

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HCM**

**VIỆN SƯ PHẠM KỸ THUẬT**

**Báo cáo giữa kỳ**

AMAZON TEXTRACT



Mã môn học & mã lớp: **CLCO332779\_22\_1\_03CLC**

Sinh viên thực hiện

**NHÓM 26**

Võ Quang Hưng 20110057

Nguyễn Minh Tuấn 20110594

Nguyễn Đình Duy 20110453

Tp. Hồ Chí Minh, tháng 10, năm 2022

Amazon Textract là một dịch vụ máy học (ML) tự động trích xuất văn bản, chữ viết tay và dữ liệu từ các tài liệu được quét. Dịch vụ này có thể xác định, hiểu rõ và trích xuất dữ liệu từ các biểu mẫu và bảng biểu chứ không đơn thuần chỉ nhận diện ký tự quang học (OCR). Ngày nay, rất nhiều công ty phải trích xuất thủ công dữ liệu từ tài liệu được quét như PDF, hình ảnh, bảng biểu và biểu mẫu hoặc thông qua các phần mềm OCR đơn giản yêu cầu cấu hình thủ công (thường phải cập nhật khi biểu mẫu thay đổi). Để loại bỏ những quy trình thủ công và tốn kém này, Textract sử dụng ML để đọc và xử lý mọi loại văn bản, trích xuất chính xác văn bản, chữ viết tay, bảng biểu và dữ liệu khác mà không cần thao tác thủ công. Bạn có thể nhanh chóng tự động hóa hoạt động xử lý tài liệu và thực hiện hành động dựa trên thông tin trích xuất được. Textract có thể trích xuất dữ liệu chỉ trong vài phút, thay vì nhiều giờ hoặc ngày liền.

Các chức năng:

+ Extract text from image (Trích xuất văn bản từ hình ảnh).

+ Key-value pairs & table information extraction from image (Thông tin của các biểu mẫu , bảng sẽ được trích xuất từ hình ảnh dưới dạng các cặp key-value ) .

+ Extract text from multi-page PDF & save it as CSV (Trích xuất văn bản từ nhiều trang PDF và lưu nó dưới dạng CSV).

+ Key-value pairs & table information extraction from PDF & save it as CSV (Thông tin của các biểu mẫu , bảng sẽ được trích xuất từ file pdf dưới dạng các cặp key-value và lưu nó dưới dạng CSV).

Form extraction: Bạn có thể phát hiện các cặp giá trị khóa trong hình ảnh tài liệu tự động và giữ lại bối cảnh mà không cần can thiệp thủ công. Một cặp giá trị khóa là một tập hợp các mục dữ liệu được liên kết. Chẳng hạn, trong một tài liệu, trường "Tên" là khóa và "Jane" là giá trị. Điều này giúp dễ dàng nhập dữ liệu được trích xuất vào cơ sở dữ liệu hoặc cung cấp nó như một biến trong một ứng dụng. Với các giải pháp OCR truyền thống, các khóa và giá trị được trích xuất dưới dạng văn bản đơn giản và mối quan hệ của chúng bị mất trừ khi các quy tắc được mã hóa cứng được viết và duy trì cho mỗi hình thức.

Table extraction: Amazon Textract bảo tồn thành phần của dữ liệu được lưu trữ trong các bảng trong quá trình trích xuất. Điều này rất hữu ích cho các tài liệu bao gồm dữ liệu có cấu trúc, chẳng hạn như báo cáo tài chính hoặc hồ sơ y tế với các bảng trong các cột và hàng. Bạn có thể tự động tải dữ liệu được trích xuất vào cơ sở dữ liệu bằng lược đồ được xác định trước. Ví dụ: các hàng số mục và số lượng trong báo cáo hàng tồn kho sẽ giữ lại liên kết của chúng để một ứng dụng quản lý hàng tồn kho có thể dễ dàng tăng tổng số mặt hàng.

Query based extraction: Amazon Textract cung cấp cho bạn tính linh hoạt để chỉ định dữ liệu bạn cần trích xuất từ ​​các tài liệu bằng các truy vấn. Bạn có thể chỉ định thông tin bạn cần dưới dạng câu hỏi ngôn ngữ tự nhiên (ví dụ: tên của khách hàng là gì) và nhận thông tin chính xác (ví dụ: John John Doe,) như một phần của phản hồi API. Bạn không cần biết cấu trúc dữ liệu trong tài liệu (bảng, biểu mẫu, trường ngụ ý, dữ liệu lồng nhau) hoặc lo lắng về các biến thể giữa các phiên bản và định dạng tài liệu. Các truy vấn Textract được đào tạo trước cho một loạt các tài liệu bao gồm PayStub, báo cáo ngân hàng, W-2S, mẫu đơn cho vay, ghi chú thế chấp, tài liệu yêu cầu và thẻ bảo hiểm. Tính linh hoạt mà các truy vấn Textract cung cấp làm giảm nhu cầu thực hiện xử lý bài, phụ thuộc vào các đánh giá thủ công của dữ liệu được trích xuất hoặc nhu cầu đào tạo các mô hình ML.

Handwriting recognition: Nhiều tài liệu, chẳng hạn như các hình thức tiếp nhận y tế và ứng dụng việc làm, bao gồm cả văn bản viết tay và in. Amazon Textract có thể trích xuất cả hai từ các tài liệu được viết bằng tiếng Anh với điểm số tự tin cao, cho dù văn bản là dạng tự do hoặc được nhúng trong các bảng. Tài liệu cũng có thể chứa một hỗn hợp văn bản được đánh máy và văn bản viết tay.

Invoices and receipts: Hóa đơn và biên lai có thể có nhiều cách bố trí, điều này gây khó khăn và tốn thời gian để trích xuất dữ liệu theo cách thủ công theo quy mô. Amazon Textract sử dụng máy học (ML) để hiểu bối cảnh của hóa đơn và biên lai và tự động trích xuất dữ liệu liên quan như tên nhà cung cấp, số hóa đơn, giá mặt hàng, tổng số tiền và điều khoản thanh toán.

Identity documents: Amazon Textract sử dụng Machine Learning (ML) để hiểu bối cảnh của các tài liệu nhận dạng như hộ chiếu Hoa Kỳ và giấy phép lái xe mà không cần các mẫu hoặc cấu hình. Bạn có thể tự động trích xuất thông tin cụ thể như ngày hết hạn và ngày sinh, cũng như xác định và trích xuất thông tin ngụ ý một cách thông minh như tên và địa chỉ. Sử dụng ID phân tích, các doanh nghiệp cung cấp dịch vụ xác minh ID và những dịch vụ tài chính, chăm sóc sức khỏe và bảo hiểm có thể dễ dàng tự động hóa việc tạo tài khoản, lập lịch hẹn, ứng dụng việc làm và hơn thế nữa bằng cách cho phép khách hàng gửi ảnh hoặc quét tài liệu nhận dạng của họ.

Các service dùng trong đề tài : Amazon textract , AWS IAM.