Assignment 3: Web Scraping

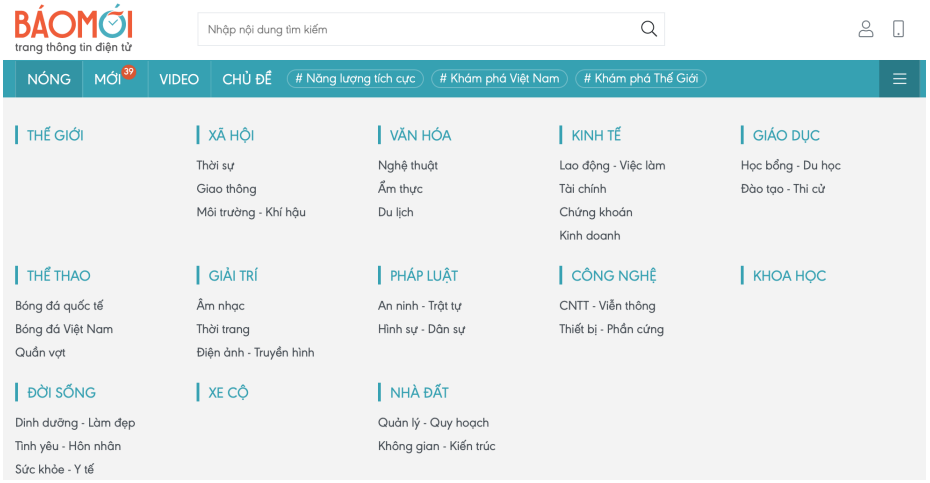
phuonglh@gmail.com

March 13, 2023

1 Problem

In this assignment, you will develop a program capable of extracting text data from a website. This is call *web scraping*, a necessary step to harvest raw data for search engines or other information retrieval systems.

Your program takes as input a start web page in a given section of the new portal `baomoi.com`. There are 13 main sections in the home page of the new portal, as shown in blue allcap titles in the following figure:



The WORLD section has the URI:

`https://baomoi.com/the-gioi.epi`

The last REAL ESTATE section has the URI:

`https://baomoi.com/nha-dat.epi`

Click on any of the 13 sections, you will see a list of news articles. There are about several dozens or a hundred news in each section, which are updated daily. It is your task to pick a section and extract main content of every news in that section and save the result into a JSON file.

2 Guide

Each news has a unique URL, for example

https://vov.vn/the-gioi/chu-tich-trung-quoc-tap-can-binh-co-the-tham-nga-trong-tuan-toi-post1007215.vov

You can use an available library to extract the main text content from it. The article above has the main content:

"*VOV.VN - Hãng tin Reuters hôm nay (13/3) đưa tin, Chủ tịch Trung Quốc Tập Cận Bình dự kiến sẽ thăm Nga trong tuần tới, sớm hơn nhiều so với thông tin được báo chí nêu trước đó. Hôm 30/1, hãng thông tấn Tass của Nga đưa tin, Tổng thống Vladimir Putin đã mời Chủ tịch Trung Quốc Tập Cận Bình sang thăm Nga trong mùa xuân này. Các nguồn tin cho biết, chuyến thăm sẽ diễn ra trong tháng 4 hoặc tháng 5. Trong khi đó, tháng trước, ông Vương Nghị - chủ nhiệm Văn phòng Ủy ban Công tác đối ngoại Trung ương Đảng Cộng sản Trung Quốc đã có chuyến thăm Nga và hội kiến Tổng thống Putin. Chuyến thăm Nga của ông Vương Nghị cũng được coi là bước chuẩn bị quan trọng cho chuyến thăm Nga của ông Tập Cận Bình. Hiện Bộ Ngoại giao Trung Quốc chưa phản hồi đề nghị xác nhận thông tin về chuyến thăm. Chủ tịch Trung Quốc Tập Cận Bình đã gặp trực tiếp Tổng thống Putin 39 lần kể từ khi trở thành Chủ tịch nước Trung Quốc, lần gặp nhau nhau gần đây nhất giữa hai nhà lãnh đạo Nga – Trung diễn ra vào tháng 9 năm ngoái, tại hội nghị thượng đỉnh ở Trung Á. Trước đó, hôm 10/3, tại phiên họp toàn thể lần thứ ba Kỳ họp thứ nhất Đại hội Đại biểu Nhân dân Toàn Quốc (tức Quốc hội Trung Quốc) khóa XIV, Tổng Bí thư Ban Chấp hành Trung ương Đảng Cộng sản Trung Quốc Tập Cận Bình đã được bầu làm Chủ tịch Trung Quốc và Chủ tịch Quân ủy Trung ương nhiệm kỳ thứ ba liên tiếp./. Hạnh Phúc/VOV1 (biên dịch) Nguồn: Reuters*"

Note that all new line characters in the main content can be kept or removed, per design.

Some useful libraries for web scraping/text extraction from webpages:

- For Python: Trafilatura – `https://trafilatura.readthedocs.io/en/latest/`
- For Scala/Java: Boilerpipe – `https://github.com/kohlschutter/boilerpipe`, or JSoup – `https://jsoup.org`

You are free to search for and use other libs for extracting main content if they are good.

First, you randomly choose a number between 1 and 13, and stick with that number, which corresponds to a section. If there are 100 news articles on that section then you should extract all of them and save to a JSON file with your name. Your section should be random because your teacher does

not want to see that many students work on the same section and extract the same articles! The JSON file has the following format, saved in UTF8 encoding:

```
[
{
  "url" : "https://...",
  "content" : "...",
  "date" :  "..."
},
{
  "url" : "https://...",
  "content" : "...",
  "date" :  "..."
},
...
]
```

3 Submission

If you develop your program in Scala, Python, or Java, submit your source file *YourFullName.scala*, *YourFullName.py*, or *YourFullName.java*, respectively; for example *NguyenPhuTrong.scala*. Your data file has the name *NguyenPhuTrong.json*. If you have multiple source files then put them in a directory and zip it to *YourFullName.zip* and submit.

- Do not submit a Notebook (say **.ipynb*). Your teacher will not run any notebook server for evaluation of this assignment. Do not submit **.rar*, your teacher does not want to be made install an external program to decompress your file.
- You need to submit the code (*.scala/*.py) and the extracted data (*.json).
- Submit your file to Classroom before the due date. You have 7 days to turn in your program.