**Lesson 13 Data Processing and Web Crawling with Python**

1. **Data Processing wit Regular Expression (Regex)**

*Regular Expression (Regex ) là gì*

Regex là những biểu thức hay còn gọi là cú pháp dùng để xử lý chuỗi. Những biểu thức Regex này có cú pháp riêng và bạn cần phải học / biết những nguyên tắc đó để sử dụng. Bạn nào đã học / biết noSQL như MongoDB chắc không còn lạ gì Regex.

Xử lý chuỗi là gì ?

Xử lý chuỗi là xử lý ( manipulate) dữ liệu dạng chuỗi (string) để biến hóa (transform) nó thành đầu ra (output) mong muốn. Xử lý chuỗi là 1 phần của công việc xử lý dữ liệu và có rất nhiều công việc phân tích đòi hỏi việc xử lý chuỗi. Ví dụ : Sentiment Analysis , Spam email , hoặc lấy dữ liệu từ internet về.

String Manipulation :

-SPLIT : tách chuỗi dựa theo điều kiện

- CONCAT : nối 2 chuỗi thành 1 chuỗi mới

- SLICING : truy xuất đoạn , chuỗi hoặc ký tự cụ thể từ chuỗi gốc

- INDEXING : truy xuất chuỗi theo index

- Adjusting Cases : viết hoa / viết thường

- Find & Replace : Tìm và thay thế

- Type conversion : Ép kiểu chuỗi

- Formatted Strings ( fstring) : In chuỗi ra theo format

- Repetitions : tần số , số lần lặp

- Metacharacter : ký tự đặc biệt

Khái niệm substrings

Practice : [Link](https://colab.research.google.com/drive/1BsDkaCvur0GwOadI6_OqM7CXm8-NHkvh?usp=sharing)

1. **Web Crawling**

Web Crawling là chúng ta sẽ cào nội dung của trang web về, và nội dung đó sẽ hiện dưới dạng response text như khi chúng ta bấm F12 để xem elements của trang web.

Vì vậy chúng ta cũng cần phải biết 1 ít khái niệm về HTML và CSS để có thể thực hiện tốt công việc này.

Có nhiều cách để crawl 1 trang web , nhưng hiện tại có 3 cách chính

**BeautifulSoup4** : Dùng để crawl những trang web có nội dung đơn giản, source code không quá phức tạp, chúng ta hoàn toàn có thể trỏ đến những nội dung cần crawl bằng các HTML tag. BeautifulSoup chỉ móc cái content của web về thôi còn lại là chúng ta tự lọc ra data trong content. Vì vậy BS4 không thực sự cào data. Và Bs4 chỉ là 1 thư viện.

**Scrapy** : Là một framework hoàn chỉnh để crawl data từ web, vì vậy phức tạp hơn BeautifulSoup , cần set up và code cũng khó hơn tuy nhiên có thể làm được nhiều việc nhanh hơn và hơn BS4. Cần nhiều thời gian hơn để master.

**Selenium** : Selenium là một automation tool.Ban đầu, Selenium được dùng cho mục đích automation web browsers chứ không phải cho việc crawl data. Tuy nhiên, chúng ta có thể sử dụng việc automation để crawl data. Có nhiều trang web chúng ta không thể crawl được bằng Bs4 hoặc Scrapy ( Web Build bằng Javascript). Tuy nhiên, Selenium có thể cân tất.

Với khóa 3 này, chúng ta sẽ học BeautifulSoup là chính. 2 thư viện còn lại các bạn có thể tự tìm hiểu và tự học được sau khi nắm rõ cách crawl data. (Trong lúc làm bài cuối kỳ các bạn có áp dụng thì cứ hỏi giáo viên sẽ giải đáp )

***Khái niệm HTML / CSS***

HTML Basics

Content HTML của 1 trang web sẽ có dạng giống thế này

Text, letter

Description automatically generated

HTML Elements

[HTML Elements (w3schools.com)](https://www.w3schools.com/html/html_elements.asp)

HTML Attributes

[HTML Attributes (w3schools.com)](https://www.w3schools.com/html/html_attributes.asp)

Practice : [Link](https://colab.research.google.com/drive/1DTTXwuv6BBMB_EjxL5mS1mp5CUcek_0q?usp=sharing)