**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**BÁO CÁO QUÁ TRÌNH LÀM VIỆC NHÓM**

**MÔN: NHẬP MÔN KHOA HỌC DỮ LIỆU**

**| Giáo viên hướng dẫn |**

**ThS. Trần Trung Kiên**

**| Sinh viên thực hiện |**

**Trần Thanh Tùng - 18120258**

**Trần Hữu Chí Bảo – 18120288**

**Thành phố Hồ Chí Minh - 2020**

# MỤC LỤC

[MỤC LỤC 2](#_Toc61279532)

[THU THẬP DỮ LIỆU 3](#_Toc61279533)

[1. Từ ngày 12-12 tới ngày 26-12: 3](#_Toc61279534)

[2. Từ ngày 27-12 tới ngày 2-1: 6](#_Toc61279535)

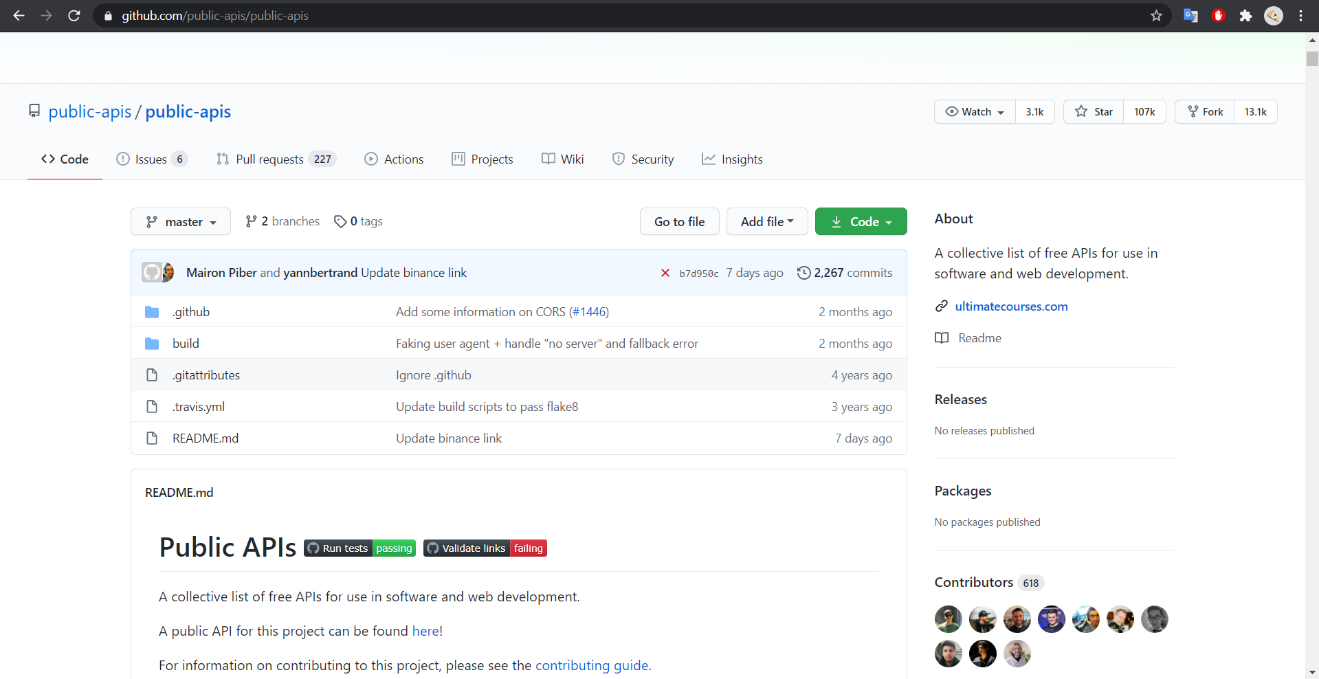
[3. Từ ngày 3-1: 6](#_Toc61279536)

