THU THẬP DỮ LIỆU TỪ WEBSITE (WEB CRAWLING)

1 Tổng quan và Mục tiêu Học tập

1.1 Tổng quan

Notebook này hướng dẫn bạn cách **thu thập dữ liệu tự động từ các website** (web crawling) sử dụng thư viện **Selenium** và **BeautifulSoup**. Chúng ta sẽ thực hành crawl tin tức từ trang web CafeF để thu thập danh sách các bài viết về tài chính quốc tế, sau đó lưu trữ dữ liệu vào file Excel để phục vụ cho các bước phân tích tiếp theo.

1.2 Mục tiêu học tập

- **Hiểu khái niệm web crawling** và ứng dụng trong khoa học dữ liệu
- **Thiết lập môi trường crawling** với Selenium và Chrome WebDriver
- **Sử dụng thành thạo Selenium** để điều khiển trình duyệt tự động
- **Áp dụng BeautifulSoup** để phân tích và trích xuất dữ liệu từ HTML
- **Xử lý và lưu trữ dữ liệu** crawl được bằng pandas và Excel
- **Nắm vững quy trình hoàn chỉnh** từ crawl đến export dữ liệu
- **Hiểu các vấn đề pháp lý và đạo đức** khi crawl dữ liệu

2 Hướng dẫn Cài đặt Môi trường

2.1 Thư viện cần thiết

Để thực hiện web crawling trong notebook này, chúng ta cần cài đặt các thư viện Python sau:

- **`selenium`**: Thư viện điều khiển trình duyệt web tự động
- **`beautifulsoup4`**: Thư viện phân tích và trích xuất dữ liệu từ HTML/XML
- **`pandas`**: Thư viện xử lý và phân tích dữ liệu
- **`openpyxl`**: Thư viện đọc/ghi file Excel
- **`chromedriver-autoinstaller`**: Tự động cài đặt ChromeDriver tương thích

2.1.1 Lệnh cài đặt

Chạy lệnh sau để cài đặt tất cả các thư viện cần thiết:

```
pip install selenium beautifulsoup4 pandas openpyxl
chromedriver-autoinstaller
```

2.1.2 Yêu cầu hệ thống

- **Python 3.7+**: Đảm bảo bạn đang sử dụng phiên bản Python 3.7 trở lên
- **Google Chrome**: Cần có trình duyệt Chrome được cài đặt trên hệ thống
- **Kết nối Internet**: Để tải ChromeDriver và truy cập website cần crawl

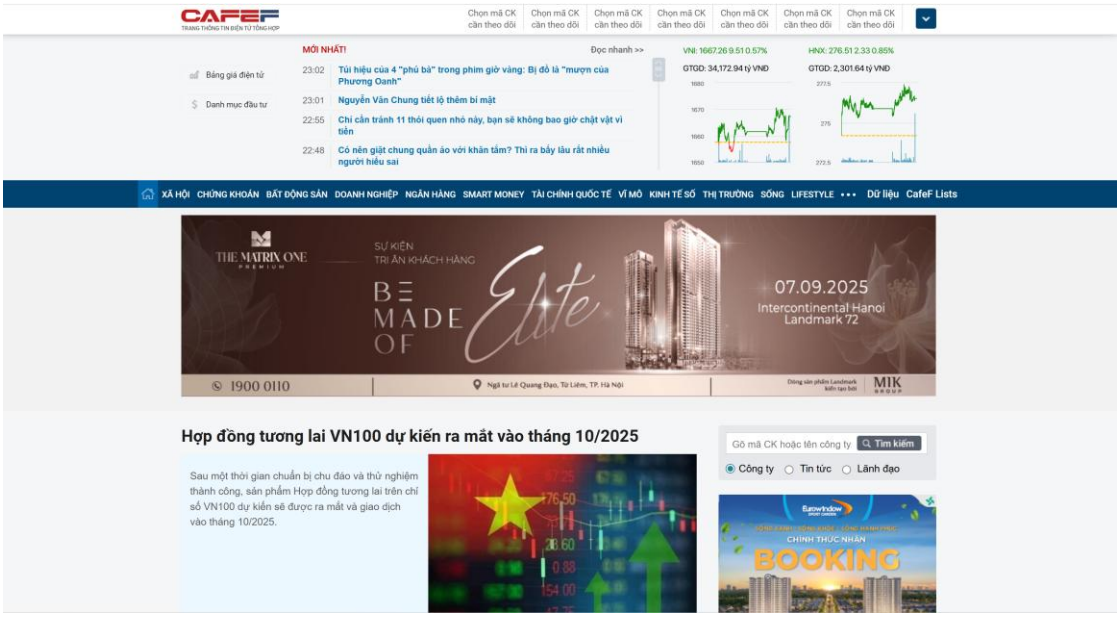
3 Quy trình crawl dữ liệu

Mục đích: Đây là bước cốt lõi - sử dụng Selenium và BeautifulSoup để truy cập website và trích xuất dữ liệu tin tức.

