

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT THÀNH PHỐ HỒ CHÍ MINH
KHOA ĐÀO TẠO CHẤT LƯỢNG CAO



HCMUTE
ĐỒ ÁN MÔN HỌC CUỐI KỲ



AMAZON TEXTTRACT
Mã môn học: CLC0332779_22_1_03CLC

Sinh viên thực hiện:

Võ Quang Hưng	20110057
Nguyễn Đình Duy	20110453
Nguyễn Minh Tuấn	20110594

Tp. Hồ Chí Minh, tháng 12 năm 2022

Mục lục

LỜI CẢM ƠN	1
I. Tìm hiểu về Textract :	2
1. Textract là gì ?	2
2. Textract có thể :	2
II. Các chức năng của Textract :	2
1. Extract text from image:	2
2. Form extraction :	2
3. Table extraction:	3
4. Query based extraction:	3
5. Handwriting recognition:	3
6. Invoices and receipts:	4
7. Identity documents:	4
III. Ưu điểm và nhược điểm của Textract:	4
1. Ưu điểm :	4
2. Nhược điểm :	4
IV. Demo chạy trên AWS	6
V. Demo chạy web	10
TÀI LIỆU THAM KHẢO :	19

LỜI CẢM ƠN

Để hoàn thành tốt đề tài và bài báo cáo này, chúng em xin gửi lời cảm ơn chân thành đến thầy Huỳnh Xuân Phụng, các quý thầy cô trong khoa Đào tạo Chất Lượng Cao nói chung và ngành Công Nghệ Thông Tin nói riêng đã tận tình truyền đạt những kiến thức cần thiết giúp em có nền tảng để làm nên đề tài này, đã tạo điều kiện để em có thể tìm hiểu và thực hiện tốt đề tài. Cùng với đó, em xin được gửi cảm ơn đến các bạn cùng khóa đã cung cấp nhiều thông tin và kiến thức hữu ích giúp em có thể hoàn thiện hơn đề tài của mình.

Đề tài và bài báo cáo được em thực hiện trong khoảng thời gian ngắn, với những kiến thức còn hạn chế cùng nhiều hạn chế khác về mặt kỹ thuật và kinh nghiệm trong việc thực hiện một dự án phần mềm. Do đó, trong quá trình làm nên đề tài có những thiếu sót là điều không thể tránh khỏi nên em rất mong nhận được những ý kiến đóng góp quý báu của các quý thầy cô để kiến thức của em được hoàn thiện hơn và em có thể làm tốt hơn nữa trong những lần sau. Em xin chân thành cảm ơn.

Cuối cùng, em kính chúc thầy luôn dồi dào sức khỏe và thành công hơn nữa trong sự nghiệp trồng người. Một lần nữa em xin chân thành cảm ơn.

TP.Hồ Chí Minh, ngày 01 tháng 12 năm 2022

I. Tìm hiểu về Textract :

1. Textract là gì ?

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.

2. Textract có thể :

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.

II. Các chức năng của Textract :

1. Extract text from image:

Trích xuất văn bản từ hình ảnh

2. 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.

3. 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.

4. 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.

5. 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.

6. 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.

7. 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.

III. Ưu điểm và nhược điểm của Textract:

1. Ưu điểm :

- **Thiết lập dễ dàng với Dịch vụ AWS:**
- **An toàn:** Amazon Textract tuân theo mô hình trách nhiệm chung AWS, bao gồm các quy định và hướng dẫn về bảo vệ dữ liệu. AWS chịu trách nhiệm bảo vệ cơ sở hạ tầng toàn cầu chạy tất cả các dịch vụ AWS; do đó chúng ta không cần lo lắng về việc dữ liệu của mình bị rò rỉ hoặc bị sử dụng bởi bất kỳ người nào khác.

