) and ends with `</li>`


## EXTRACTING MATCHING PATTERNS

```
import re

html2 = "<html><body>
  <ul><li>johndoe@gmail.com</li><li>janedoe@gmail.com</li>
  <li>Robin</li></ul></body></html>"

pattern2 = "<li>(.*?)</li>"

list_values = re.findall(pattern2, html2)
list_values
```



Pattern to match any string that starts with `<li>` followed by (zero or more characters) and ends with `</li>`

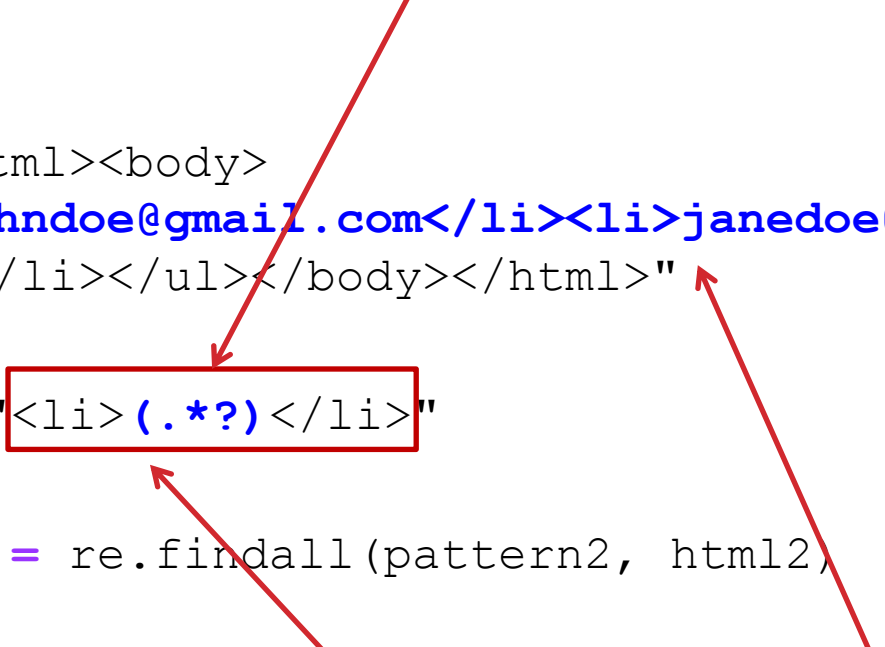
## EXTRACTING MATCHING PATTERNS

```
import re

html2 = "<html><body>
<ul><li>johndoe@gmail.com</li><li>janedoe@gmail.com</li>
<li>Robin</li></ul></body></html>"

pattern2 = "<li>(.*?)</li>"

list_values = re.findall(pattern2, html2)
list_values
```



Notice that there is a `?` which indicates that we are doing a non-greedy matching (3 matches in total)

`.*` will match greedily (i.e. highlighted section)

