MINING WEB DATA

STQD6414 PERLOMBONGAN DATA

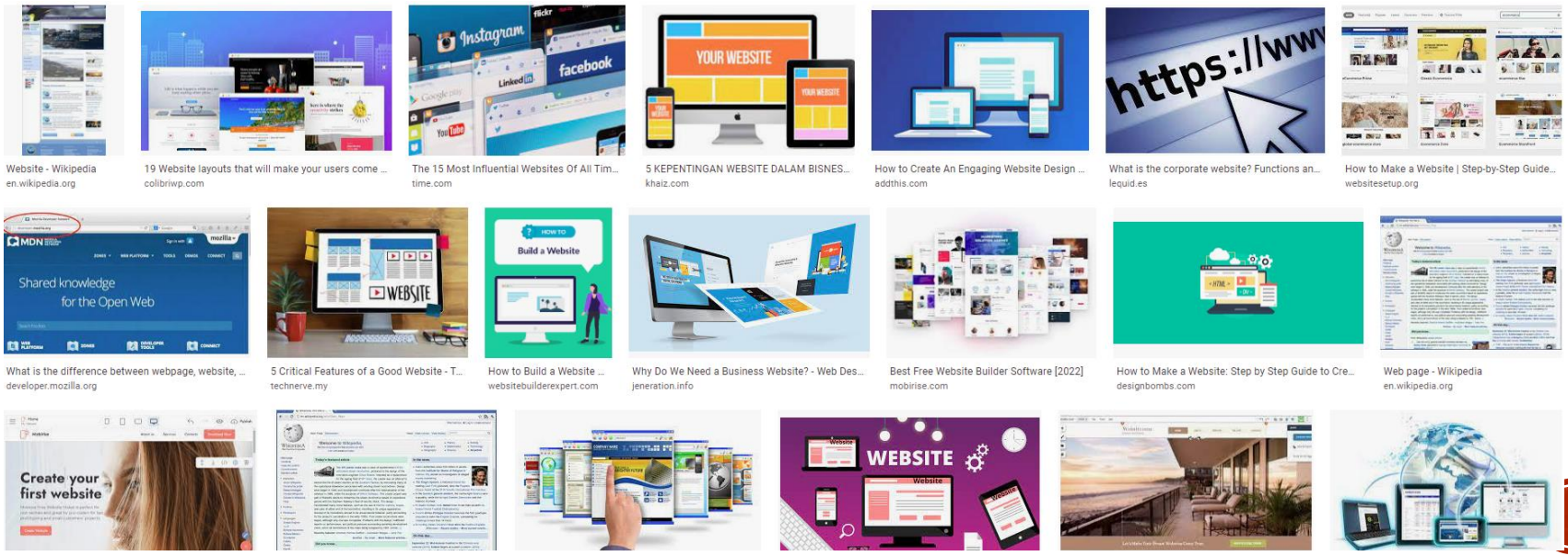
Assoc. Prof. Dr. Nurulkamal Masseran

Department of Mathematical Sciences

Universiti Kebangsaan Malaysia

INTRODUCTION:

- Nowadays, web is one of the biggest data sources for data mining analysis.
- Web mining aims to discover useful information or knowledge from the web hyperlink structure or website pages.
- Web data analysis require the knowledge of artificial intelligence, machine learning, statistics, pattern recognition, and data mining.
- Web data indicate the characteristics of heterogeneous, semistructured or unstructured data.



INTRODUCTION:

- Among the web mining techniques:

i) Web structure mining:

- This technique aims to find useful information or valuable structural summaries about the sites and pages from the hyperlinks between the web pages.

ii) Web content mining:

- This technique aims to extract useful information from contents on any particular web page.

iii) Web usage mining:

- This technique aims to discover user access patterns from web logs for the purpose of intrusion detection, fraud detect, and attempted break-in.

