**Data Crawling**

## Giới thiệu.

* Mục đích
  + Thu thập dữ liệu về chữ nôm (các văn bản hán nôm, từ điển hán nôm) từ trang web nomfoundation.org để phục vụ nghiên cứu và phát triển ứng dụng liên quan tới hàn nôm, bao gồm fine-tune mô hình AI, …
* Phạm vi:
  + Tập trung vào từ điển chữ Hán-Nôm, các văn bản cổ được viết bằng Hán-Nôm.

## Mô tả nguồn dữ liệu.

* Tên trang web: nomfoundation.org
* Loại dữ liệu thu thập.
  + Chữ Hán-Nôm (các từ, các câu, bài thơ,…)
* Cách dữ liệu được tổ chức trên trang web.
  + Dữ liệu từ diển được chia thành nhiều trang và nhiều danh mục (cấu trúc khá phức tạp).
  + Dữ liệu được phân trang sẵn từ server và được nhúng các đoạn mã Javascript để lấy dữ liệu.

## Quy trình thu thập dữ liệu.

* Công cụ sử dụng:
  + Ngôn ngữ Python.
  + Thư viện sử dụng BeautifulSoup, Request kết hợp với Selenium để xử lí phân trang và các đoạn mã chứa javascript.
* Phương pháp thu thập:
  + Phân tích cấu trúc HTML của các trang mục tiêu
  + Lập trình một crawler để trích xuất dữ liệu cần thiết (có thể ta sẽ cần lập trình nhiều crawler để trích xuất với từng loại trang khác nhau).
  + Lưu dữ liệu bằng định dạng TXT.
* Kich thước dự kiến: số lượng các mục ước tính trên trang web

## Quy tắc thu thập và gới hạn

* Tốc độ crawl:
  + Mỗi request đến trang web sẽ > 1s đảm bảo tuân thủ tốc độ truy cập và không gây quá trải cho trang web
* Giới hạn:
  + Tuân thủ các quy định về bản quyền khi truy cập và thu thập dữ liệu.

## Mô tả quá trình thu thập.

1. Phân tích và trích xuất các đường link từ trang web chình (<http://nomfoundation.org>)

* Ta gửi request đến trang web bằng thư viện request và sử dụng beautifull soup để chuyển đổi đoạn mã HTML.
* Sau khi đã có được đoạn mã HTML, ta phân tích và nhận thấy các đường link ta cần được lưu trữ trong thẻ ‘a’ và có class là ‘has\_submenu’.
* Ta trích xuất và đoạn mã và nhận được các đường link cần thiết.

A blue and white text on a white background

Description automatically generated

* Trong phạm vi của đồ án ta chỉ cần thu thập thêm 3 đường link cuối (vì một số dữ liệu đã được cung cấp – thu thập trước đó) .

1. Thu thập trên **History of Greater VietNam**.

* Đường link chữa dữ liệu chính: (<http://nomfoundation.org/nom-project/History-of-Greater-Vietnam/Fulltext>)

A screenshot of a computer

Description automatically generated

* Sau khi truy cập vào đường link trên: Ta có thể thấy rằng trang dữ liệu được chia ra thành nhiều quyển và mỗi quyển đều chứa các thông tin cần thiết.
* Vì vậy trước hết ta cần phân tích đoạn mã HTML từ trang web này đề có thể trích xuất ra đường link chữa dữ liệu của từng quyển.
* Sau khi phân tích đoạn mã HTML trên thì ta nhận được các đường link và tên của từng quyển mà ta cần thu thập.

A close-up of a blue text

Description automatically generated

* Tiếp theo ta cần tiến hành phân tích từng đường link dữ liệu để có thể lấy được các dữ liệu cần thiết.
* Ta truy cập vào 1 đường link cụ thể để tiến hành phân tích.

A screenshot of a computer

Description automatically generated

* Sau khi truy cập vào đường link, quan sát sơ bộ ta có thể dễ dạng nhận thấy được mỗi quyển được phân thành nhiều trang. Và để có dữ liệu của từng trang ta cần truy cập vào các trang kế tiếp để lấy được hoàn toàn.
* Tuy nhiên ta lại gặp một vấn đề Tại đây sự phân trang không sử dụng các đường link như các page khác mà sử dụng các đoạn mã javascript được nhúng vào trang và khi gọi các đoạn mã thì server sẽ trả về các trang khác **javascript:GotoPage(??).**
* Vì thế ta cần sự hỗ trợ của công cụ Selenium để giả lập trình duyệt để truy cập và thực hiện:

A computer code with red text

Description automatically generated

* Sau khi hoàn thành xử lý trên 1 đường link ta tiến hành cài đặt để chạy trên các đường link còn lại và lưu chúng dưới định dạng TXT.

## Phụ mục.

* Các đường link khác sẽ được phân tích dựa trên đặc tính của chúng.
* Mã nguồn để thu thập dữ liệu sẽ được lưu trữ trong file .ipynb.
* Môi tường chạy: sử dụng google colab (để dễ dàng xử lí và cài đặt thêm các mỗi trường mới, tránh xung đột với các trường trình khác)
* Tối ưu: Mã nguồn được viết theo hướng xử lý tuần tự. Vì vậy bạn có thể cải tiến bằng cách chạy song song hoặc Tách dữ liệu thành từng nhóm và tiến hành xử lí độc lập.