# THU THẬP DỮ LIỆU

1. **Các mốc thời gian:**

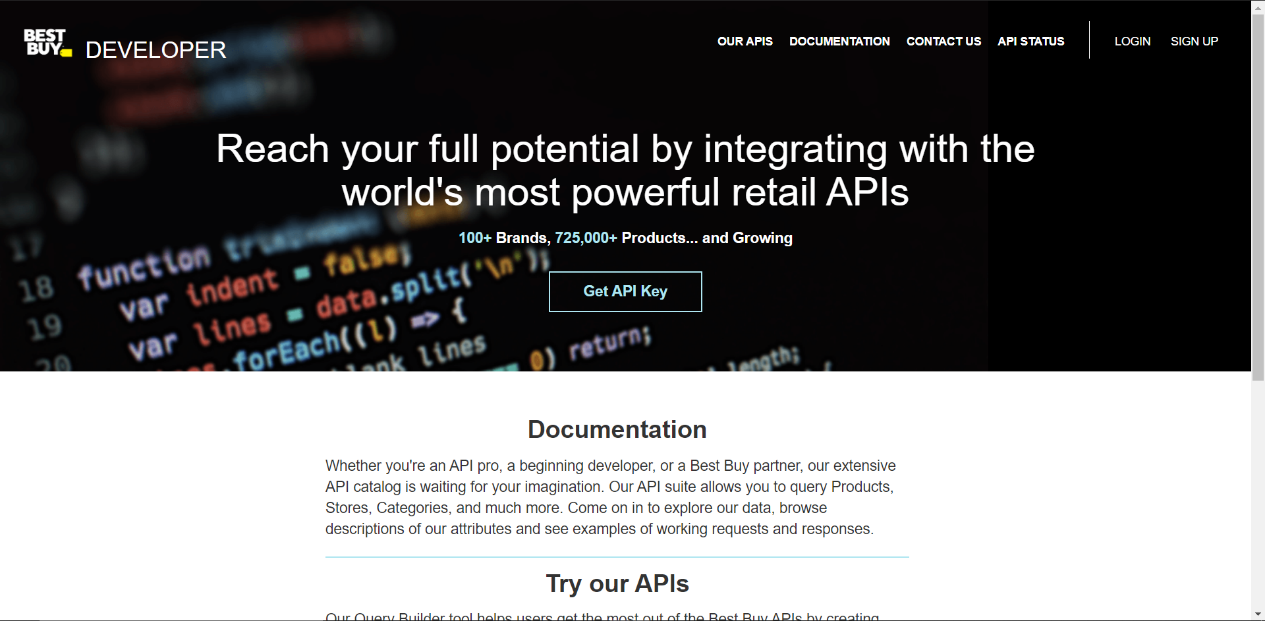
## Từ ngày 12-12 tới ngày 26-12:

Tìm hiểu các nguồn dữ liệu từ internet, đầu tiên vào link tổng hợp các public API trên github : https://github.com/public-apis/public-apis.

