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

**KHOA CÔNG NGHỆ THÔNG TIN**

🙠🙟🕮🙝🙢



Ảnh có chứa quảng trường

Mô tả được tạo tự động

**BÁO CÁO ĐỒ ÁN**

**MÔN: ĐIỆN TOÁN ĐÁM MÂY**

**Đề tài:**

**TÌM HIỂU VÀ VIẾT ỨNG DỤNG MINH HỌA VỀ AWS TEXTRACT**

**GVHD:** *TS. Huỳnh Xuân Phụng*

**Lớp Học Phần:** *CLCO332799\_22\_1\_09*

**SVTH:**  *Lê Văn Anh Đức 20110634*

*Lại Văn Quý 20110708*

*Thành phố Hồ Chí Minh, tháng 12 năm 202*

**NHẬN XÉT**

## 

# LỜI CÁM ƠN

Để hoàn thành đề tài này, nhóm em xin gửi lời cảm ơn chân thành đến:

Ban giám hiệu trường Đại Học Sư phạm kỹ thuật Thành phố Hồ Chí Minh vì đã tạo mọi điều kiện về cơ sở vật chất, nền tảng trực tuyến, hiện đại, đa dạng các loại sách, tài liệu thuận lợi cho việc tìm kiếm, nghiên cứu thông tin.

Xin cảm ơn giảng viên bộ môn - Thầy Huỳnh Xuân Phụng đã giảng dạy tận tình, chi tiết để nhóm có đủ kiến thức và vận dụng chúng vào đề tài này.

Do chưa có nhiều kinh nghiệm làm để tài cũng như còn những hạn chế về kiến thức, nên đề tài chắc chắn sẽ không tránh khỏi những thiếu sót. Rất mong nhận được sự nhận xét, ý kiến đóng góp từ phía Thầy để đề tài của nhóm được hoàn thiện hơn

# MỤC LỤC

[LỜI CÁM ƠN 1](#_Toc121747651)

[MỤC LỤC 2](#_Toc121747652)

[DANH MỤC CÁC BẢNG 4](#_Toc121747653)

[MỞ ĐẦU 5](#_Toc121747654)

[1. Lí do chọn đề tài 5](#_Toc121747655)

[2. Mục đích nghiên cứu 5](#_Toc121747656)

[3. Đối tượng và phạm vi nghiên cứu 6](#_Toc121747657)

[*4.* Phương pháp nghiên cứu 6](#_Toc121747658)

[Chương 1: TÌM HIỂU VỀ AMAZON TEXTRACT 7](#_Toc121747659)

[1.1. Tổng quan về Amazon Textract 7](#_Toc121747660)

[1.1.1. Khái niệm 7](#_Toc121747661)

[1.1.2. Đặc điểm và tính năng 7](#_Toc121747662)

[1.1.3. Lợi ích của việc sử dụng Amazon Textract 8](#_Toc121747663)

[1.2. Hình thức hoạt động của Amazon Textract 9](#_Toc121747664)

[1.2.1. Xác định và phân tích tài liệu 9](#_Toc121747665)

[1.2.2. Giới hạn 10](#_Toc121747666)

[1.2.3. Chi phí sử dụng 11](#_Toc121747667)

[Chương 2: ỨNG DỤNG AMAZON TEXTRACT VÀO ĐỀ TÀI 12](#_Toc121747668)

[2.1. Giới thiệu sản phẩm của đề tài 12](#_Toc121747669)

[2.2. Kết quả thực nghiệm sản phẩm của đề tài: 13](#_Toc121747670)

[KẾT LUẬN 18](#_Toc121747671)

[1. Những kết quả đạt được 18](#_Toc121747672)

[2. Hạn chế 18](#_Toc121747673)

[3. Hướng phát triển 18](#_Toc121747674)

[3.1. Hướng khắc phục các hạn chế 18](#_Toc121747675)

[3.2. Hướng mở rộng đề tài 18](#_Toc121747676)

[DANH MỤC TÀI LIỆU THAM KHẢO 19](#_Toc121747677)

# DANH MỤC CÁC BẢNG

[Bảng 1: Bảng giới hạn của Amazon Textract 11](#_Toc90241765)

# MỞ ĐẦU

## Lí do chọn đề tài

Hiện nay ngày càng nhiều các công ty, doanh nghiệp, các đơn vị, tổ chức, cá nhân cần phải ứng dụng công nghệ thông tin vào hoạt động quản lý của mình. Các ngân hàng là một trong những tổ chức sử dụng nhận dạng kí tự quang học OCR để để tự động trích xuất văn bản, chữ viết tay và dữ liệu từ các tài liệu được quét từ hình ảnh. Số hóa tài liệu trong lĩnh vực ngân hàng là một tiện ích tuyệt vời. Nhiều ngân hàng sử dụng công nghệ để đạt được an ninh giao dịch và quản lý rủi ro tốt hơn.

Việc sử dụng phần mềm OCR trong doanh nghiệp cũng có thể quét nhiều tài liệu bảo lãnh viết tay quan trọng của khách hàng như tài liệu vay của họ và hơn thế nữa. Khả năng OCR tự động để nhập dữ liệu mang lại lợi ích như thế nào đối với hoạt động kinh doanh và quy trình làm việc. Tuy nhiên, các phần mềm OCR hoạt động kém với các tài liệu viết tay hoặc các tài liệu được đánh máy có phông chữ giống như chữ viết tay. Các tập tin của bạn phải ở trong tình trạng tốt hoặc chất lượng với ánh sáng, độ tương phản tốt, không bị bẩn hoặc nhòe. Nếu không, lỗi có thể được tạo ra trong các tệp đầu ra ví dụ như một số từ sẽ bị thiếu. Chính vì vậy, ta sẽ tìm hiểu về một cách thức hiệu quả, chính xác hơn, đó là dịch vụ Amazon Textract, đây không chỉ là OCR vì nó được thiết kế để phân tích và trích xuất dữ liệu từ các biểu mẫu, bảng và các tài liệu khác. Nó có khả năng trích xuất các cặp key-value quan trọng, bảng và các chuỗi khóa khác.