Quy trình crawl dữ liệu:

1. **Truy cập URL**: Sử dụng WebDriver để mở trang web CafeF
2. **Chờ tải trang**: Đảm bảo trang web được tải hoàn toàn bằng `time.sleep()`
3. **Phân tích HTML**: Sử dụng BeautifulSoup để parse mã nguồn HTML
4. **Tìm kiếm elements**: Sử dụng CSS selectors để định vị các thẻ HTML chứa tin tức
5. **Trích xuất dữ liệu**: Lấy tiêu đề và đường link của từng bài viết
6. **Lưu trữ dữ liệu**: Tổ chức dữ liệu vào list Python để xử lý tiếp

3.1 Gửi request đến website



Trang web được chọn để crawl dữ liệu

- Sử dụng `requests.get()` để tải nội dung HTML từ URL
- Kiểm tra status code để đảm bảo request thành công.
- Tìm thẻ <h3> bên trong mỗi bài viết, sau đó tìm tiếp thẻ <a> chứa tiêu đề và đường dẫn
- **Khái niệm CSS Selector**: `"div.tlitem.box-category-item"` có nghĩa là tìm thẻ `<div>` có các class `tlitem` và `box-category-item`.

Kết quả

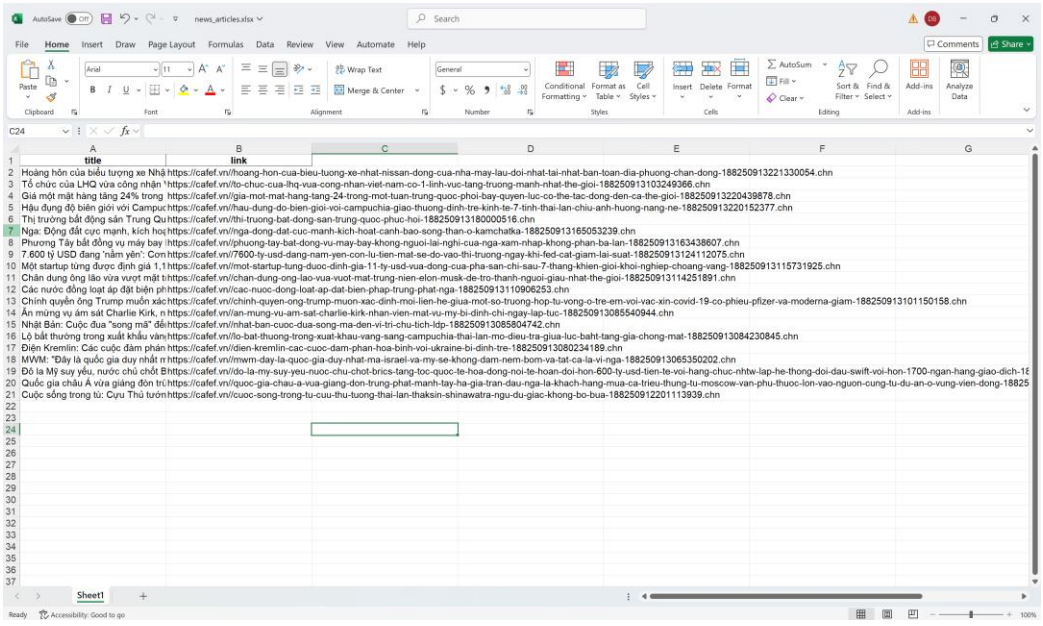
```
Đang truy cập: https://cafef.vn//tai-chinh-quoc-te.chn
Vui lòng đợi trong giây lát...
Đã tải trang thành công!
Đang phân tích cấu trúc HTML...
Đã tìm thấy 20 bài viết trên trang!
Đang trích xuất thông tin từng bài viết...

Hoàn thành crawl! Đã thu thập 20 bài viết thành công!
{'title': 'Hoàng hôn của biểu tượng xe Nhật: Nissan đóng cửa nhà máy lâu đời nhất tại Nhật Bản, toàn địa phương chấn động', 'link': 'htt
{'title': 'Tổ chức của LHQ vừa công nhận Việt Nam có 1 lĩnh vực "tăng trưởng mạnh nhất thế giới"', 'link': 'https://cafef.vn//to-chuc-cu
{'title': 'Giá một mặt hàng tăng 24% trong một tuần: Trung Quốc phơi bày quyền lực, có thể tác động đến cả thế giới', 'link': 'https://c
{'title': 'Hậu đụng độ biên giới với Campuchia: Giao thương đình trệ, kinh tế 7 tỉnh Thái Lan chịu ảnh hưởng nặng nề', 'link': 'https://
{'title': 'Thị trường bất động sản Trung Quốc phục hồi', 'link': 'https://cafef.vn//thi-truong-bat-dong-san-trung-quoc-phuc-hoi-18825091
{'title': 'Nga: Động đất cực mạnh, kích hoạt cảnh báo sóng thần ở Kamchatka', 'link': 'https://cafef.vn//nga-dong-dat-cuc-manh-kich-hoat
{'title': 'Phương Tây bất đồng vụ máy bay không người lái nghi của Nga xâm nhập không phận Ba Lan', 'link': 'https://cafef.vn//phuong-ta
{'title': "7.600 tỷ USD đang 'nằm yên': Cơn lũ tiền mặt sẽ đổ vào thị trường ngay khi Fed cắt giảm lãi suất?", 'link': 'https://cafef.vn