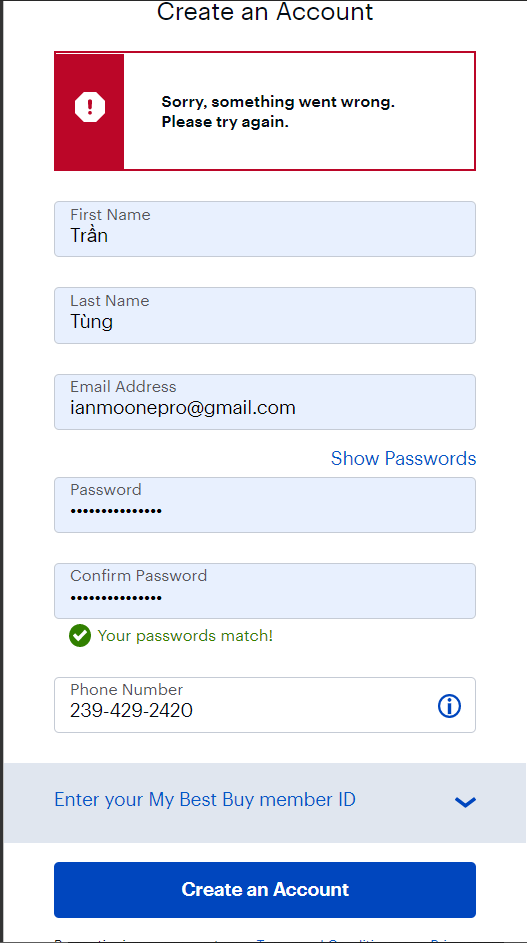


Đầu tiên định làm đề tài shopping, thu thập sản phẩm trên các trang thương mại điện tử bestbuy.com hoặc eBay.com.

Tuy nhiên bestbuy.com yêu cầu API key mà API key chỉ có thể có qua việc đăng ký tài khoản

