PERLOMBONGAN DATA WEB

STQD6414 PERLOMBONGAN DATA

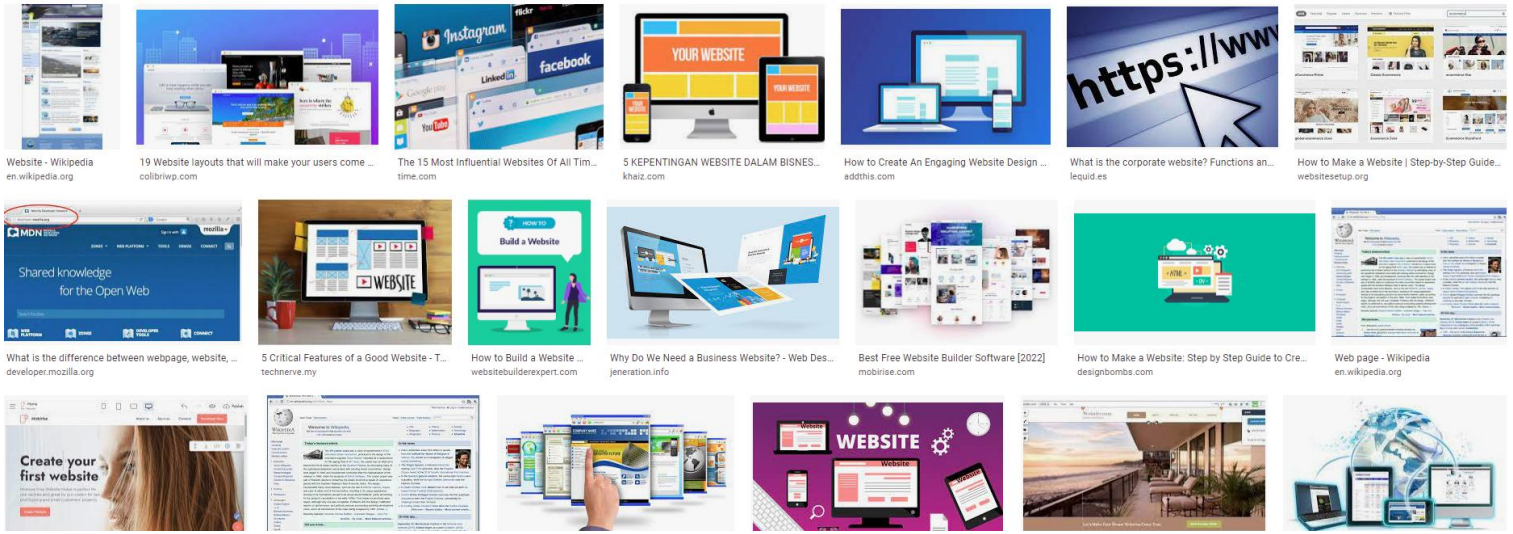
Assoc. Prof. Dr. Nurulkamal Masseran

Jabatan Sains Matematik

Universiti Kebangsaan Malaysia

PENGENALAN:

- Pada masa kini, web adalah salah satu sumber data terbesar untuk analisis perlombongan data.
- Perlombongan web bertujuan untuk mencari maklumat atau pengetahuan yang berguna daripada struktur hiperpautan web (*web hyperlink*) atau halaman laman web (*website pages*).
- Analisis data web memerlukan pengetahuan tentang kecerdasan buatan (AI), pembelajaran mesin, statistik, pengecaman corak dan perlombongan data.
- Data web mempunyai ciri-ciri data heterogen, separuh berstruktur atau tidak berstruktur.



PENGENALAN :

- Antara teknik perlombongan web:

i) Perlombongan struktur Web:

- Teknik ini bertujuan untuk mencari maklumat atau ringkasan struktur berkaitan tapak (*sides*) dan halaman (*pages*) daripada hiperpautan antara halaman web.

ii) Perlombongan kandungan Web:

- Teknik ini bertujuan untuk mengekstrak maklumat berguna daripada kandungan laman web tertentu.

iii) Perlombongan penggunaan Web:

- Teknik ini bertujuan untuk menyelidiki corak capaian pengguna log web untuk tujuan pengesanan pencerobohan, pengesanan penipuan dan percubaan pecah masuk (*break-in*).