CHARACTERISTICS OF WEB DATA:

- The characteristics of web data are:

i) The information in web are heterogeneous. Any type of data can be contained in the Web. Either structured or unstructured data.

ii) Information on the Web is constantly changing.

iii) The amount of data in Web always growing.

iv) Vast amounts of information on the web is linked.

v) The data is noisy.

HYPERTEXT MARKUP LANGUAGE (HTML):

- In order to to scrape data from websites, we need to understand how the web pages are structured.
- The foundation for website structure is HTML.
- HTML organizes the web browser (browser) for the way the web page is displayed, the content in the web page, etc..
- Example: HTML

```
<html>
<head>
  <title>Page title</title>
</head>
<body>
  <h1 id='first'>A heading</h1>
  <p>Some text &amp;amp; <b>some bold text.</b></p>
  <img src='myimg.png' width='100' height='100'>
</body>
```

- Thus, the we need to understand the underlying HTML structure before we can scrape it.

BASIC STRUCTURE IN HTML:

- HTML has a hierarchical structure formed by elements which consist of:

i) a start tag (e.g. <tag>)

ii) optional attributes (id='first'),

iii) contents.

iv) end tag1 (like </tag>)

- The symbol of < and > are used for start and end tags.

ELEMENTS & ATTRIBUTES IN HTML:

- HTML element is defined by a starting tag. If the element contains other content, it ends with a closing tag.
- Example: is starting tag of a paragraph and is closing tag of the same paragraph.

- Some of the important HTML elements are:

i) HTML must have two main components: <head>, which contains document metadata like the page title, and <body>, which contains the content you see in the browser.

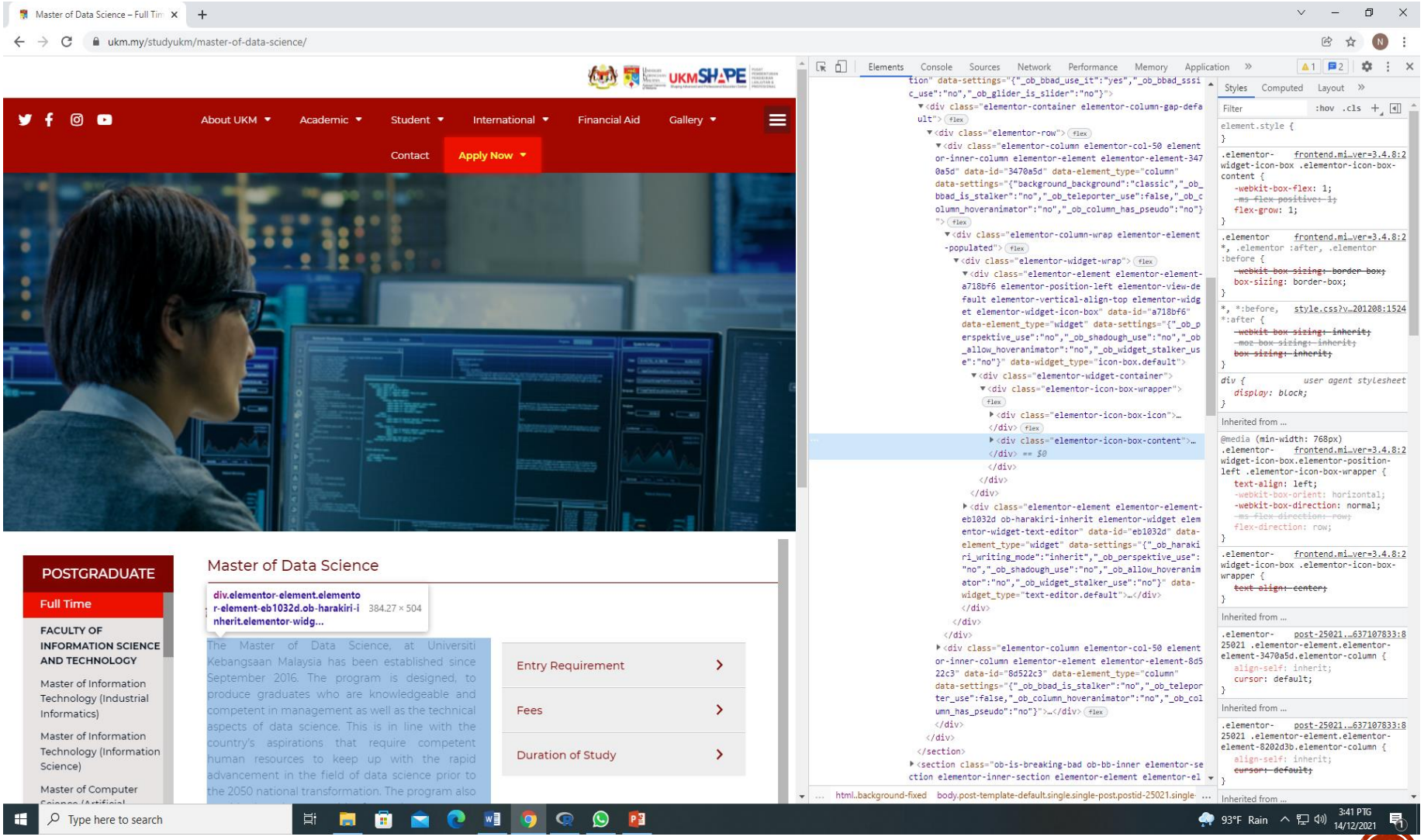
ii) Block tags like <h1> (heading 1), (paragraph), and <ol> (ordered list) form the overall structure of the page.