Đối với các công ty vẫn thực hiện nhập dữ liệu thủ công, Textract có thể giúp bạn tiết kiệm tiền rất nhiều tiền bạc, giảm cả thời gian làm việc để gõ trên bàn phím và thực tế là nó có thể xử lý hàng loạt nhiều mục cùng một lúc, làm tăng đáng kể tốc độ nhập dữ liệu.

## Mục đích nghiên cứu

Tìm hiểu về khái niệm, đặc điểm và phương thức hoạt động của việc nhận dạng, trích xuất dữ liệu của Amazon Textract. Từ đó ứng dụng dịch vụ này một cách hiệu quả, tiết kiệm chi phí vào việc phân tích tài liệu để cho kết quả chính xác và đầy đủ.

## Đối tượng và phạm vi nghiên cứu

*Đối tượng nghiên cứu của đề tài*

- Cloud Computing (Điện toán đám mây)

- Amazon Web Services (AWS)

- Dịch vụ Amazon Textract

*Phạm vi nghiên cứu*

Các vấn đề cơ bản của Amazon Textract, phương thức hoạt động, chi phí sử dụng và ứng dụng của Amazon Textract vào thực tiễn.

## Phương pháp nghiên cứu

Nghiên cứu các tài liệu, giáo trình, bài báo về điện toán đám mây, dịch vụ Amazon Textract đã được công bố và các vấn đề khác có liên quan.

*Kết cấu của báo cáo*

* Chương 1. Tìm hiểu về Amazon Textract
* Chương 2. Ứng dụng Amazon Textract vào đề tài

# Chương 1: TÌM HIỂU VỀ AMAZON TEXTRACT

## 1.1. Tổng quan về Amazon Textract

### 1.1.1. Khái niệm

Amazon Textract là Amazon Textract là một dịch vụ máy học (ML) sử dụng tính năng nhận diện ký tự quang học (OCR) để tự động trích xuất văn bản, chữ viết tay và dữ liệu từ các tài liệu được quét như PNG.

Với Amazon Textract, ta chỉ phải trả tiền cho những gì ta sử dụng. Không yêu cầu mức phí tối thiểu và không cần cam kết trả trước. Bất kể là ta trích xuất văn bản, văn bản dạng bảng và/hoặc dữ liệu biểu mẫu, Amazon Textract cũng chỉ tính phí cho các trang đã qua xử lý.

Amazon Textract giúp ta dễ dàng thêm tính năng phát hiện và phân tích văn bản tài liệu vào các ứng dụng của mình. API phát hiện văn bản Amazon Textract có thể phát hiện văn bản trong nhiều loại tài liệu bao gồm báo cáo tài chính, hồ sơ y tế và biểu mẫu thuế.

### 1.1.2. Đặc điểm và tính năng

Đối với tài liệu có dữ liệu có cấu trúc, ta có thể sử dụng API phân tích tài liệu của Amazon Textract để trích xuất văn bản, biểu mẫu và bảng. Amazon Textract dựa trên cùng một công nghệ Deep Learning (học sâu) đã được chứng minh, có khả năng mở rộng cao, được phát triển bởi các nhà khoa học thị giác máy tính của Amazon để phân tích hàng tỷ hình ảnh và video hàng ngày.

Amazon Textract bao gồm các API đơn giản, dễ sử dụng có thể phân tích dữ liệu hình ảnh, văn bản và nhiều kiểu dữ liệu tệp khác nhau.

Dưới đây là các trường hợp sử dụng phổ biến sử dụng Amazon Textract:

- Tạo một tìm kiếm thông minh: Amazon Textract cho phép ta tạo thư viện văn bản được phát hiện trong tệp hình ảnh và PNG.

- Sử dụng tính năng trích xuất văn bản thông minh để xử lý ngôn ngữ tự nhiên (NLP): ta có thể sử dụng Amazon Textract để trích xuất văn bản thành các từ và dòng. Nó cũng nhóm văn bản theo các ô trong bảng nếu tính năng phân tích bảng tài liệu Amazon Textract được bật. Amazon Textract cung cấp cho ta quyền kiểm soát cách văn bản được nhóm làm đầu vào cho NLP.

- Tăng tốc thu thập và chuẩn hóa dữ liệu từ các nguồn khác nhau: Amazon Textract cho phép trích xuất dữ liệu dạng bảng và văn bản từ nhiều loại tài liệu, chẳng hạn như tài liệu tài chính, báo cáo nghiên cứu và ghi chú y tế. Với API phân tích tài liệu của Amazon Textract, ta có thể dễ dàng và nhanh chóng trích xuất dữ liệu có cấu trúc và phi cấu trúc từ tài liệu của mình.

- Tự động thu thập dữ liệu từ các biểu mẫu: Amazon Textract cho phép trích xuất dữ liệu có cấu trúc từ các biểu mẫu. Với API phân tích văn bản của Amazon, ta có thể xây dựng khả năng trích xuất vào quy trình công việc kinh doanh hiện có để dữ liệu người dùng được gửi thông qua biểu mẫu có thể được trích xuất thành định dạng có thể sử dụng được.

### 1.1.3. Lợi ích của việc sử dụng Amazon Textract

**Một số lợi ích của việc sử dụng Amazon Textract bao gồm:**

- Tích hợp tính năng phát hiện văn bản tài liệu vào ứng dụng của ta**:** Amazon Textract loại bỏ sự phức tạp của việc xây dựng khả năng phát hiện văn bản vào ứng dụng của ta bằng cách cung cấp phân tích mạnh mẽ và chính xác với một API đơn giản. Ta không cần thị giác máy tính hoặc kiến ​​thức chuyên môn sâu để sử dụng tính năng phát hiện văn bản tài liệu của Amazon Textract. Với API văn bản của Amazon Textract, ta có thể dễ dàng xây dựng tính năng phát hiện văn bản vào bất kỳ ứng dụng web, thiết bị di động hoặc thiết bị được kết nối nào.

- Phân tích tài liệu có thể mở rộng: Amazon Textract cho phép ta phân tích và trích xuất dữ liệu nhanh chóng từ hàng triệu tài liệu, điều này có thể đẩy nhanh quá trình ra quyết định.