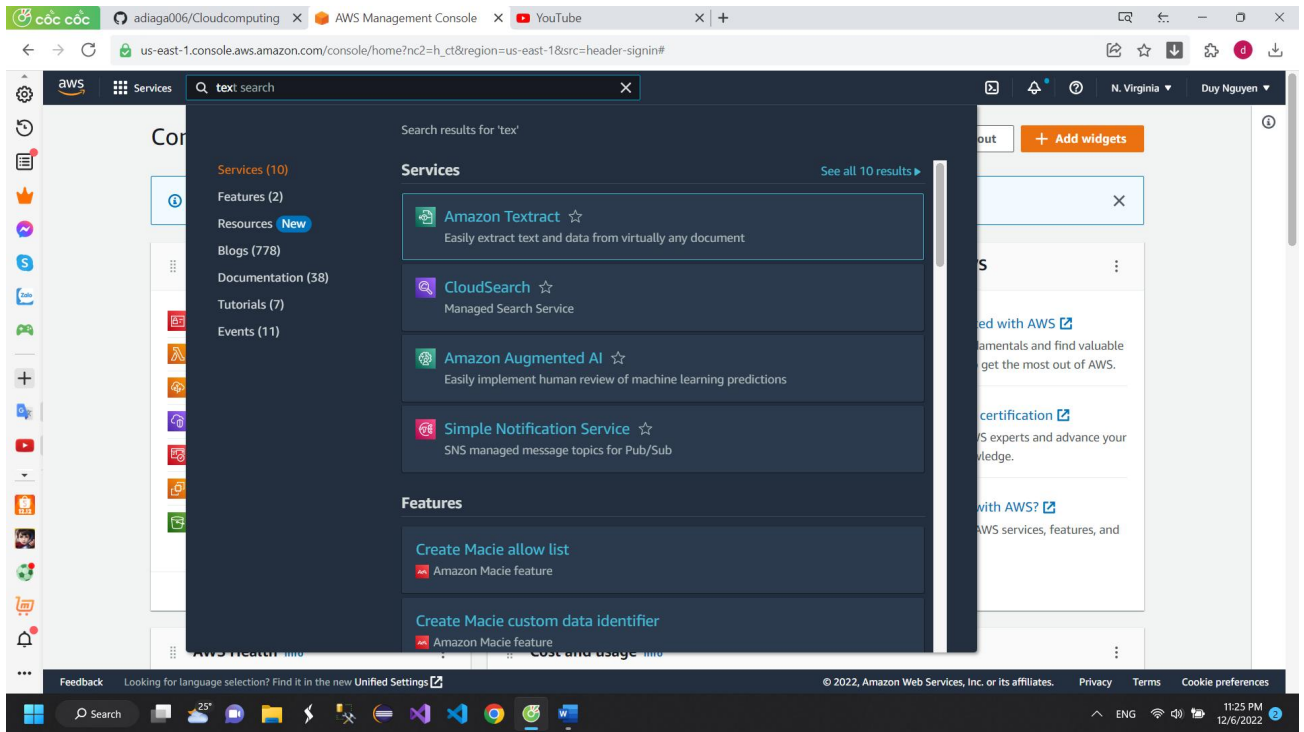
2. Nhược điểm :

- **Không có khả năng trích xuất các trường tùy chỉnh:** Có thể có nhiều trường dữ liệu trong một hóa đơn nhất định, chẳng hạn như ID hóa đơn, Ngày đến hạn, Ngày giao dịch, v.v. Những trường này là một cái gì đó phổ biến trong hầu hết các hóa đơn. Nhưng nếu chúng ta muốn trích xuất một trường tùy chỉnh từ hóa đơn, chẳng hạn như số GST hoặc thông tin ngân hàng, thì Textract thực hiện một cách kém hiệu quả.
- **Tích hợp với các nhà cung cấp khác:** Textract không cho phép bạn tích hợp với các nhà cung cấp khác nhau một cách dễ dàng
- **Khả năng xác định tiêu đề bảng:** Đối với các tác vụ trích xuất bảng, textract không cho phép bạn xác định tiêu đề bảng. Do đó, sẽ không dễ dàng để tìm kiếm hoặc tìm một cột hoặc một bảng cụ thể trong một tài liệu nhất định.
- **Kiểm tra Không gian lận:** Các OCR hiện đại giờ đây có thể tìm xem một tài liệu nhất định là tài liệu gốc hay giả bằng cách xác thực ngày tháng và tìm các vùng có pixel. AWS Textract không đi kèm với điều này, công việc duy nhất của nó là chọn tất cả văn bản từ một tài liệu đã tải lên.
- **Không có trích xuất văn bản dọc:** Trong một số tài liệu, số hóa đơn hoặc địa chỉ có thể được tìm thấy theo chiều dọc. Hiện tại, AWS chỉ hỗ trợ trích xuất văn bản theo chiều ngang
- **Giới hạn ngôn ngữ:** Amazon Textract hỗ trợ phát hiện văn bản tiếng Anh, tiếng Tây Ban Nha, tiếng Đức, tiếng Pháp, tiếng Ý và tiếng Bồ Đào Nha.
- **Đám mây cho mọi lưu trữ :** Mọi tài liệu được xử lý bằng Textract sẽ được đưa vào đám mây, chỉ hỗ trợ một vài vùng. Tuy nhiên, một số công ty có thể không quan tâm đến việc đưa tài liệu của họ lên đám mây vì những lý do như yêu cầu