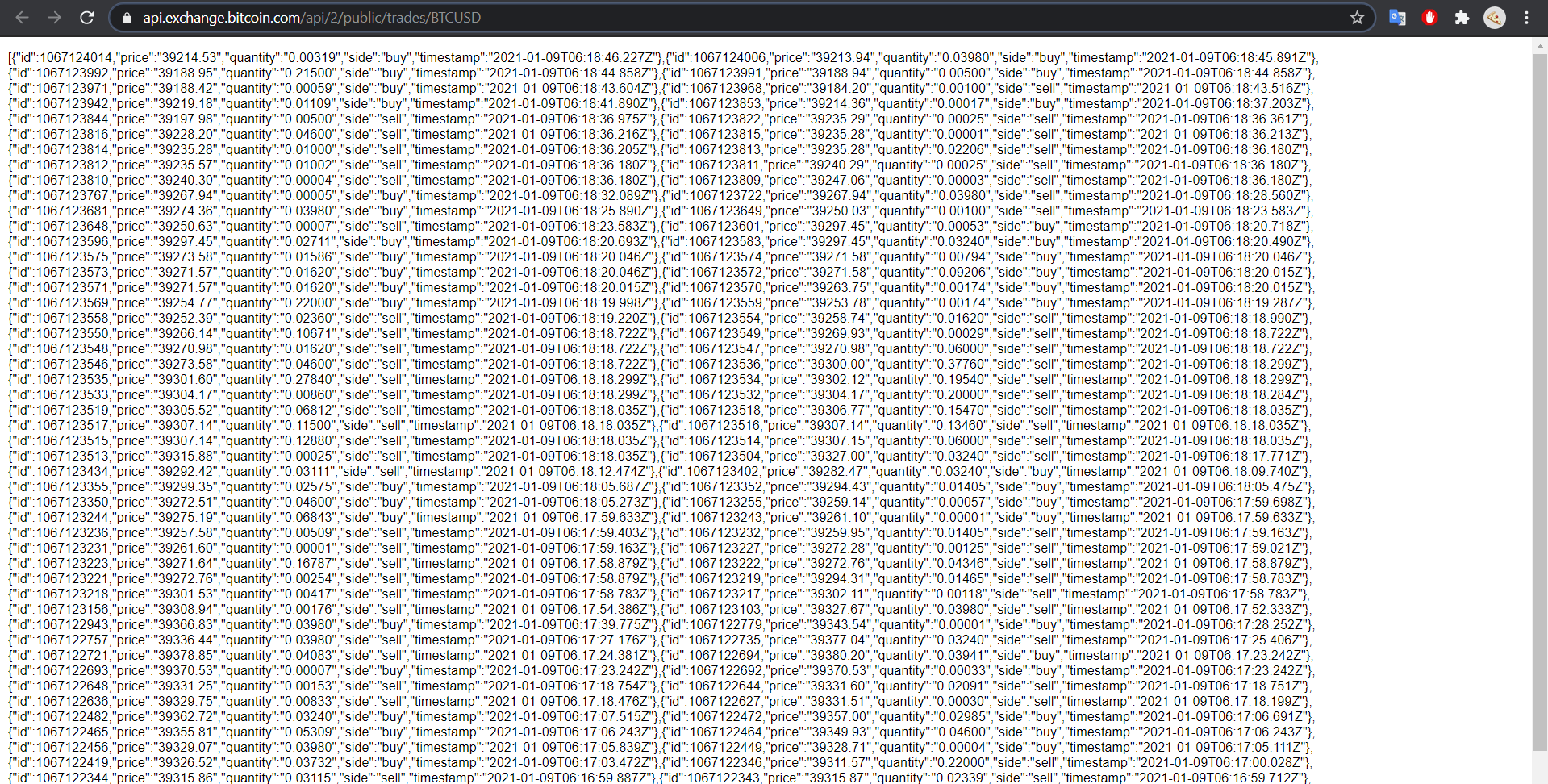


Nhưng bestbuy lại không cho đăng ký tài khoản:



Quyết định chuyển sang ebay.com. Ebay không yêu cầu API key, tuy nhiên Ebay chỉ là 1 cái “chợ” ở đó người dùng đăng sản phẩm họ cần bán và tự đặt mức giá à dữ liệu không đáng tin cậy.

Chuyển hướng đề tài sang thị trường tiền ảo, dùng api <https://api.exchange.bitcoin.com/> để thu thập thông tin về thị trường bitcoin.