- Chi phí thấp: Với Amazon Textract, ta chỉ trả tiền cho các tài liệu ta phân tích. Không có phí tối thiểu hoặc cam kết trả trước. Ta có thể bắt đầu miễn phí và tiết kiệm nhiều hơn khi phát triển với mô hình định giá theo tầng của Amazon Textract.

Với quá trình xử lý đồng bộ, Amazon Textract có thể phân tích các tài liệu một trang cho các ứng dụng có độ trễ là rất quan trọng. Amazon Textract cũng cung cấp các hoạt động không đồng bộ để mở rộng hỗ trợ cho các tài liệu nhiều trang.

## 1.2. Hình thức hoạt động của Amazon Textract

Amazon Textract cho phép ta phát hiện và phân tích văn bản trong các tài liệu đầu vào đơn trang hoặc nhiều trang. Amazon Textract cung cấp các hoạt động để phát hiện văn bản và các hoạt động để phân tích văn bản phức tạp nhằm tìm ra các mối quan hệ dữ liệu sâu hơn, chẳng hạn như dữ liệu biểu mẫu và bảng.

Amazon Textract cung cấp các hoạt động đồng bộ để xử lý các tài liệu nhỏ, một trang và để nhận được phản hồi gần như thời gian thực.

Amazon Textract cũng cung cấp các thao tác không đồng bộ mà ta có thể sử dụng để xử lý các tài liệu nhiều trang, lớn hơn. Các phản hồi không đồng bộ không theo thời gian thực.

Khi một thao tác Amazon Textract xử lý một tài liệu, kết quả sẽ được trả về trong một mảng của phần được gọi là các đối tượng “Block”. Đối tượng Block chứa thông tin được phát hiện về các mục, bao gồm vị trí của chúng trên tài liệu và mối quan hệ của chúng với các mục khác trên tài liệu.

### 1.2.1. Xác định và phân tích tài liệu

Khi Amazon Textract xử lý tài liệu, nó sẽ tạo ra một danh sách các khối Block Object để xác định hoặc phân tích dữ liệu văn bản. Mỗi khối chứa thông tin về dữ liệu được xác định, vị trí của chúng và mối quan hệ giữa các dữ liệu này với nhau.

Một tài liệu thông thường sẽ được tạo dựng theo các kiểu Block Object phổ biến sau: PAGE, LINE và WORD , TABLE,CELL và SELECTION ELEMENT.

* 1. Xác định tài liệu

Amazon Textract cung cấp các hoạt động đồng bộ và không đồng bộ chỉ trả về văn bản được phát hiện trong tài liệu. Đối với cả hai kiểu hoạt động trên, thông tin sau được trả về thành các khối Block Objects như:

* Các dòng và từ của văn bản trong một văn bản.
* Mối quan hệ giữa các dòng và từ của văn bản được phát hiện
* Trang mà văn bản được phát hiện xuất hiện trên đó
* Vị trí của các dòng và từ của văn bản trên trang tài liệu
  1. Phân tích tài liệu

Amazon Textract có thể phân tích các tài liệu và biểu mẫu để tìm mối quan hệ giữa các văn bản được phát hiện. Các thao tác phân tích textract trả về 3 danh mục trích xuất văn bản bao gồm: văn bản, biểu mẫu và bảng.

**Trích xuất dạng văn bản**

Văn bản thô được trích xuất từ ​​một tài liệu.

**Trích xuất dạng biểu mẫu**

Dữ liệu biểu mẫu là các mục văn bản được liên kết được trích xuất từ ​​một tài liệu. Amazon Textract biểu thị dữ liệu biểu mẫu dưới dạng các cặp giá trị chính.

**Trích xuất dạng bảng Table**

Amazon Textract có thể trích xuất bảng, ô bảng và các mục trong ô bảng.

Ta có thể sử dụng các thao tác đồng bộ hoặc không đồng bộ để phân tích văn bản trong tài liệu bằng các thao tác phổ biến:

Để phân tích văn bản một cách đồng bộ, sử dụng thao tác **AnalyzeDocument** và chuyển một tài liệu làm đầu vào. AnalyzeDocument trả về toàn bộ tập hợp kết quả.

Để phát hiện văn bản không đồng bộ, hãy sử dụng **StartDocumentAnalysis** để bắt đầu xử lý.

Để có được kết quả, hãy gọi **GetDocumentAnalysis.** Kết quả sẽ được trả về trong một hoặc nhiều câu trả lời.

### 1.2.2. Giới hạn

Tuy có nhiều đặc điểm, tính năng nhưng Amazon Textract sẽ có những điểm hạn chế như sau:

|  |  |
| --- | --- |
| **Giới hạn** | **Mô tả** |
| Định dạng tệp cho phép | - Xử lý hoạt động đồng bộ với: PNG & JPEG (không hỗ trợ JPEG2000)  - Xử lý hoạt động không đồng bộ với: PNG & JPEG (không hỗ trợ JPEG2000) |
| Kích cỡ tệp | - JPEG & PNG: 10MB |
| Căn lề | - Tài liệu văn bản căn lề theo chiều ngang  - Không hỗ trợ văn bản căn lề theo chiều dọc |
| Ngôn ngữ hỗ trợ | - Tiếng Anh, Pháp, Đức, Ý, Bồ Đào Nha và Tây Ban Nha |
| Kích cỡ đối tượng | - Độ cao tối đa nên là 15 pixels |
| Kiểu dữ liệu đối tượng | - Hỗ trợ cả chữ viết tay và tệp in ấn |

Bảng 1: Bảng giới hạn của Amazon Textract

### 1.2.3. Chi phí sử dụng

Amazon Textract có ba API khác nhau: API Detect Document Text, API Analyze Document và API Analyze Expense.

*API Detect Document Text:* sử dụng công nghệ OCR để trích xuất văn bản và chữ viết tay từ tài liệu được cung cấp

*API Analyze Document:* có hai chức năng, gồm biểu mẫu và bảng, với các mức định giá khác nhau.