{'title': 'Một startup từng được định giá 1,1 tỷ USD vừa đóng cửa, phá sản chỉ sau 7 tháng khiến giới khởi nghiệp choáng váng', 'link':
{'title': 'Chân dung ông lão vừa vượt mặt trung niên Elon Musk để trở thành người giàu nhất thế giới', 'link': 'https://cafef.vn//chan-d
{'title': 'Các nước đồng loạt áp đặt biện pháp trừng phạt Nga', 'link': 'https://cafef.vn//cac-nuoc-dong-loat-ap-dat-bien-phap-trung-pha
{'title': 'Chính quyền ông Trump muốn xác định mối liên hệ giữa một số trường hợp tử vong ở trẻ em với vắc xin Covid-19: Có phiếu Pfizer
{'title': 'Ăn mừng vụ ám sát Charlie Kirk, nhân viên mật vụ Mỹ bị đình chỉ ngay lập tức', 'link': 'https://cafef.vn//an-mung-vu-am-sat-c
{'title': 'Nhật Bản: Cuộc đua "song mã" đến vị trí chủ tịch LDP?', 'link': 'https://cafef.vn//nhat-ban-cuoc-dua-song-ma-den-vi-tri-chu-t
{'title': 'Lộ bất thường trong xuất khẩu vàng sang Campuchia, Thái Lan mở điều tra giữa lúc baht tăng giá chóng mặt', 'link': 'https://c
{'title': 'Điện Kremlin: Các cuộc đàm phán hòa bình với Ukraine bị đình trệ', 'link': 'https://cafef.vn//dien-kremlin-cac-cuoc-dam-phan-
{'title': 'MWM: "Đây là quốc gia duy nhất mà Israel và Mỹ sẽ không dám ném bom, và tất cả là vì Nga"', 'link': 'https://cafef.vn//mwm-da
{'title': 'Đô la Mỹ suy yếu, nước chủ chốt BRICS tăng tốc 'quốc tế hóa' đồng nội tệ: Hoán đổi hơn 600 tỷ USD tiền tệ với hàng chục NHTW,
{'title': 'Quốc gia châu Á vừa giáng đòn trừng phạt, mạnh tay ép giá dầu Nga: Là khách hàng mua cả triệu thùng từ Moscow, vẫn phụ thuộc
{'title': 'Cuộc sống trong tù: Cựu Thủ tướng Thái Lan Thaksin Shinawatra ngủ đủ giấc, không bỏ bữa', 'link': 'https://cafef.vn//cuoc-sor
Đã đóng trình duyệt và giải phóng tài nguyên.
```

3.2: Xuất Dữ liệu ra File Excel

Mục đích: Chuyển đổi dữ liệu từ list Python sang **pandas DataFrame** và xuất ra file Excel để lưu trữ và phân tích sau này.

Ưu điểm của pandas DataFrame: - Cấu trúc dữ liệu dạng bảng mạnh mẽ, dễ thao tác - Hỗ trợ nhiều định dạng export (Excel, CSV, JSON…) - Tích hợp tốt với các thư viện phân tích dữ liệu khác

Tại sao chọn Excel: Format phổ biến, dễ mở bằng nhiều phần mềm, phù hợp để chia sẻ dữ liệu với người không chuyên về lập trình.



