Bài tập thực hành:

Scraper với jsoup

Thời gian làm bài: 1 tuần (Xem dealine trong link nộp bài trên moodle)

Nộp bài:

- Nộp bài lên moodle.
- Đặt tên bài nộp theo định dạng MSSV.rar. Trong đó bao gồm:
  - Tập tin báo cáo.
  - Các tập tin dữ liệu theo yêu cầu của bài tập
- Nếu sử dụng code trên Internet, trước đoạn code sử dụng phải có chú thích đường dẫn đến trang web chứa đoạn code. Trường hợp 2 bài có đoạn code giống nhau, bài không chú thích đường dẫn đến trang web chứa source code sẽ bị tính là gian lận, và bị 0 điểm.

Các hành vi sử dụng toàn bộ/một phần bài làm của người khác sẽ bị 0 điểm cho toàn bộ phần thực hành

1 Jsoup

Download tại: https://jsoup.org/download

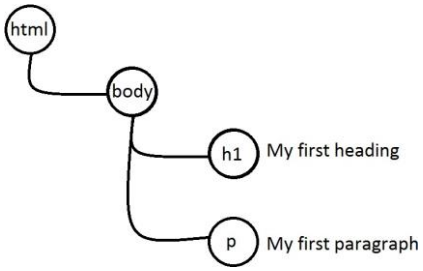
2 Cấu trúc của HTML

```
<!DOCTYPE html>
<html>
<body>

<h1>My First Heading</h1>
<p>My first paragraph.</p>

</body>
</html>
```

Nội dung tập tin html thường gồm nhiều node, mỗi nốt gồm một cặp tab mở <> - tab đóng </>. Khoảng giữa tab mở và tab đóng là nội dung của nốt. Nội dụng có thể là văn bản hoặc nốt khác. Với đoạn html trên, ta có cây như sau:



3 Thuộc tính của một node

```
<img src="w3schools.jpg" alt="W3Schools.com" width="104" height="142">
```

Một node sẽ có loại và các thuộc tính được định nghĩa trong tab mở <>. Ở đây, node này có loại là "img" và có các thuộc tính: "src", "alt", "width" và "height". Các giá trị của thuộc tính được ghi dưới dạng chuỗi ngay sau dấu "=" sau tên thuộc tính.