- API Analyze Document cho bảng sử dụng công nghệ OCR để trích xuất văn bản và chữ viết tay từ tài liệu được cung cấp. API này cũng trích xuất dữ liệu được trình bày ở dạng bảng hoặc bảng được tổ chức theo cột và hàng.

API Analyze Expense: trích xuất dữ liệu từ các hóa đơn và biên nhận.

*Ví dụ: Tính giá API Phát hiện văn bản tài liệu – Detect Document Text*

Giả sử bạn muốn trích xuất văn bản từ 100.000 trang báo cáo nghiên cứu bằng API Detect Document Text. Giá mỗi trang ở khu vực Miền Tây Hoa Kỳ (Oregon) cho một triệu trang đầu tiên là 0,0015 USD, vì vậy, để xử lý 100.000 trang, tổng chi phí mỗi tháng phải trả là 150 USD. Xem cách tính dưới đây:

Tổng số trang đã xử lý = 100.000

Giá mỗi trang = 0,0015 USD

Tổng mức phí mỗi tháng = 0,0015 USD\*100.000 = 150 USD

# Chương 2: ỨNG DỤNG AMAZON TEXTRACT VÀO ĐỀ TÀI

## 2.1. Giới thiệu sản phẩm của đề tài

Trong cuộc sống, vấn đề chuyển đổi tài liệu từ dạng hình ảnh để sử dụng, lưu trữ các thông tin quan trọng là một điều cấp thiết cho con người, nhất là các bạn sinh viên, với việc thường xuyên tiếp cận với các tài liệu học tập, nghiên cứu, tham khảo với đa dạng các định dạng, trong đó có cả dịnh dạng hình ảnh PNG, JPEG.

Ứng dụng minh họa ra đời nhằm hỗ trợ tối ưu giúp các tổ chức, cá nhận, đặc biệt là sinh viên quét và xử lý tài liệu trong thực tiễn học tập, làm việc - có ứng dụng Amazon Textract nhận dạng, phân tích và trích xuất tài liệu để đọc dữ liệu và xác định các bảng trong tài liệu PNG, JPEG,.

Ứng dụng này sử dụng ngôn ngữ Python.

Nền tảng công nghệ:

**- Boto3** giúp ta dễ dàng tích hợp ứng dụng, thư viện hoặc tập chỉ lệnh với các dịch vụ AWS, bao gồm Amazon S3, Amazon EC2, Amazon DynamoDB và nhiều dịch vụ khác.

**- Flask Framworks** cung cấp cho ta tất cả các công cụ, thư viện và công nghệ cho phép bạn xây dựng một ứng dụng web.

**- Amazon Textract**

**\*Source code : https://github.com/HiImLeVanAnhDuc/Amazon-Textract-Project-IC-**

Với các chức năng chính như sau:

1. Đăng nhập tài khoản

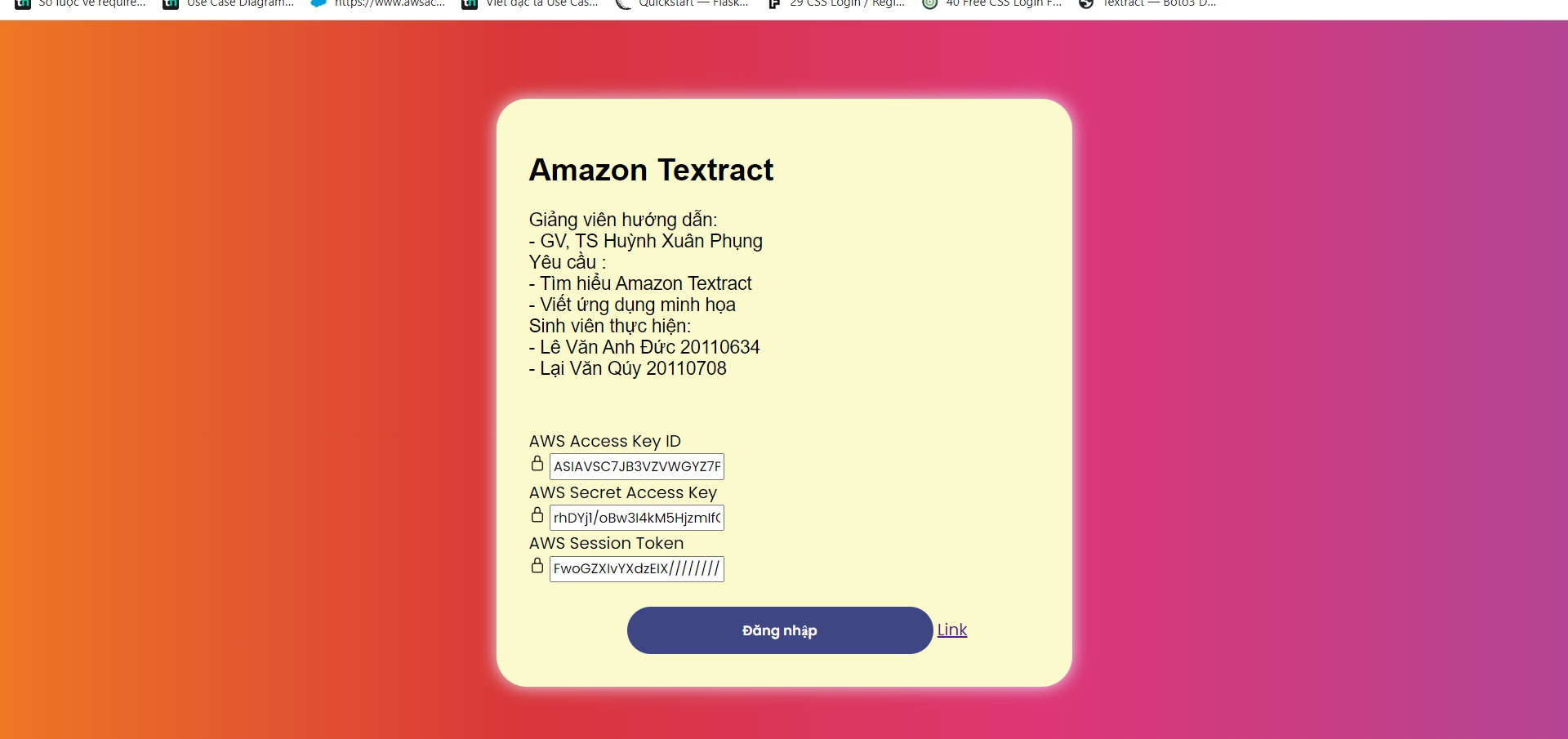
2. Nhận dạng tài liệu dạng .PNG, .JPEG

3. Trích xuất tài liệu dạng bảng chuẩn

4. Chuyển đổi tài liệu hình ảnh sang dạng văn bản .docx

5. Chuyển đổi tài liệu hình ảnh sang dạng .csv

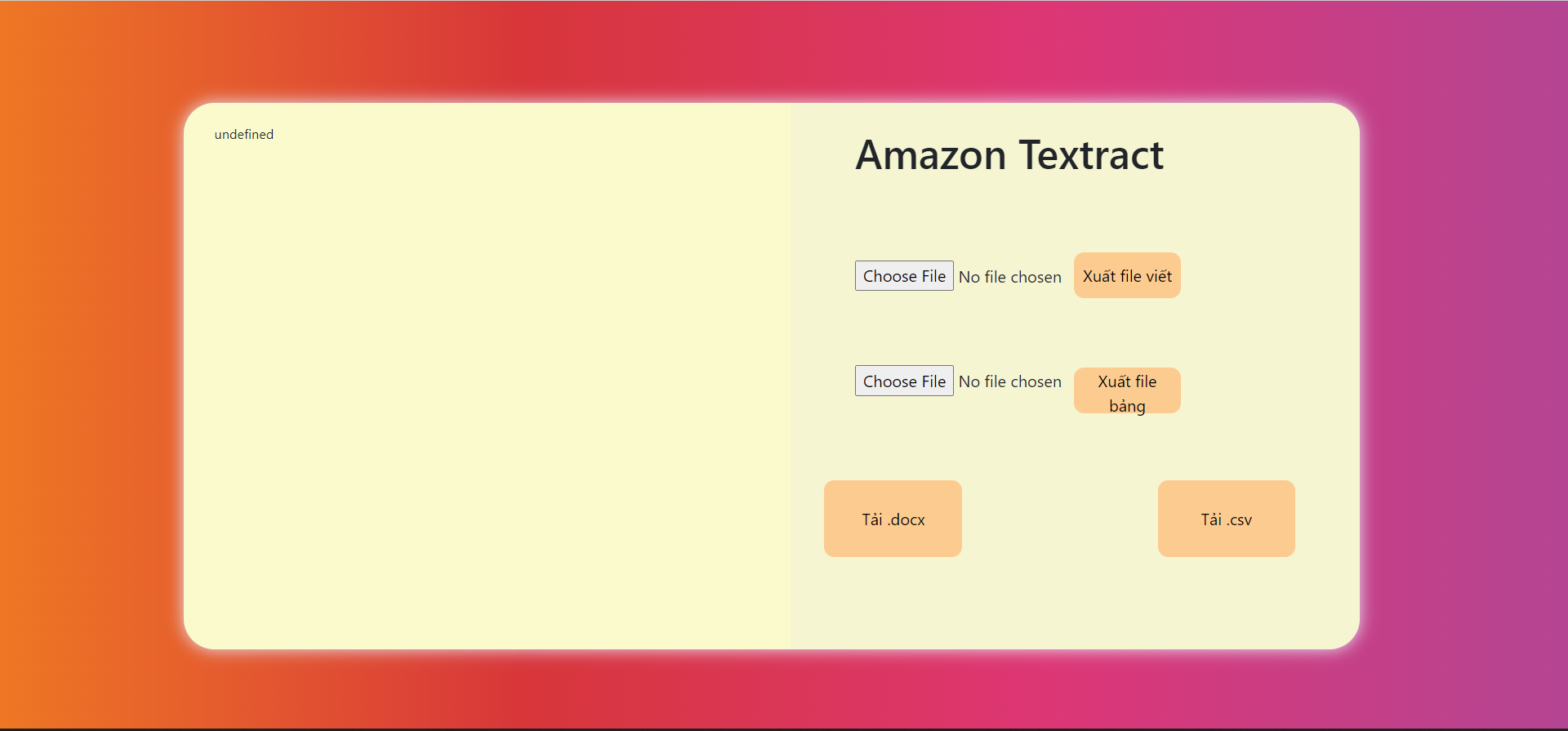
## 2.2. Kết quả thực nghiệm sản phẩm của đề tài:



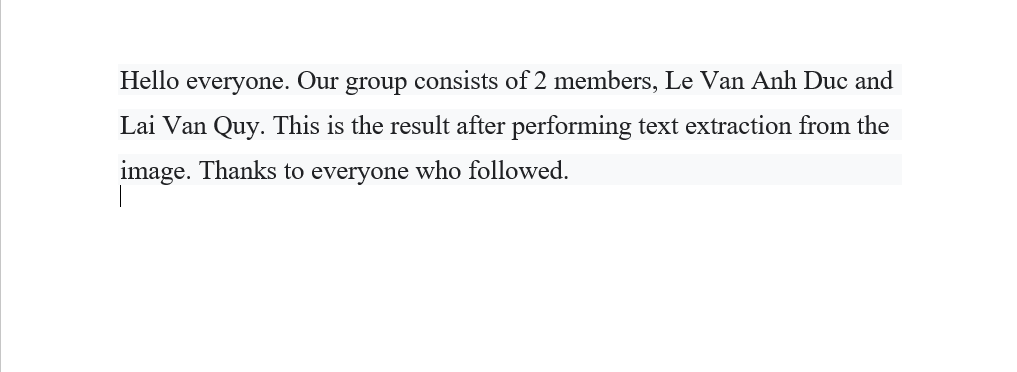
Giao diện trang đăng nhập

Thực hiện đăng nhập bằng các thông số aws\_access\_key\_id, aws\_secret\_access\_key