CIRI-CIRI DATA WEB:

- Antara ciri-ciri data web ialah:

i) Maklumat dalam web adalah heterogen (*heterogeneous*). Sebarang bentuk jenis data boleh terkandung dalam Web. Sama ada data berstruktur atau tidak berstruktur.

ii) Maklumat di Web sentiasa berubah.

iii) Amaun data dalam Web sentiasa bertambah.

iv) Sebilangan besar maklumat di web mempunyai pautan.

v) Data mempunyai hingar (*noisy*).

BAHASA PENANDA HIPERTEKS (HTML):

- Untuk skrap (*scrape*) data daripada tapak web, kita perlu memahami bagaimana halaman web distrukturkan.
- Asas untuk struktur laman web ialah Bahasa Penanda Hiperteks (*Hypertext Markup Language, HTML*).
- HTML mengorganisasi pelayar web (*browser*) untuk cara paparan halaman web, kandungan dalam laman web, dan lain-lain.
- Contoh: 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>
```

- Oleh itu, kita perlu memahami struktur asas HTML sebelum kita boleh mengskrap sebarang data dari web.

STRUKTUR ASAS DALAM HTML:

- HTML mempunyai struktur hierarki yang dibentuk oleh elemen yang terdiri daripada:

i) tag mula (contoh: <tag>)

ii) Atribut-atribut pilihan (contoh: id='first'),

iii) Kandungan

iv) Tag akhir (contoh: like </tag>)

ELEMENS & ATRIBUT DALAM HTML:

- Elemen HTML ditakrifkan oleh tag permulaan. Jika elemen mengandungi kandungan lain, ia berakhir dengan tag penutup.
- Contoh: ialah tag permulaan perenggan dan ialah tag penutup perenggan yang sama.

- Antara elemen HTML yang penting ialah:

i) HTML mesti mempunyai dua komponen utama: <head>, yang mengandungi metadata dokumen seperti tajuk halaman, dan <body>, yang mengandungi kandungan yang anda boleh lihat dalam pelayar.

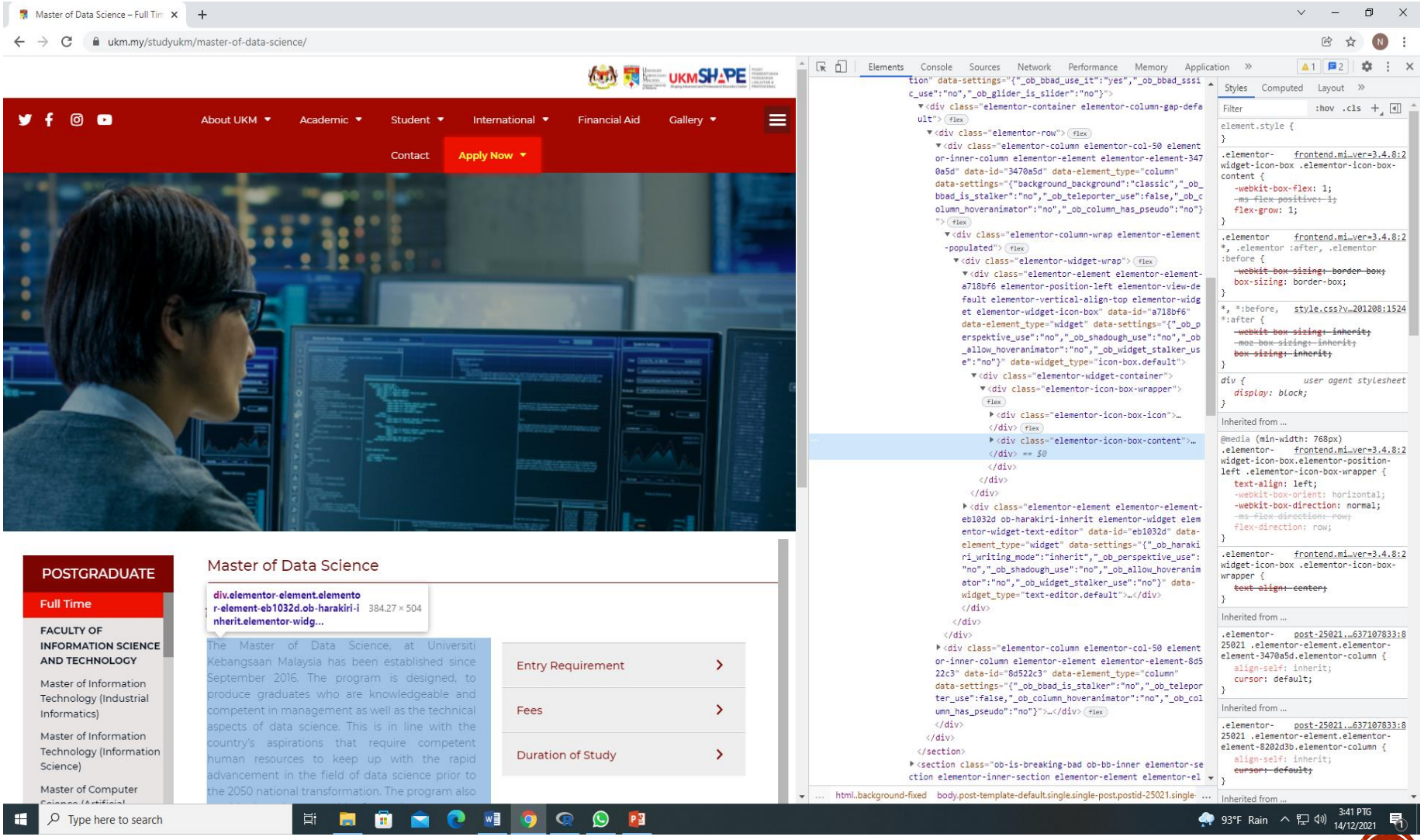