iii) Inline tags like <b> (bold), <i> (italics), and <a> (links) formats text inside block tags.

CSS & JAVASCRIPT

- HTML provide a content for a web page.
- However, the HTML content is just a plain text.
- Thus, to make the display of content in the website more attractive, a CSS and Javascript need to be integrated.
- CSS stand for Cascading Style Sheets.
- In other word, CSS is a language used to describe the formatting of a document written in HTML (XML, XTML and etc.)
- For example, CSS is useful for adding style such as; fonts, colors, and spacing into a web documents.
- On the other hand, Javascript is a language used to manage the behavior of a web page.

EXAMPLE: HTML, CSS & JAVASCRIPT



WEB SCRAPING:

- Web scraping is a technique for converting the data present in unstructured format (HTML tags) in the web to the structured format which can easily be accessed and used.
- In order to scrap the data from website, we need to know the hierarchical structure presented in the website.
- This hierarchical structure known as DOM (Document Object Model).
- DOM defines the logical structure of a document and the way it is accessed and manipulated.

- Apart from that, other important tool is Xpath.
- XPath refer to XML Path Language.
- It is a query language for selecting nodes from an XTML or XML document.

REFERENCES:

- Aydin, O. (2018). *R Web Scraping Quick Start Guide*. Packt Publisher.
- Khalil, S. (2021). *Rcrawler: Web Crawler and Scraper*. R package version 0.1.9-1.
- Liu, B. (2011). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer
- Munzert, M., Rubba, C., Meißner, P., Nyhuis, D. (2014). *Automated Data Collection With R : A Practical Guide To Web Scraping And Text Mining*. Wiley.
- Patel, J.M. (2020). *Getting Structured Data from the Internet: Running Web Crawlers/Scrapers on a Big Data Production Scale*. Apress Publisher
- Wickham, H. (2021). *rvest: Easily Harvest (Scrape) Web Pages*. R package version 1.0.2.

FINAL EXAM:

DATE: 22 JANUARY 2024 (MONDAY)

TIME: 9.00 AM - 12.00 PM

VENUE: GAMMA LABORATORY, DEPARTMENT OF MATHEMATICAL SCIENCES

GOOD LUCK!!!