bảo mật hoặc pháp lý. Tuy nhiên, rất tiếc, AWS Textract không hỗ trợ bất kỳ triển khai tại chỗ nào để xử lý tài liệu.

IV. Demo chạy trên AWS

Bước 1: Tìm Amazon Textract ở AWS



Bước 2: Chọn Analyze Document (phân tích tài liệu), AWS tự chọn tài liệu mẫu để nhận diện

The screenshot shows the Amazon Textract console interface. On the left is a navigation menu with options like 'Amazon Textract', 'Analyze Lending', 'Demos', 'Service quotas', and 'Additional resources'. The main area displays a sample document titled 'paystub' which has been analyzed. The document content is organized into several sections: 'Earnings Statement', 'Deductions', 'Other Benefits and Information', and 'Important Notes'. The 'Results' panel on the right shows the extracted data in a structured format, including fields like 'CO. FILE DEPT. CLOCK NUMBER', 'ANY COMPANY CORP.', 'Period ending: 7/18/2008', 'Pay date: 7/25/2008', 'Social Security Number: 987-65-4321', 'Taxable Marital Status: Married', 'JOHN STILES', 'Exemptions/Allowances: 101 MAIN STREET', 'Federal: 3. \$25 Additional Tax', 'ANYTOWN, USA 12345', 'State: 2', 'Local: 2', 'Earnings', 'rate', 'hours', 'this period', and 'year to date'.

Textract nhận diện theo Raw text (theo từng dòng) , Form (biểu mẫu) , Table (bảng) , Queries(câu hỏi truy vấn).

Có thể chọn các loại tài liệu mẫu khác như giấy vay , bảo hiểm y tế,, hoặc tự upload tài liệu từ máy.

This screenshot shows the same Amazon Textract console interface, but with the 'Choose a sample document' dropdown menu open. The menu lists several sample document types: 'Paystub', 'Vaccination card', 'Loan application', 'Health insurance card', and 'Form 1005'. The 'Upload document' section on the right provides instructions: 'Documents must be fewer than 11 pages, smaller than 5 MB, and one of the following formats: JPEG, PNG, or PDF.' The 'Results' panel on the far right shows a detailed view of the extracted data from a sample document, including fields like 'Local: 2', 'Earnings', 'rate', 'hours', 'this period', 'year to date', 'Other Benefits and Information', 'Regular', '10.00', '32.00', '320.00', '16,640.00', 'Information', 'this period', 'total to date', 'Overtime', '15.00', '1.00', '15.00', '780.00', 'Group Term Life', '0.51', '27.00', 'Holiday', '10.00', '8.00', '80.00', '4,160.00', 'Loan Amt Paid', '840.00', 'Tuition', '37.43*', '1,946.80', 'Gross Pay', '\$ 452.43', '23,526.80', 'Vac Hrs', '40.00', 'Sick Hrs', '16.00', 'Deductions', 'Statutory', 'Title', 'Operator', 'Federal Income Tax', '-40.60', '2,111.20', 'Social Security Tax', '-28.05', '1,458.60', 'Medicare Tax', '-6.56', '341.12', 'Important Notes', 'NY State Income Tax', '-8.43', '438.36', 'EFFECTIVE THIS PAY PERIOD YOUR REGULAR', 'NYC Income Tax', '-5.94', '308.88', 'HOURLY RATE HAS BEEN CHANGED FROM \$8.00', 'NY SUI/SDI Tax', '-0.60', '31.20', 'TO \$10.00 PER HOUR', '&', 'Other', 'Bond', '-5.00', '100.00'.

