BÁO CÁO KẾT QUẢ CÔNG VIỆC

# Công việc được giao

* Xây dựng từ điển dựa trên dữ liệu về các điều luật
* Xây dựng chương trình kiểm tra lỗi chính tả cho văn bản đầu vào
* Xây dựng chương trình kiểm tra ngữ pháp cho văn bản đầu vào

# Nghiên cứu

## Thao tác với file dữ liệu JSON

* Đọc file kích thước lớn
* Truy cập
* Xác định cấu trúc, thành phần
* Thao tác với dữ liệu
* Lọc dữ liệu

## Xử lý ngôn ngữ tự nhiên

* Nguồn tham khảo: <https://huggingface.co/learn/nlp-course>
* Tiền xử lý văn bản
* Loại bỏ ký tự thừa
* Loại bỏ từ dừng
* Quy văn bản về định dạng chung

## Thư viện sử dụng

* pivy

# Xây dựng từ điển

## Đọc dữ liệu:

Chia file thành các file nhỏ (10GB -> 50MB) để mở bằng trình hiển thị và xác định cấu trúc dữ liệu thủ công







*\* Xác định được cấu trúc và thành phần file JSON: không phân cách dòng, các object là các document được chuyển đổi*

## Đọc file: Theo khối 1MB

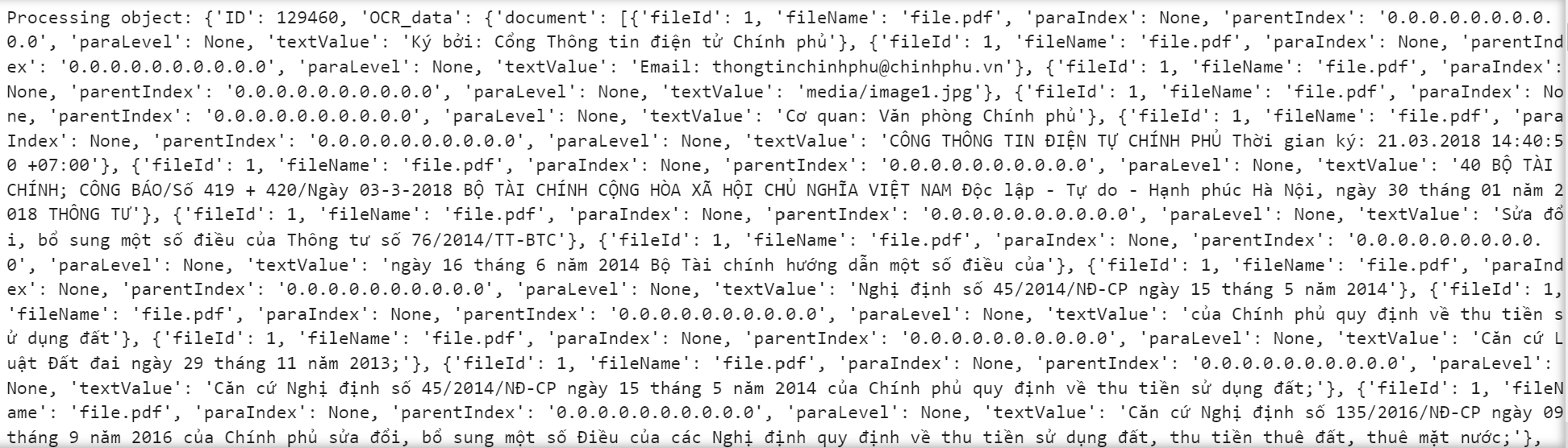


*\* Kiểm tra dữ liệu đúng định dạng mới tiến hành xử lý*

## Xử lý dữ liệu theo từng đối tượng

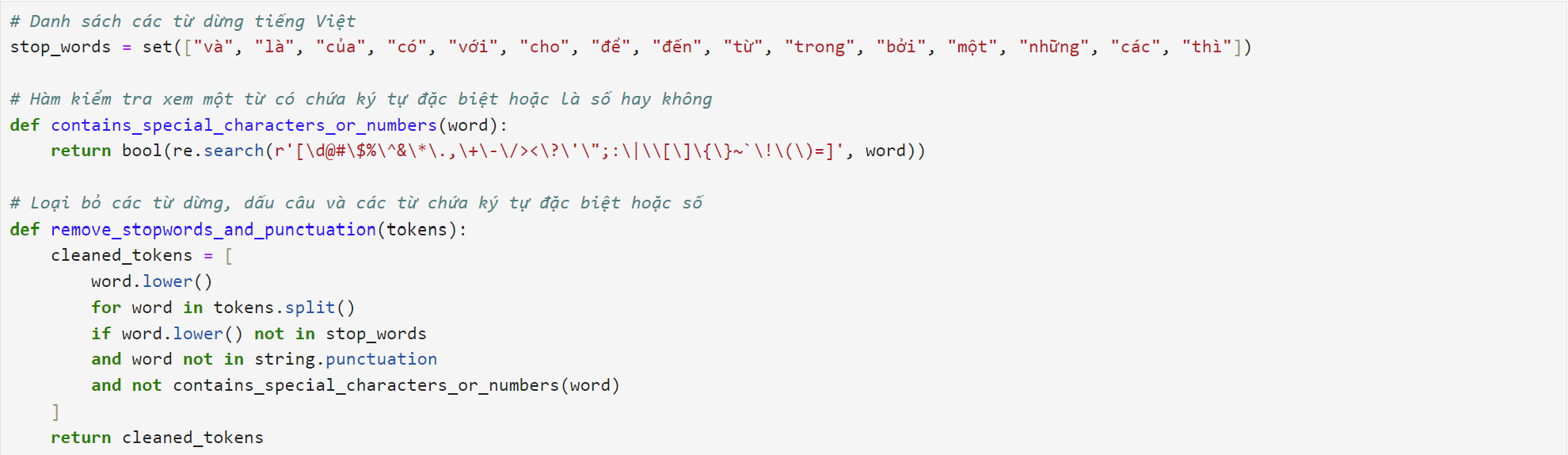


*\* Chuyển đổi văn bản OCR sang dạng thường*