ii) Tag blok (*Block tags*) seperti <h1> (heading 1), (paragraph), dan <ol> (ordered list) yang membentuk struktur keseluruhan halaman.

iii) Tag sebaris (*Inline tags*) seperti <b> (bold), <i> (italics), dan <a> (links) yang memformat teks dalam tag blok.

CSS & JAVASCRIPT

- HTML menyediakan kandungan untuk halaman web.
- Walau bagaimanapun, kandungan HTML hanyalah teks biasa (*plain text*).
- Oleh itu, untuk menjadikan paparan kandungan dalam laman web lebih menarik, CSS dan Javascript perlu diintegrasikan.
- CSS merujuk kepada *Cascading Style Sheets*.
- Dengan kata lain, CSS ialah bahasa yang digunakan untuk menerangkan pemformatan dokumen yang ditulis dalam HTML (XML, XTML dan lain-lain.)
- Contoh: CSS digunakan untuk menambah gaya seperti; jenis fon (*font*) tulisan, warna dan jarak ke dalam dokumen web.
- Manakala, Javascript ialah bahasa yang digunakan untuk mengurus tingkah laku halaman web.

CONTOH: HTML, CSS & JAVASCRIPT



MENGSKRAP WEB:

- Mengskrap web ialah teknik untuk menjelmakan data yang ada dalam format tidak berstruktur (tag HTML) dalam web kepada format berstruktur yang boleh diakses dan digunakan dengan mudah.
- Untuk mengskrap data dari tapak web, kita perlu mengetahui struktur hierarki yang wujud dalam laman web.
- Struktur hierarki ini dikenali sebagai DOM (*Document Object Model*).
- DOM mentakrifkan struktur logik bagi dokumen dan cara ia boleh diakses dan dimanipulasi.
- Selain itu, alat (*tool*) lain yang penting ialah Xpath.
- XPath merujuk kepada XML Path Language.
- Ia ialah bahasa pertanyaan (*query language*) untuk memilih nod daripada dokumen XTML atau XML.

RUJUKAN:

- 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.

TOPIK SETERUSNYA:

Perlombongan Proses