Bước 3 : Chọn Analyze Expense (phân tích biểu mẫu chi phí) , AWS tự chọn tài liệu mẫu để nhận diện

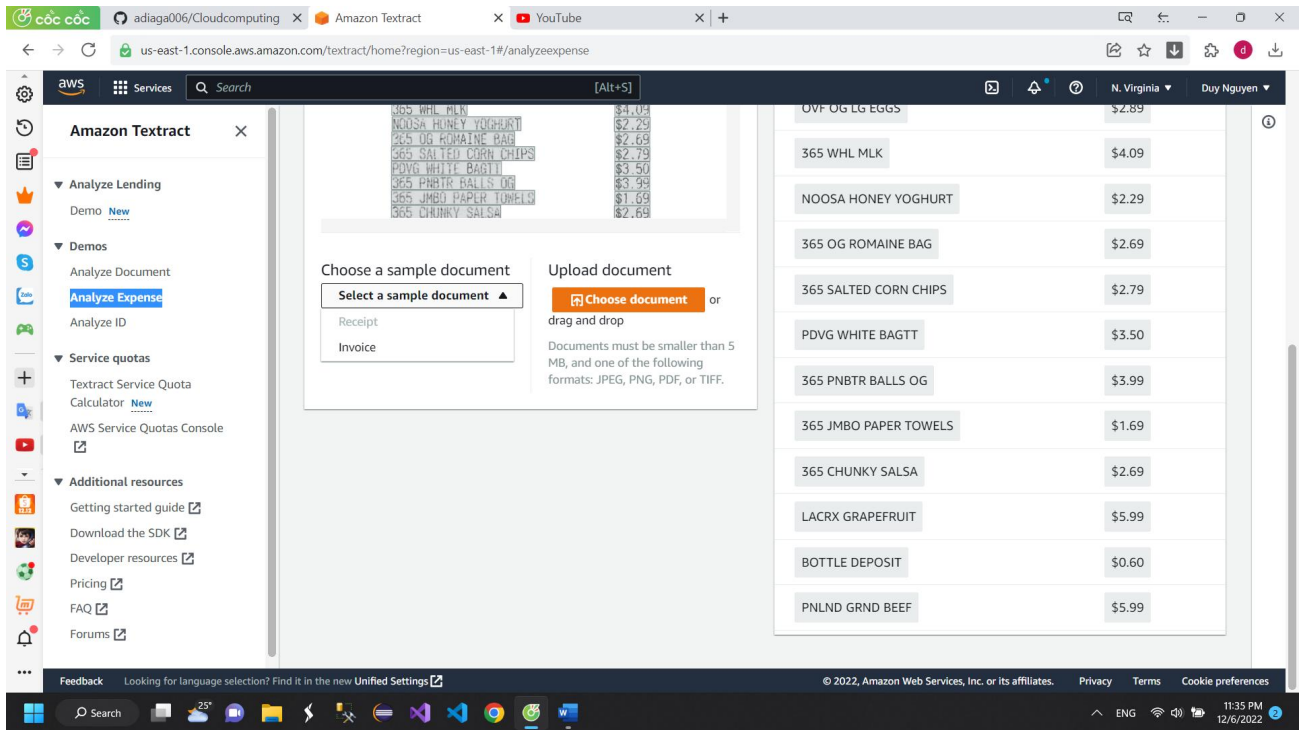
The screenshot shows the Amazon Textract console with the 'Analyze Expense' demo selected. The 'Sample receipt' is a Whole Foods Market receipt. The 'Summary fields' tab is active, showing the following extracted data:

Field	Value
(VENDOR_PHONE)	917-728-5700
(VENDOR_NAME)	WHOLE FOODS MARKET
(VENDOR_ADDRESS)	Bryant Park BPK 1095 6th Ave New York, NY 10036
(INVOICE_RECEIPT_DATE)	04/02/2019
(NAME)	WHOLE FOODS MARKET
(ADDRESS_BLOCK)	1095 6th Ave New York, NY 10036

Textract nhận diện theo Summary fields (theo dạng key- value) và Line item fields (theo dạng hàng – cột , có 2 cột là Item và Price)