## EXTRACTING MAIL

**Question:** Suppose we only want to extract emails. How can we change the pattern to achieve this?

```
import re

html2 = "<html><body>
<ul><li>johndoe@gmail.com</li><li>janedoe@gmail.com</li>
<li>Robin</li></ul></body></html>"

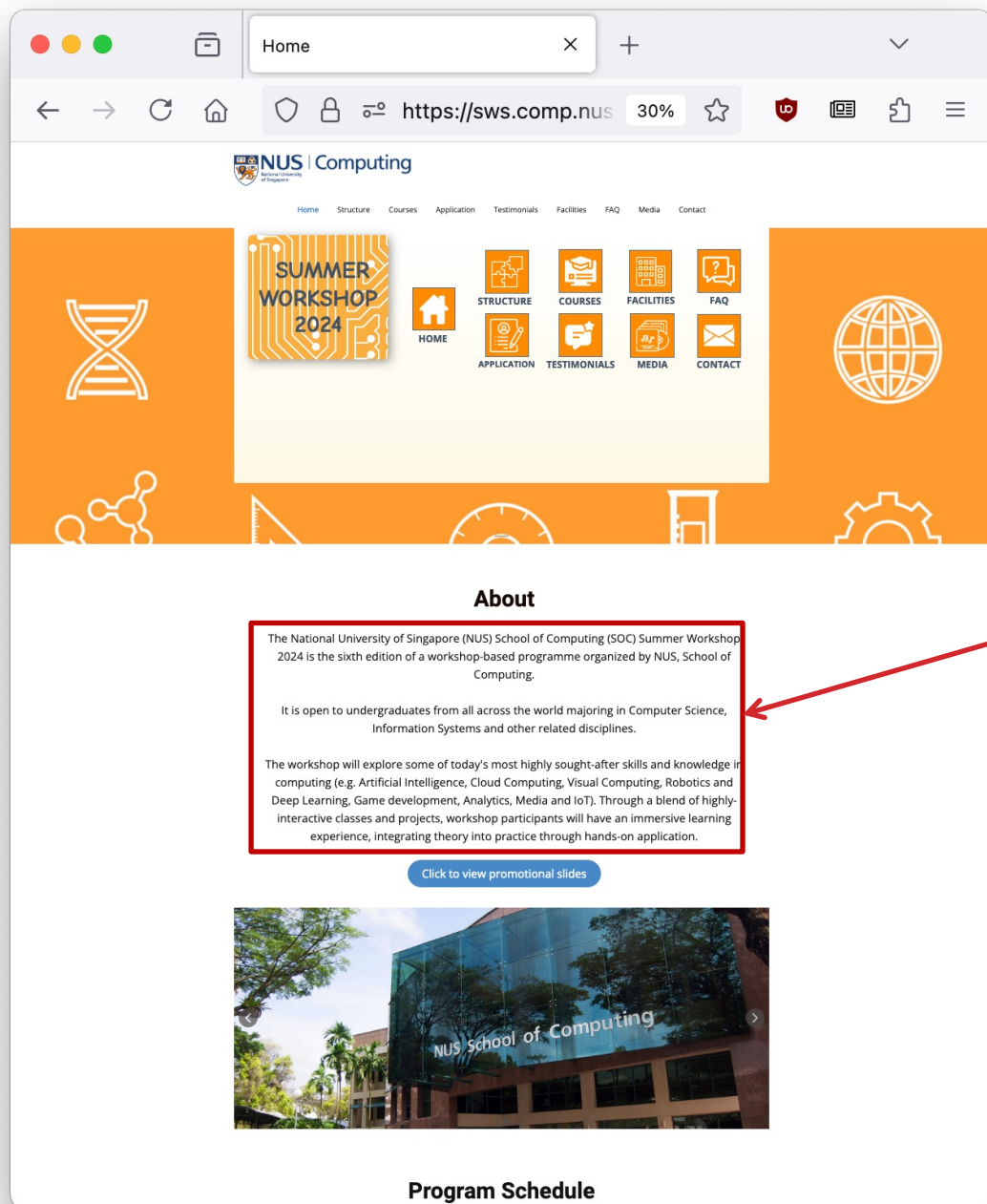
email_pattern = "<li>(.*?)</li>"

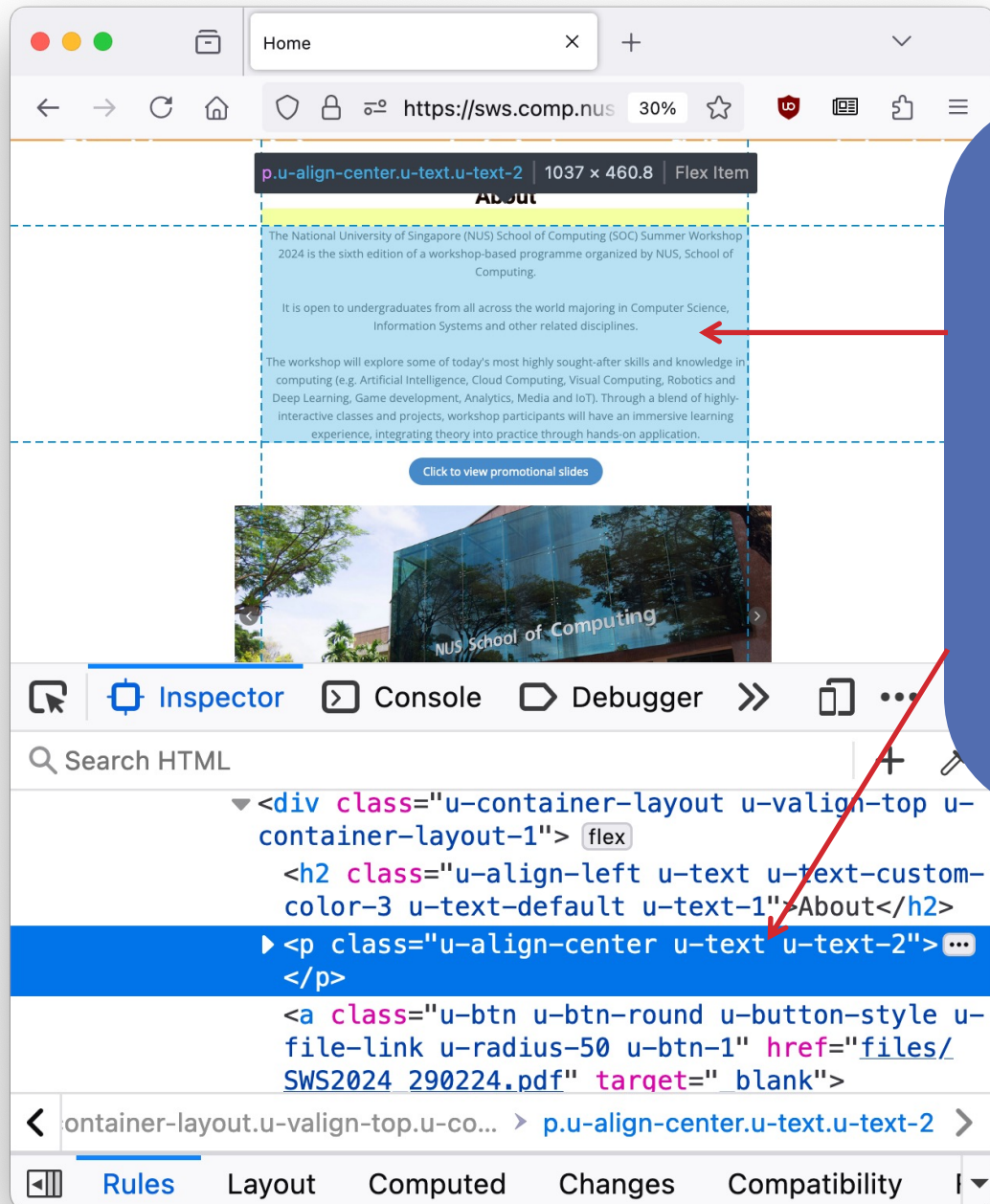
emails = re.findall(email_pattern, html2)
emails
```



Another example  
(Real website)

Suppose we want to  
extract the contents under  
the About





First figure out the rough position in the page using the web browser's **Inspect Element** feature

Then find it View Source window (because the browser might clean up the page slightly such as removing unnecessary spaces)

```
import re
import requests

page = requests.get("https://sws.comp.nus.edu.sg/")

html3 = page.content.decode("utf-8")

#. by default does not match for newline characters
#re.DOTALL - will make . match even for newline
pattern3 = '<p class="u-align-center u-text u-text-2">(.*?)</p>'
results = re.search(pattern3, html3, re.DOTALL)

extracted_text = results.group(1)

#remove some html
#\n - refers to group 1
extracted_text = re.sub("<br>", "\n", extracted_text)
extracted_text = re.sub("&nbsp;", " ", extracted_text)
extracted_text = re.sub("<.*?>(.*?)</.*?>", "\\1",
    extracted_text)
extracted_text = extracted_text.strip()
```

# **SUMMARY**

## **Introduction to Web Mining**

- Web Scraper, Web Crawler

## **Web Basics**

- HTML

## **Techniques for performing Web Content Mining**

- Extracting Content from HTML Source

# WHAT'S NEXT?

## Mining Web Content II