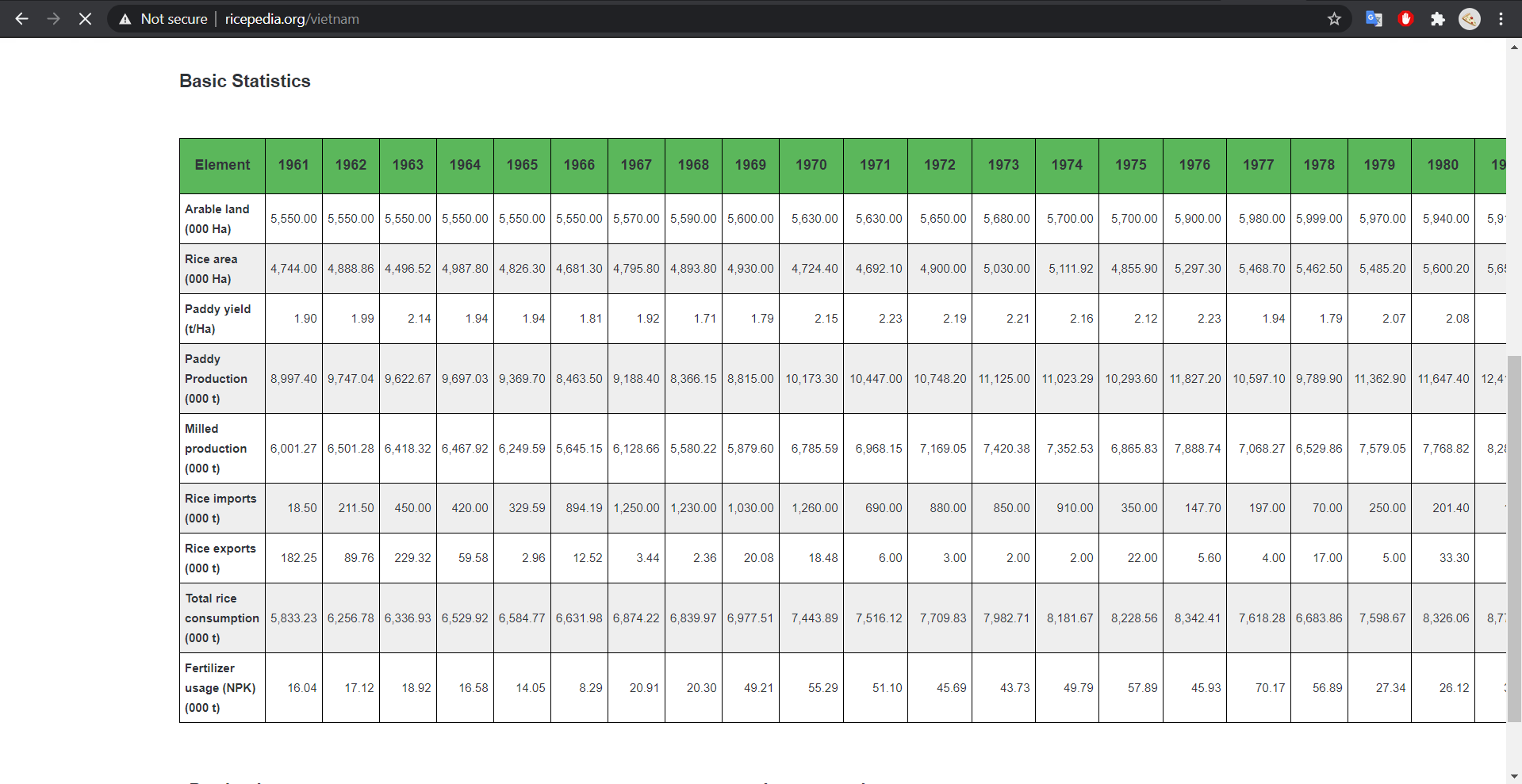


Tuy nhiên không đưa ra được câu hỏi từ dữ liệu này.

Tiếp tục chuyển hướng sang nhiều lĩnh vực khác như thời tiết, bóng đá, game nhưng cũng không đưa được ra câu hỏi vừa ý.

Chuyển sang làm về nông nghiệp, tuy nhiên rất nhiều dữ liệu nông nghiệp phải mua chứ không miễn phí.

Vô tình tìm được website ricepedia.org, có tổng hợp khá chi tiết và đầy đủ dữ liệu về tình hình sản xuất gạo của các quốc gia thành các bảng:



Quyết định làm về đề tài sản xuất gạo ở Việt Nam và parse trang <http://ricepedia.org/vietnam> để lấy dữ liệu về.

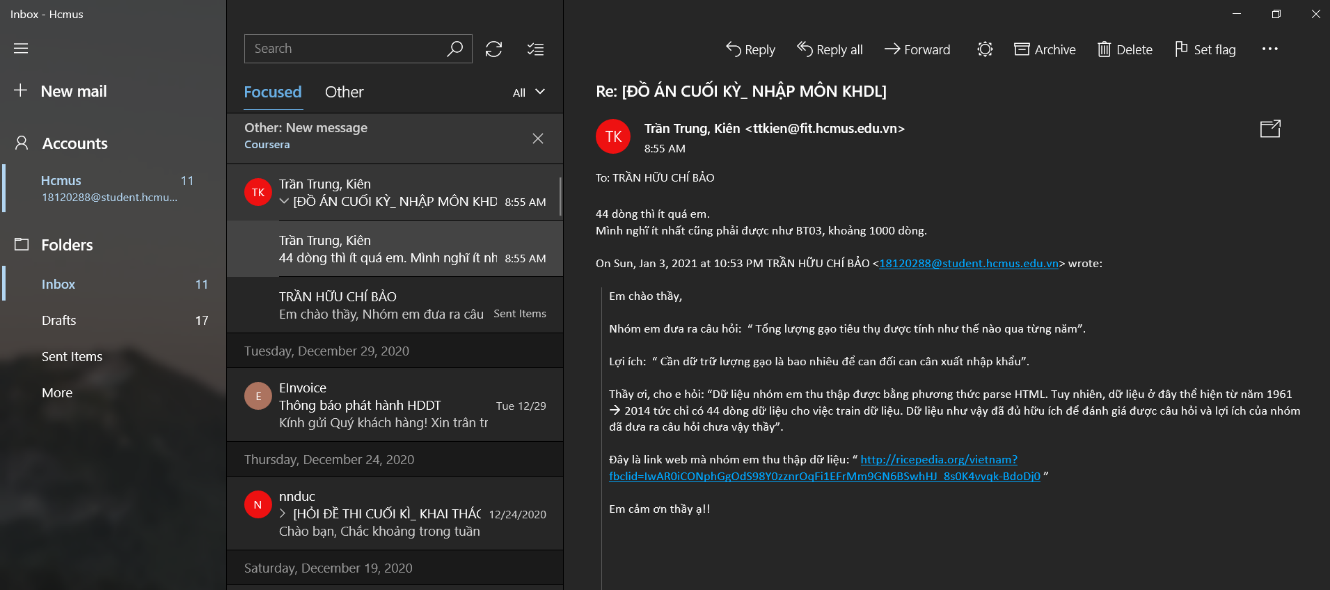
## Từ ngày 27-12 tới ngày 2-1:

Áp dụng quy trình khoa học dữ liệu đã được học, tiền xử lý dữ liệu raw ban đầu parse được.

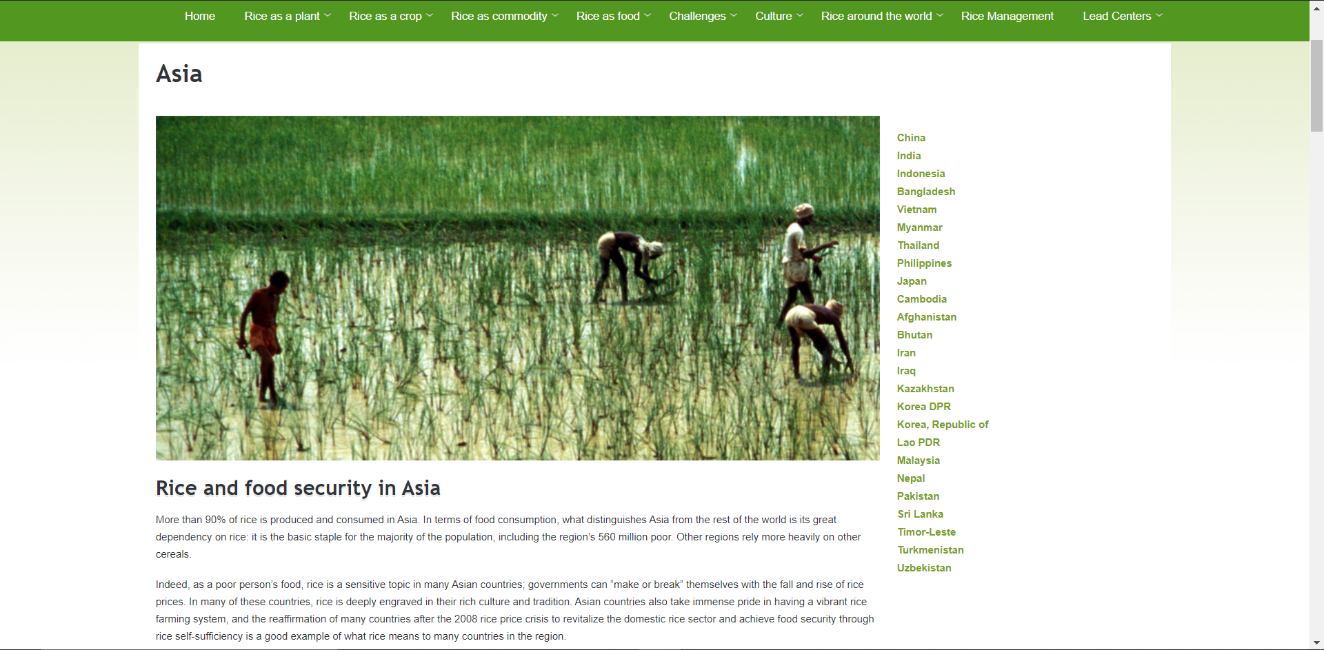
Đồng thời thử nghiệm một số mô hình để dự đoán.