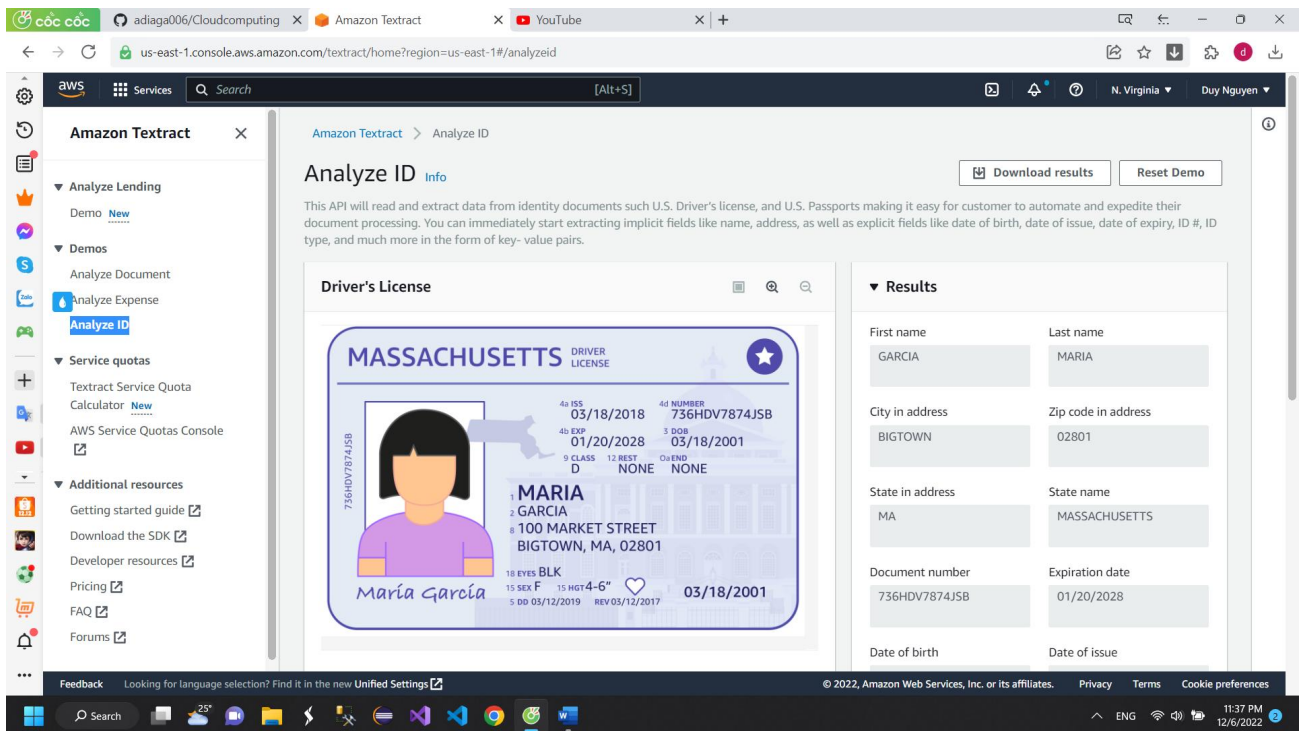
The screenshot shows the Amazon Textract console with the 'Analyze Expense' demo selected. The 'Line item fields' tab is active, showing the following extracted data:

ITEM	PRICE
BROU BROWN ALE	\$10.99
BOTTLE DEPOSIT	\$0.30
DRSCL STRAWBERRIES	\$3.49
OVF OG LG EGGS	\$2.89
365 WHL MLK	\$4.09
NOOSA HONEY YOGHURT	\$2.29
365 OG ROMAINE BAG	\$2.69



