Bài tập Tuần 3

Bài 1: Thu thập dữ liệu từ 1 URL và các URL liên quan, trích rút tất cả địa chỉ email và số điện thoại có trong trang web.

1. Tìm 1 trang web có chứa số điện thoại và email (thường có trên các diễn đàn)
2. Lấy nội dung từ 1 URL
3. Tìm các URL có trong trang web (Chỉ cần lấy 10 URL cùng mục)
    a. Ví dụ các URL cùng 1 chuyên mục
    b. Đối với bài viết trên Diễn đàn, lấy các trang trong topic hiện tại
4. Truy cập vào tất cả các URL đã thu thập được
    a. Tìm và lấy địa chỉ email, số điện thoại
    b. Lưu trữ vào file

Bài 2: Trích rút văn bản từ website theo một chủ đề đã chọn trước

1. Chọn 1 địa điểm: Đà Nẵng, Hà Nội, Quảng Bình…
2. Chọn 1 trang web tin tức (vnexpress, tuoitre, thanhnien, vietnamnet, dantri…)
3. Tìm các url và thu thập 100 bài viết bất kỳ.
4. Lấy nội dung của bài viết, tách thành các câu và lưu vào file
5. Với dữ liệu đã có, hãy xác định các từ ghép và nối với nhau bởi ký tự "_", xác định các danh từ riêng (tên địa điểm, tên người…)
    a. Sử dụng thư viện underthesea
    b. https://pypi.org/project/underthesea/
6. Nếu nội dung trang web chứa các từ khóa liên quan đến địa điểm đã chọn, hãy lưu tất cả các tiêu đề và link bài viết này vào 1 file, đặt tên "danang.txt" hoặc "hanoi.txt" hoặc "quangbinh.txt"
7. Hãy thử tìm cách phân loại bài viết đó thuộc danh mục nào (Thể thao, Pháp luật, Kinh tế, Giáo dục…) và lưu thành các danh sách bài viết tương ứng.

Lưu ý: Chức năng phân loại văn bản này underthesea có hỗ trợ, tuy nhiên chỉ chạy được trên Ubuntu, và cần tải "model" trước khi sử dụng.

```
$ underthesea download-model TC_GENERAL
$ underthesea list-data --all
```