4 Chạy jsoup

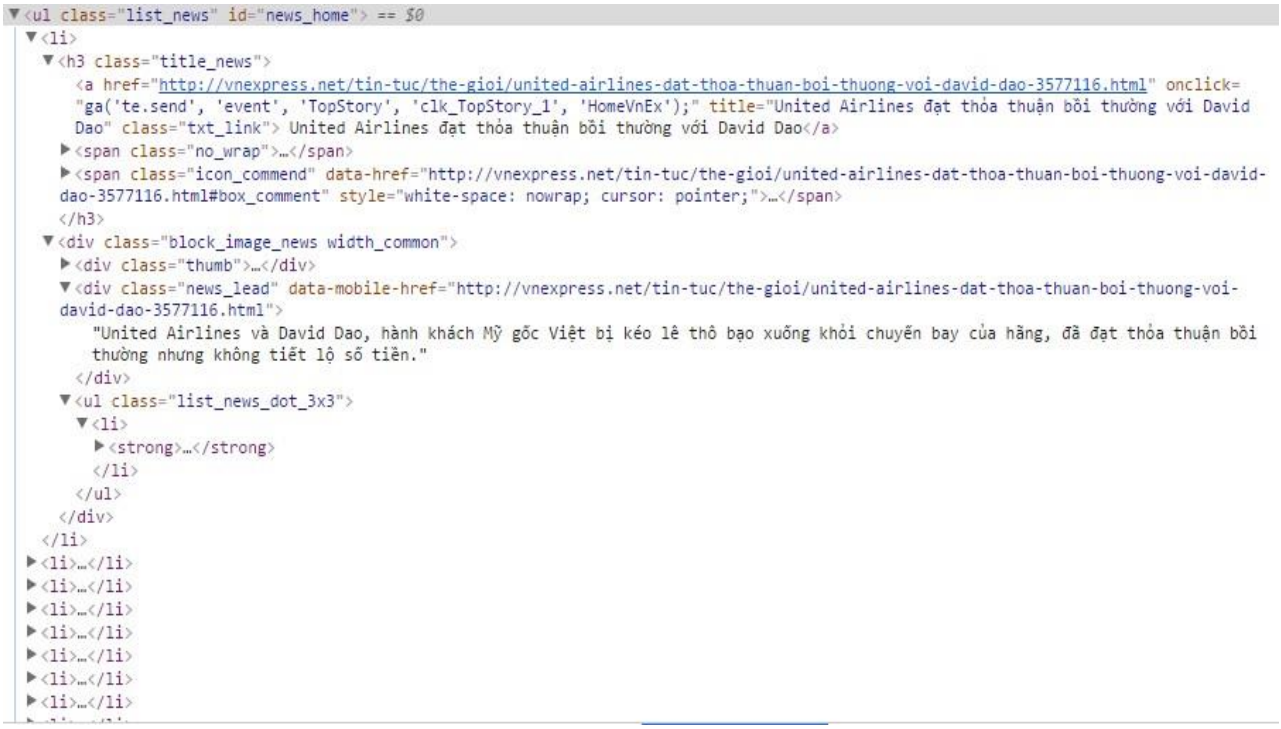
Giả sử ta đã đọc trang web và ghi nội dung vào String content của một trang thuộc vnexpress.net.

```
Document doc = Jsoup.parse(content);
Elements page_news = doc.getElementsByClass("list_news");
```

Đoạn code này giúp ta lấy được node có thuộc tính **class** là "list_news" trong toàn bộ html. Ta sẽ duyệt từng node và xử lý bằng:

```
for (Element news : page_news){
            ...
}
```

Ta xét 1 node có class là "list_news" như sau:



Ở đây, node này là một danh sách các tin tức. Mỗi tin tức có title nằm trong node **h3** với **class** là "title_news". Nội dung của tin tức là nội dung text của node **div** với **class** là "news_leads".

Để lấy các node con theo loại, ta sử dụng hàm getElementsByClass:

```
Elements children = news.getElementsByTag("loại thẻ");
```

Để lấy các node con theo class, ta sử dụng hàm `getElementsByClass`:

```
Elements children = news.getElementsByClass("tên class");
```

Để lấy giá trị các thuộc tính của node, ta sử dụng hàm attr():

```
String attr_value = news.attr("tên thuộc tính");
```

Để lấy danh sách các thuộc tính của node, ta sử dụng hàm `attributes()`:

```
Attributes attr = news.attributes();
for(Attribute a : attr){
    System.out.print(a.getKey());
    System.out.println(a.getValue());
}
```

Ví dụ

Với node ul có class là **list_news** như trên, nếu ta thực hiện:

```
Elements list_items = news.children();
```

Với đoạn code này, ta sẽ lấy được các node con trực tiếp của node trên là các node li.

Tuy nhiên, nếu ta thực hiện:

```
Elements list_items = news.getElementsByTag("li");
```

Với đoạn code này, ngoài những node con trực tiếp, ta còn lấy được các node **li** nằm trong các node con này.

5 Bài tập

Với một trang vnexpress.net cho trước, hãy lọc ra các tin tức và ghi vào một tập tin news.txt. Nội dung tập tin news.txt này có dạng:

```
Title: <Title của tin tức 1>
Text: <Nội dung tin tức 1>
Title: <Title của tin tức 2>
Text: <Nội dung tin tức 2>
...
```

Nộp lại báo cáo, mã nguồn và thư mục chứa kết quả thực thi.

Tài liệu tham khảo:

[1] https://www.w3schools.com