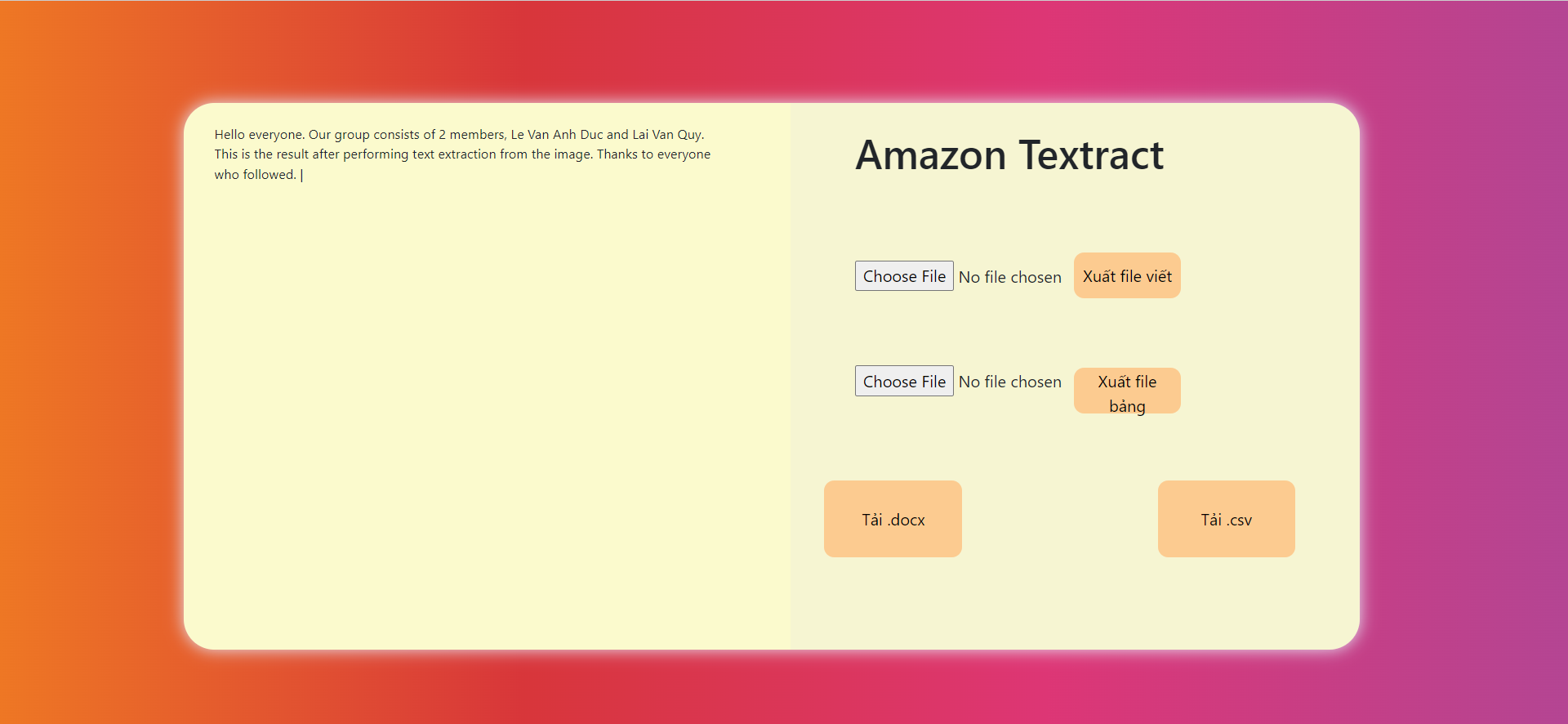
, aws\_session\_token để sử dụng các dịch vụ của AWS.