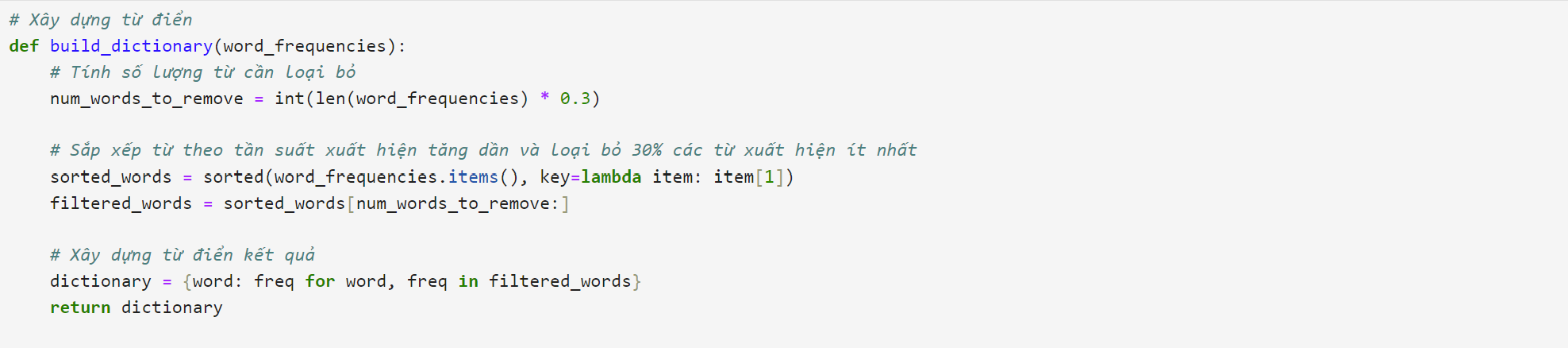


*\* Lọc chỉ xét dữ liệu nội dung (parentIndex != 0.0.0.0.0.0.0.0.0.0.0)*

*\* Chuyển về các token rồi làm sạch và đếm từ, cập nhật vào Counter word\_freq*

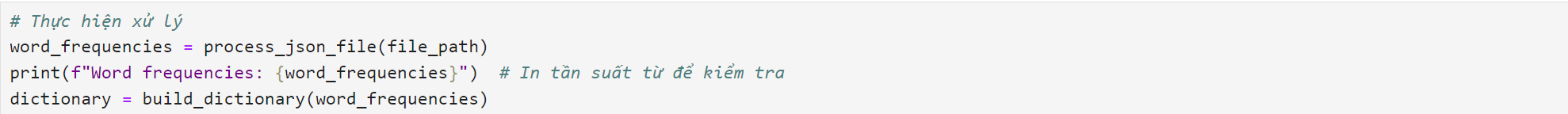


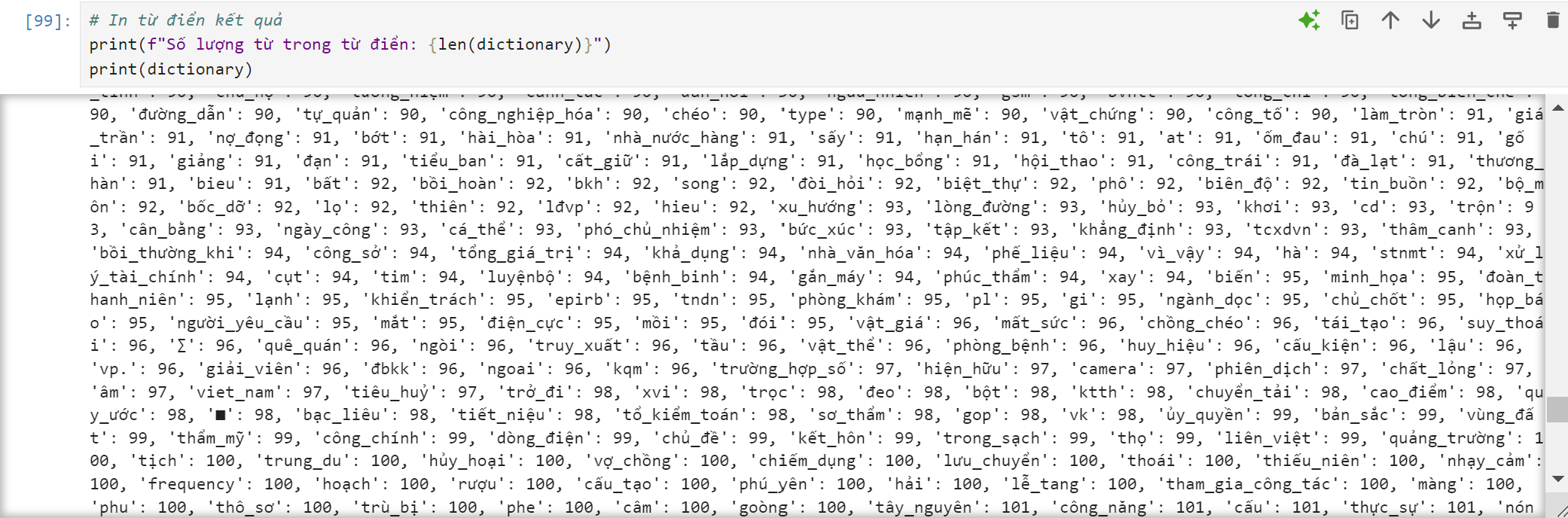
## Xây dựng từ điển

**

*\* Loại bỏ 30% số từ sắp xếp theo tần suất xuất hiện*

## Chạy

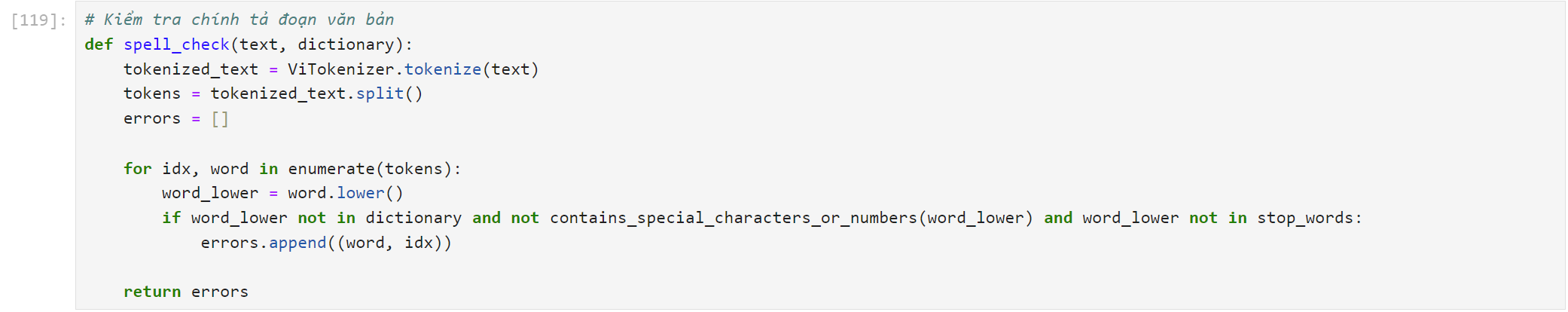




*\* Số lượng từ trong từ điển: 29883*

# Xây dựng chương trình kiểm tra chính tả

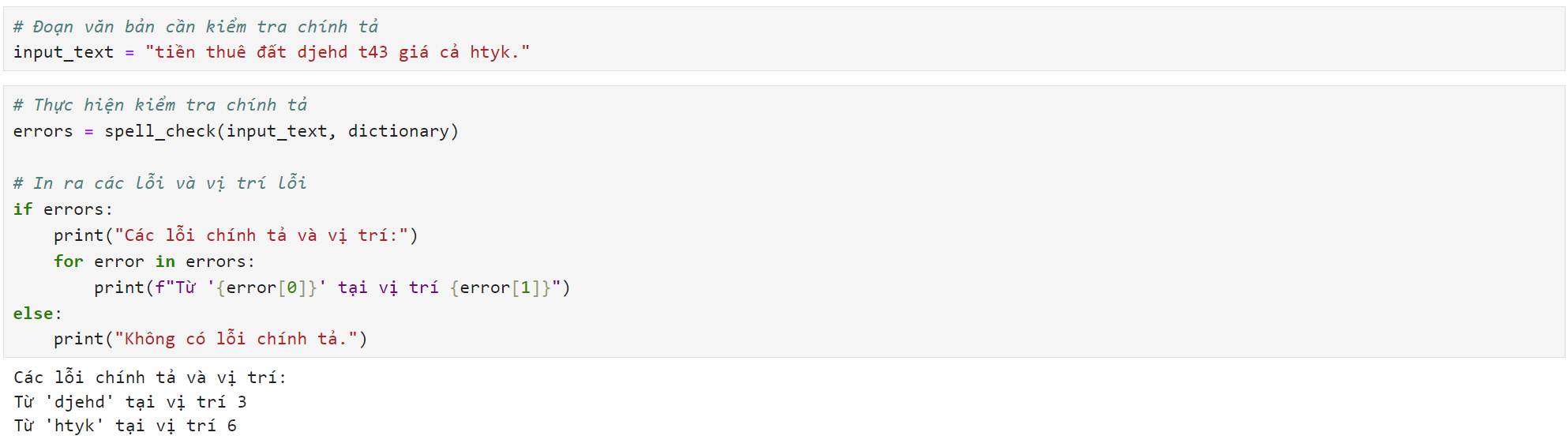
## Hàm kiểm tra



*\* Token hóa văn bản đầu vào*

*\* Quy về định dạng chung và làm sạch (bỏ các chuỗi chứa số, các ký tự đặc biệt, các từ dừng)*

## Kiểm thử



# Xây dựng chương trình kiểm tra ngữ pháp (Chưa tiến hành)

## Ý tưởng

Kiểm tra quan hệ của các từ đứng cạnh nhau (của + chủ thể, là + danh từ)

Kiểm tra cấu trúc câu: Trong câu phải có động từ ngăn cách chủ ngữ vị ngữ

## Vướng mắc

Em kém văn, tiếng Việt phức tạp biến ảo khó lường (Câu bị động, câu ghép, cụm danh từ, động từ, tính từ, kiểu từ trùng lặp)

VD: Tôi thích chạy bộ. (vừa danh từ vừa động từ)