4 Phân tích Kết quả

4.1 Kiểm tra và Xem Dữ liệu đã Thu thập

- **Mục đích**: Sau khi crawl và lưu dữ liệu, chúng ta cần **kiểm tra chất lượng** và **hiểu rõ cấu trúc** của dữ liệu đã thu thập được.
- **Tại sao cần phân tích kết quả?**
  - **Đảm bảo chất lượng**: Kiểm tra xem dữ liệu có đầy đủ, chính xác không
  - **Hiểu cấu trúc**: Nắm rõ các trường dữ liệu và định dạng
  - **Phát hiện vấn đề**: Tìm ra các bài viết bị lỗi, link không hợp lệ
  - **Lập kế hoạch xử lý**: Chuẩn bị cho các bước phân tích tiếp theo

4.2 Phân tích và Diễn giải Kết quả

- **Kết quả thu được**: Chúng ta đã thành công crawl danh sách các bài viết tin tức từ trang CafeF.
- **Ý nghĩa của dữ liệu**: - **Cột `title`**: Chứa tiêu đề của từng bài viết, giúp hiểu nội dung chính **Cột `link`**: URL đầy đủ dẫn đến bài viết chi tiết, có thể dùng để crawl nội dung đầy đủ sau này
- **Đánh giá chất lượng dữ liệu**: - **Completeness**: Tất cả bài viết đều có đầy đủ tiêu đề và link
  - **Accuracy**: Các URL được tạo đúng format (base_url + relative_path)
  - **Consistency**: Cấu trúc dữ liệu đồng nhất cho tất cả bài viết
- **Ứng dụng thực tế**:
  - **Phân tích xu hướng**: Theo dõi các chủ đề hot trong tài chính quốc tế
  - **Crawl nâng cao**: Sử dụng danh sách link này để thu thập nội dung đầy đủ
  - **Phân tích cảm xúc**: Xử lý ngôn ngữ tự nhiên trên tiêu đề tin tức
  - **Tự động hóa**: Thiết lập crawl định kỳ để cập nhật tin tức mới nhất

5 Kết luận và Gợi ý Mở rộng

5.1 Tóm tắt

Trong notebook này, chúng ta đã **thành công xây dựng một quy trình web crawling hoàn chỉnh** từ A đến Z:

1. **Thiết lập môi trường**: Cài đặt Chrome, ChromeDriver và các thư viện Python
2. **Cấu hình Selenium**: Tạo WebDriver với các options tối ưu cho crawling
3. **Thực hiện crawl**: Truy cập website, phân tích HTML và trích xuất dữ liệu tin tức
4. **Lưu trữ dữ liệu**: Chuyển đổi sang pandas DataFrame và export ra Excel
5. **Phân tích kết quả**: Kiểm tra chất lượng và hiểu rõ cấu trúc dữ liệu

Kết quả đạt được: Thu thập thành công danh sách các bài viết tin tức tài chính quốc tế từ CafeF với đầy đủ tiêu đề và đường link.

5.2 Gợi ý Mở rộng

Cải tiến kỹ thuật:

1. **Crawl nội dung đầy đủ**: Sử dụng danh sách link để thu thập toàn bộ nội dung bài viết
2. **Xử lý nâng cao**: Thêm try-catch để xử lý lỗi, retry mechanism khi request thất bại
3. **Tối ưu hiệu suất**: Sử dụng threading/multiprocessing để crawl song song nhiều trang

Phân tích dữ liệu:

1. **Phân tích từ khóa**: Sử dụng TF-IDF để tìm từ khóa quan trọng trong tiêu đề
2. **Phân tích cảm xúc**: Áp dụng sentiment analysis để đánh giá tone của tin tức
3. **Phân loại chủ đề**: Sử dụng clustering hoặc classification để nhóm tin tức theo chủ đề

Ứng dụng thực tế:
1. **Crawler định kỳ**: Thiết lập cron job để crawl tin tức mới hàng ngày
2. **Dashboard tin tức**: Tạo web app hiển thị tin tức mới nhất với Flask/Streamlit
3. **Alert system**: Thiết lập cảnh báo khi có tin tức quan trọng xuất hiện

Lưu ý đạo đức và pháp lý:
- Luôn kiểm tra robots.txt của website trước khi crawl
- Respect rate limiting để không làm quá tải server
- Chỉ crawl dữ liệu công khai và tuân thủ Terms of Service