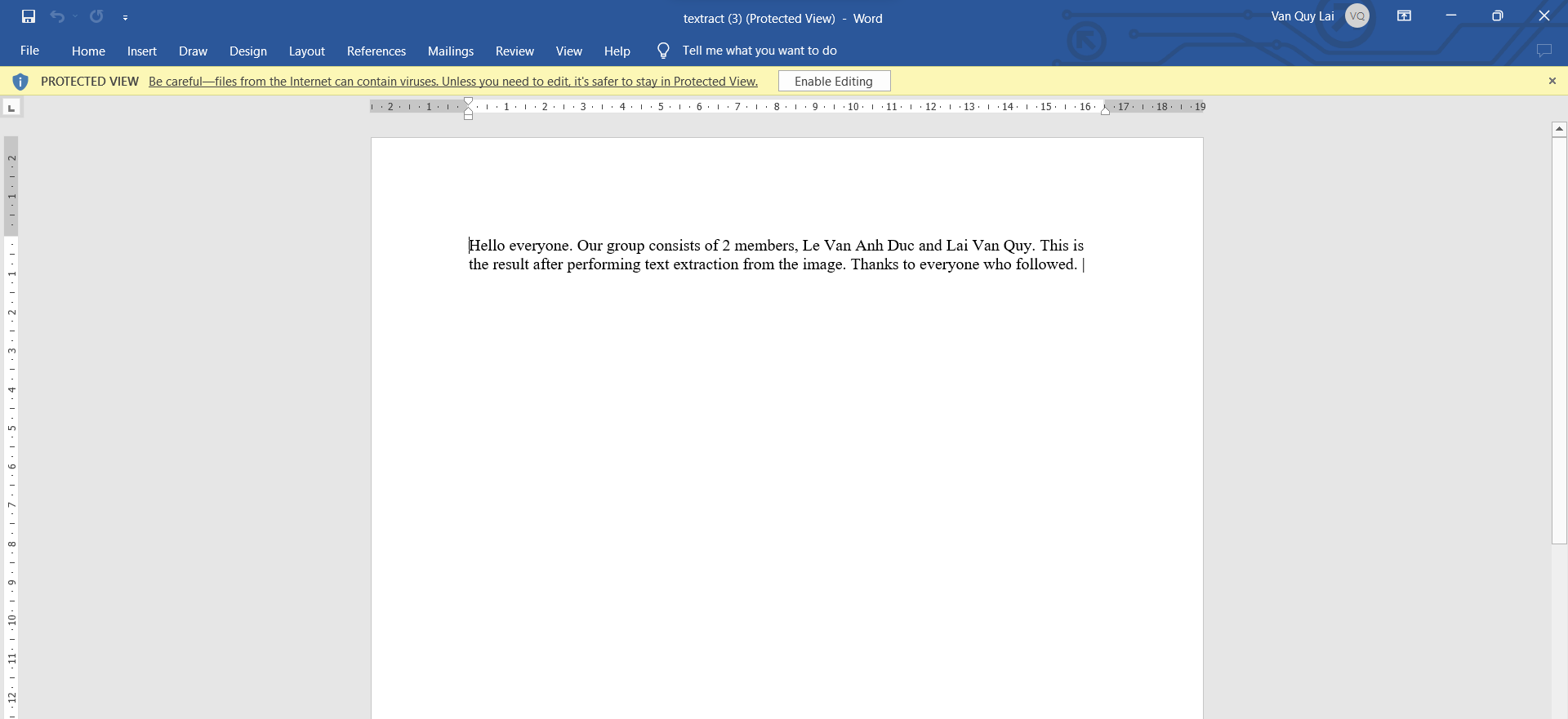
Giao diện trang chính của ứng dụng



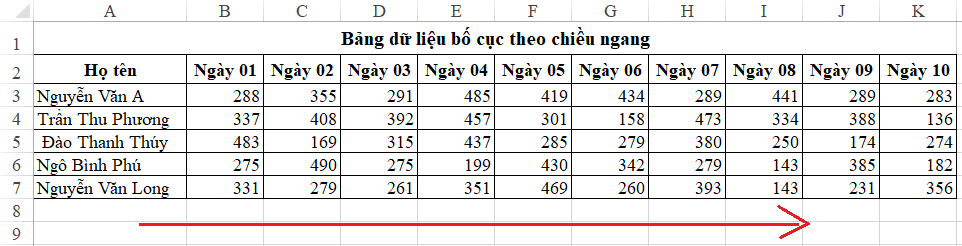
Ảnh demo chức năng trích xuất văn bản



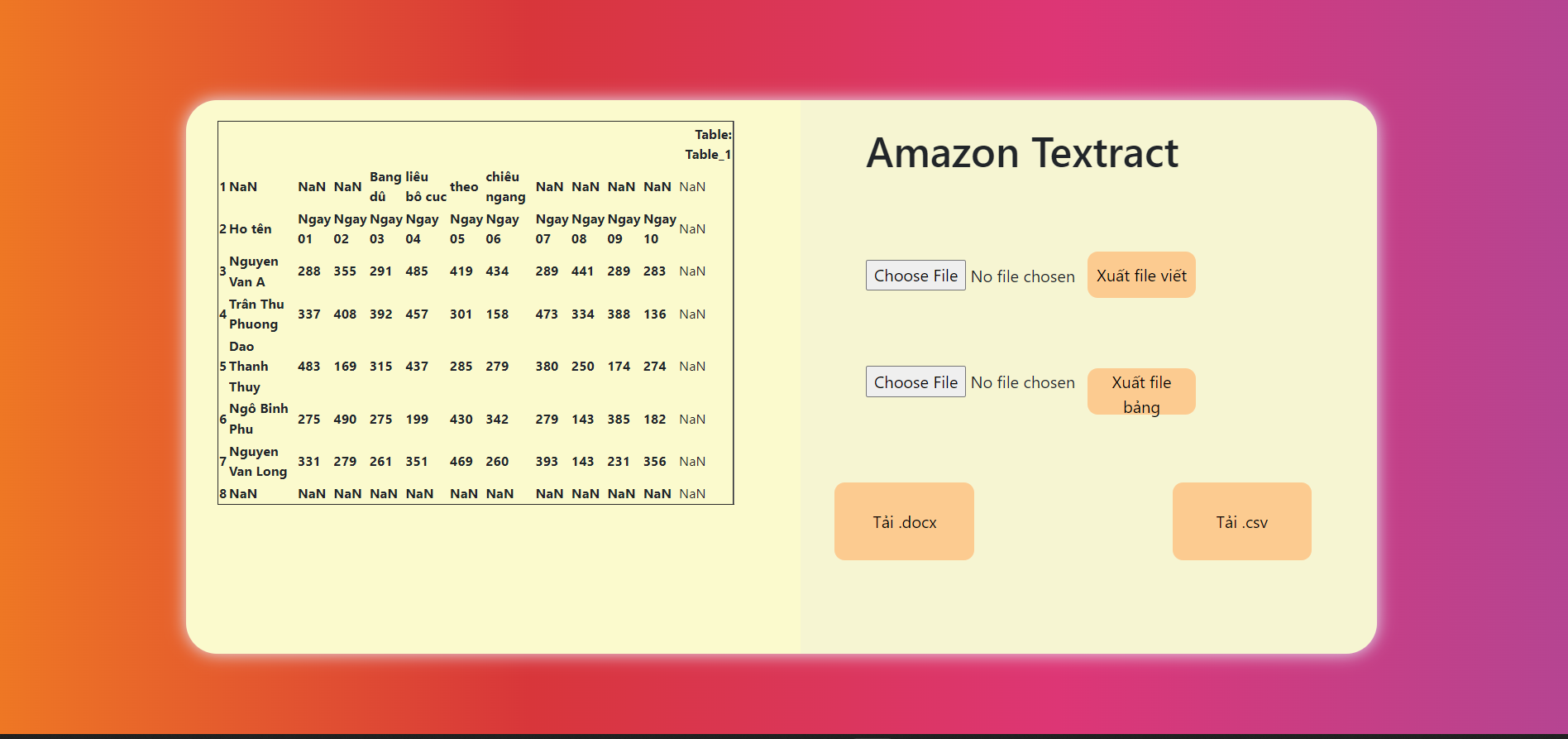
Tính năng xuất định dạng từ tệp .png sang tệp .docx