## Từ ngày 3-1:

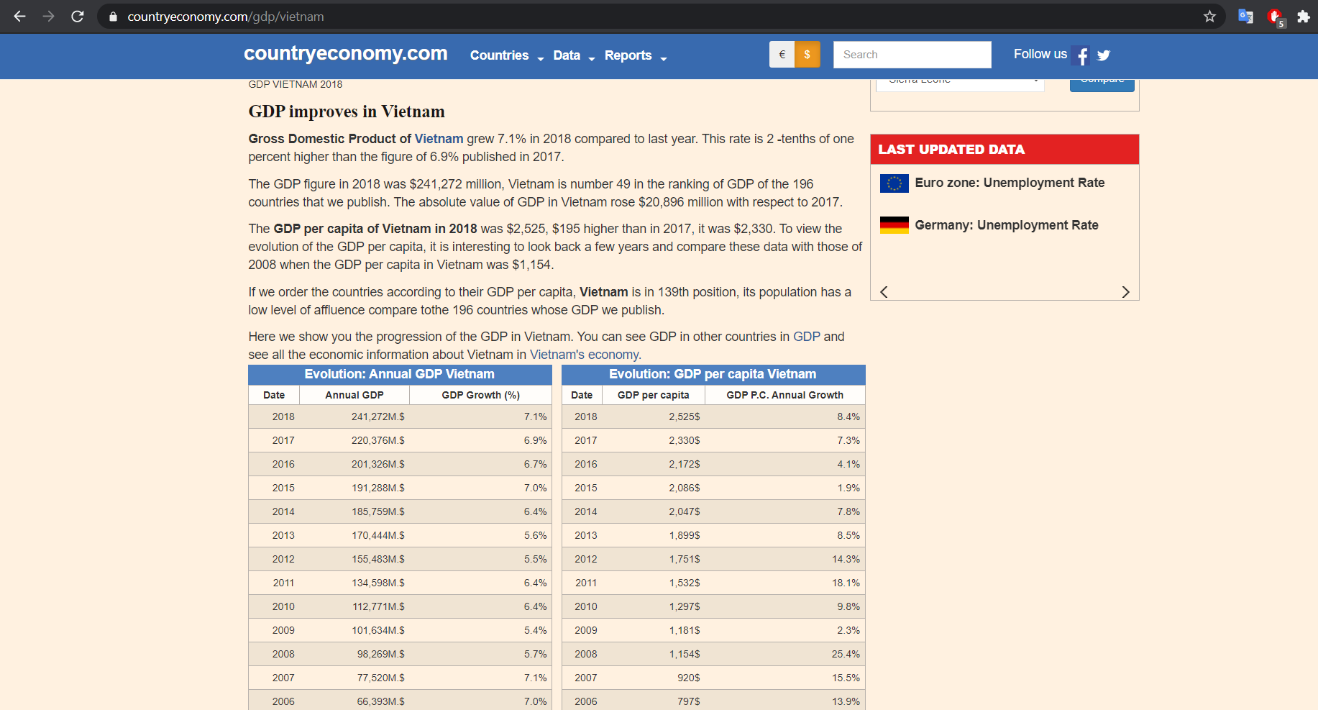
Ghi nhận góp ý của thầy:



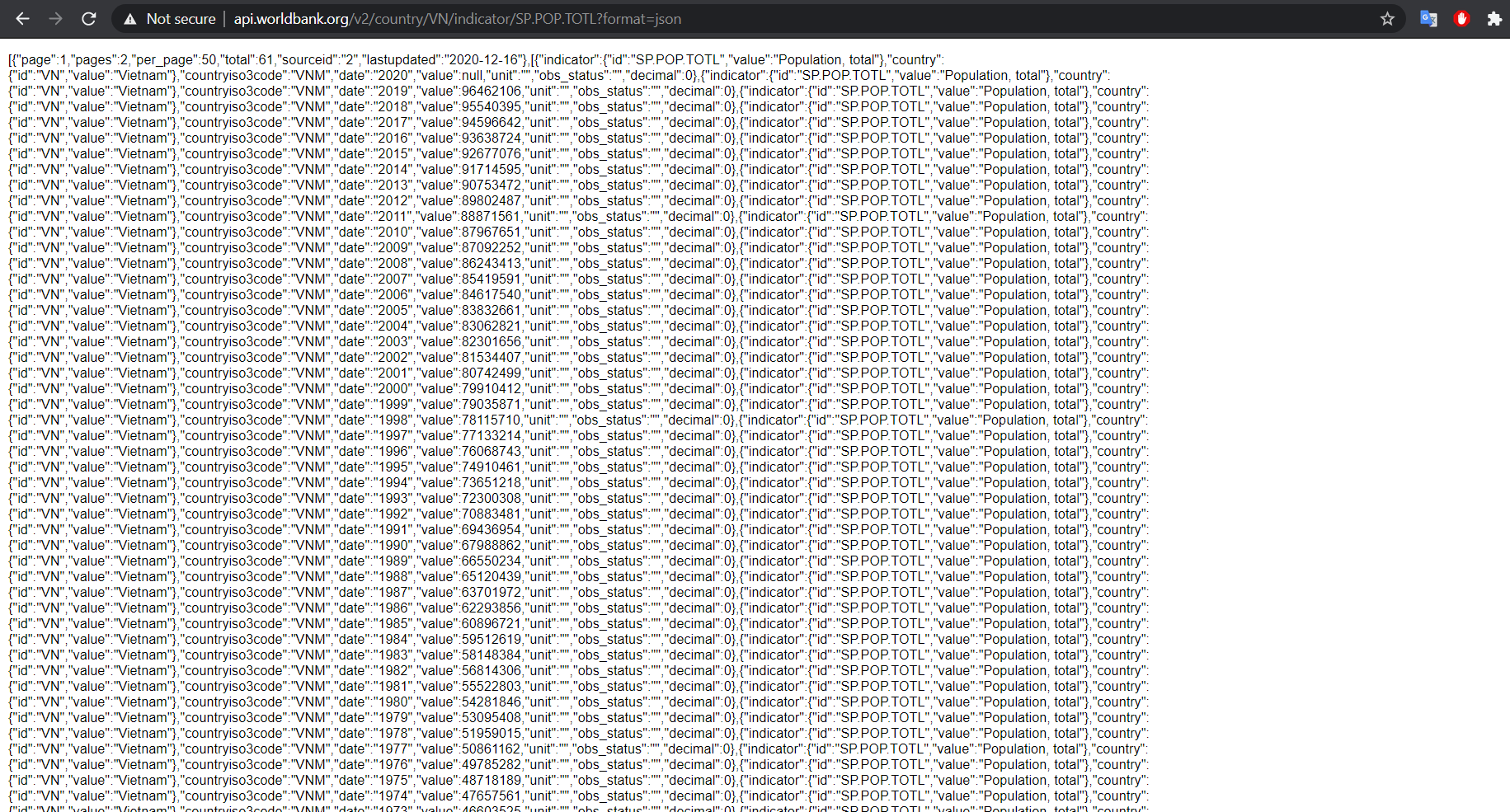
Thay đổi đề tài từ sản xuất gạo ở Việt Nam thành sản xuất gạo ở các quốc gia châu Á và mỹ latinh. Sau đó parse dữ liệu của các quốc gia này trên ricepedia về làm.



Bổ sung thêm thuộc tính mới là GDP bình quân đầu người. Thu thập dữ liệu này bằng các parse bảng trên website <https://countryeconomy.com>



Bổ sung thêm một thuộc tính mới nữa là dân số của quốc gia qua các năm. Thu thập dữ liệu này bằng các sử dụng API của world bank.



Cuối cùng là hoàn thiện mô hình với dữ liệu mới và viết báo cáo.