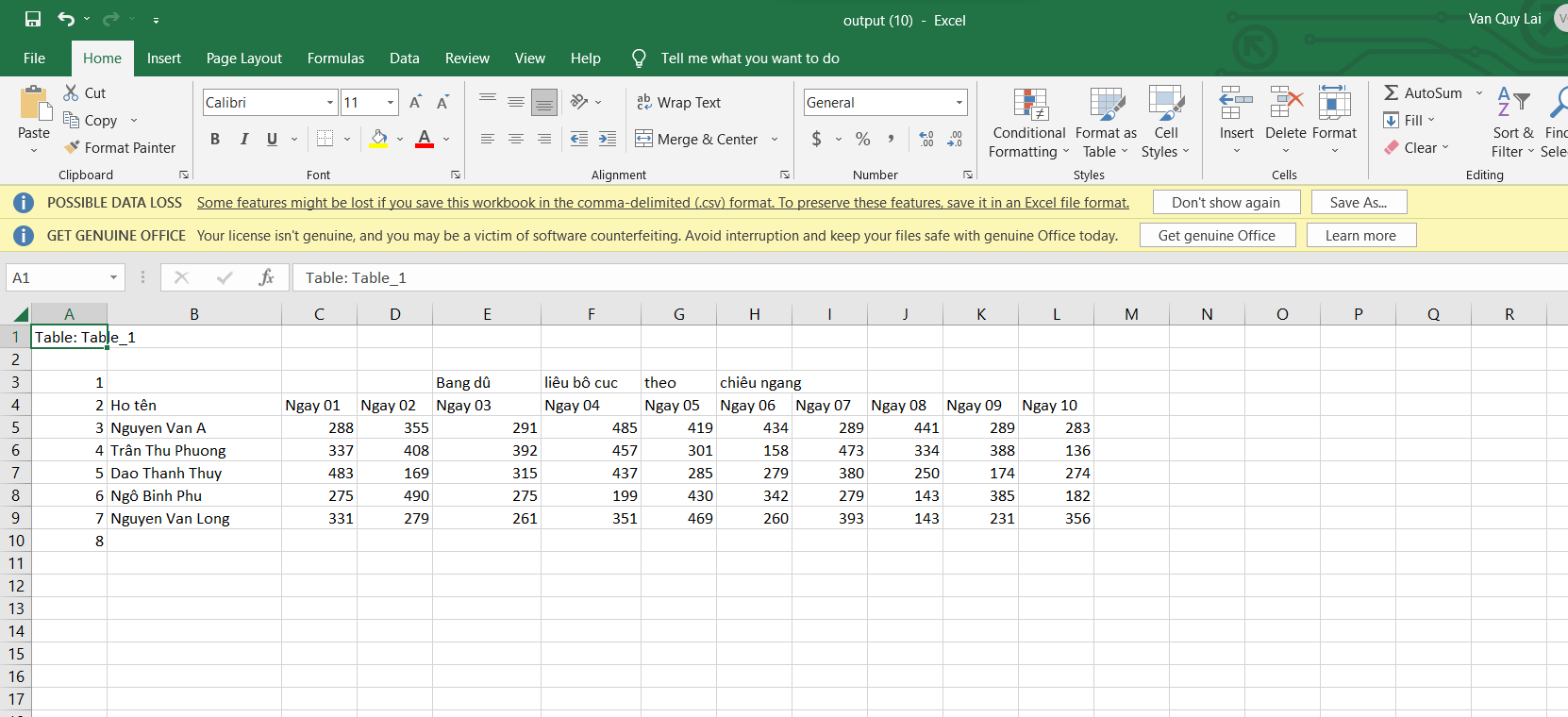
Tệp hình văn bản .png sau khi được Extract thành dạng .doc



Ảnh ban đầu để demo chức năng



Kết quả sau khi trích xuất bảng trên giao diện web



Tải kết quả trích xuất về dạng csv

KẾT LUẬN

Báo cáo trên đã tìm hiểu tổng quan về dịch vụ Amazon Textract, mô tả cách thức hoạt động, giới hạn, chi phí sử dụng và ứng dụng thực tiễn của đề tài để nhận dạng, phân tích và trích xuất tài liệu cho các doanh nghiệp tổ chức, cá nhân cần sử dụng để xử lý, lưu trữ dữ liệu.

## 1. Những kết quả đạt được

- Nhận dạng, phân tích được tài liệu đầu vào dạng PNG, JPEG

- Nhận dạng, phân tích được tài liệu dạng Table (Bảng) từ hình ảnh sang dạng Table chuẩn.

- Chuyển đổi được tài liệu từ dạng ảnh sang dạng tệp văn bản .doc

- Chuyển đổi được tài liệu từ dạng ảnh sang dạng tệp trang tính .csv

- Kết quả nhận dạng, phân tích và chuyển đổi định dạng với độ chính xác cao.

## 2. Hạn chế

- Kết quả nhận dạng kém với tài liệu viết tay hoặc tài liệu đánh máy có phông chữ viết tay

- Xử lý kém với các tệp hình ảnh có chất lượng phân giải thấp, thiếu sáng, độ tương phản thấp hoặc bị bẩn, nhòe.

- Chưa phân tích và chuyển đổi tài liệu thành dạng biểu mẫu, hóa đơn, giấy tờ tùy than,...

## 3. Hướng phát triển

### 3.1. Hướng khắc phục các hạn chế

- Cải thiện khả năng nhận dạnh tài liệu có chữ viết tay hay các tài liệu đánh máy có sử dụng phông chữ dạng viết tay

- Cải thiện khả năng nhận dạng các tài liệu phức tạp, chất lượng hình ảnh tài liệu kém.

### 3.2. Hướng mở rộng đề tài

- Thêm giao diện để xem tài liệu trực tiếp sau khi phân tích thành các dạng khác

- Thêm tính năng cho phép tải lên và xử lý nhiều tệp tài liệu cùng một lúc

- Thêm tính năng nhận dạng, phân tích bảng giá, chi phí cho hóa đơn giấy, hóa đơn điện tử, giấy tờ tùy thân

# DANH MỤC TÀI LIỆU THAM KHẢO

[1]: Amazon Web Services Inc, Amazon Textract Developer Guide, 2019

[2]: Stephen V.Rice, Optical Character Recognition: An Illustrated Guide to Frontier, Springer, 1999