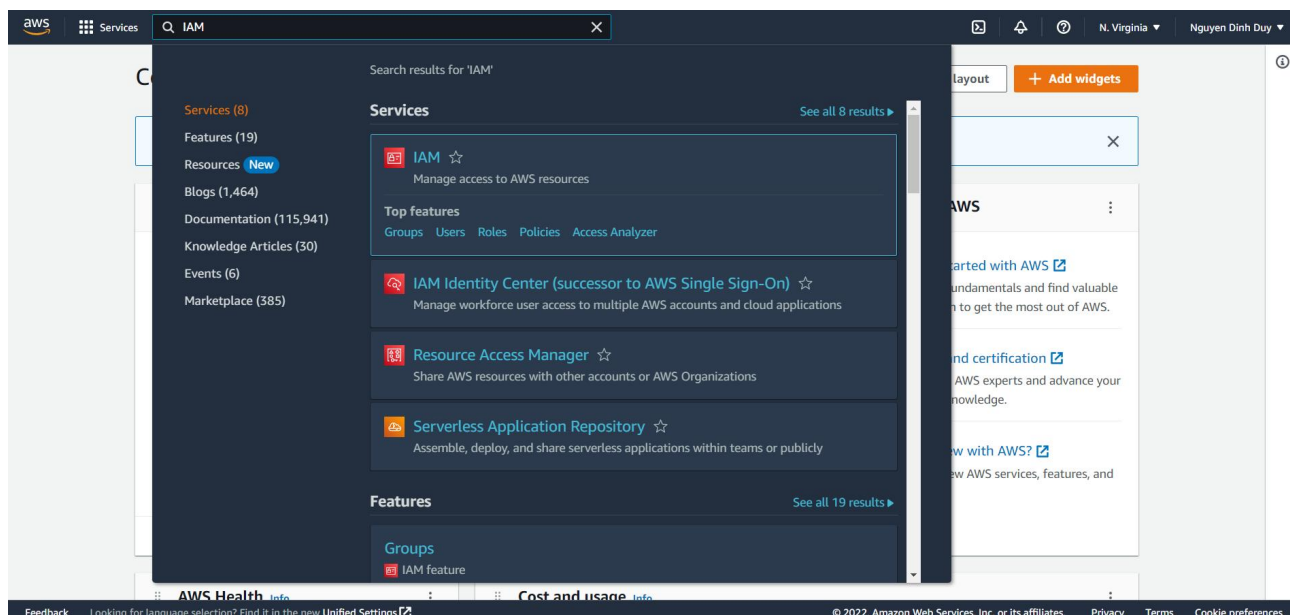
Có thể chọn các loại tài liệu mẫu khác như hóa đơn và biên nhận hoặc tự upload tài liệu từ máy.

Bước 4: Chọn Analyze ID , nhận diện chính xác Passport và giấy phép lái xe Hoa Kỳ

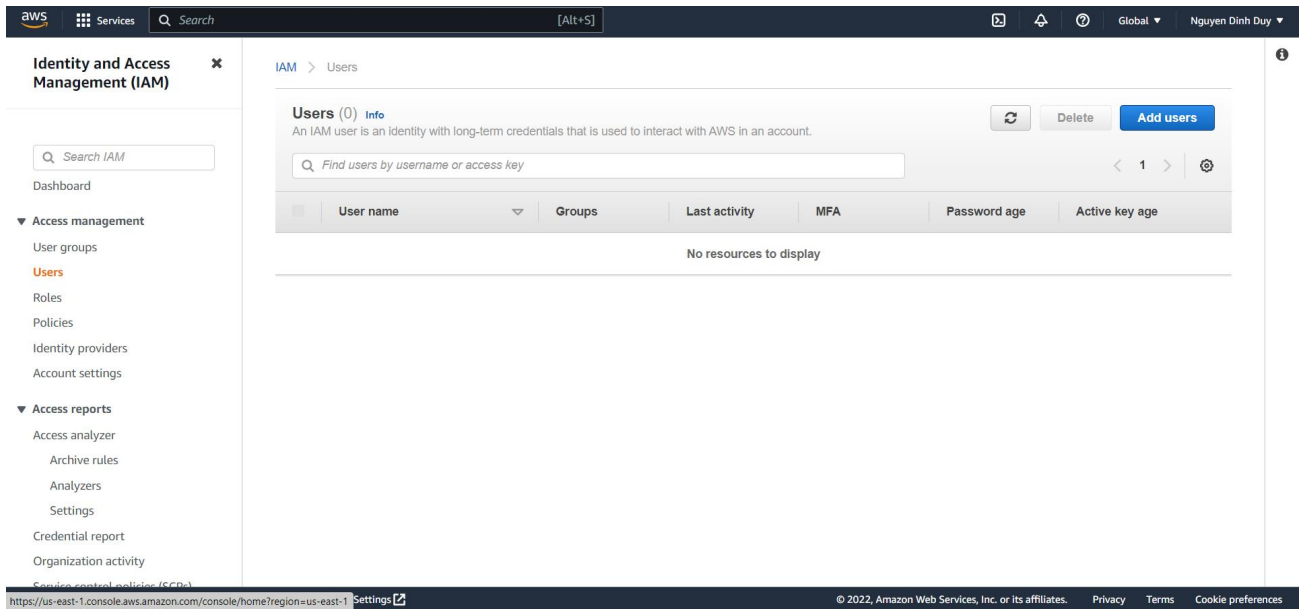


V. Demo chạy web

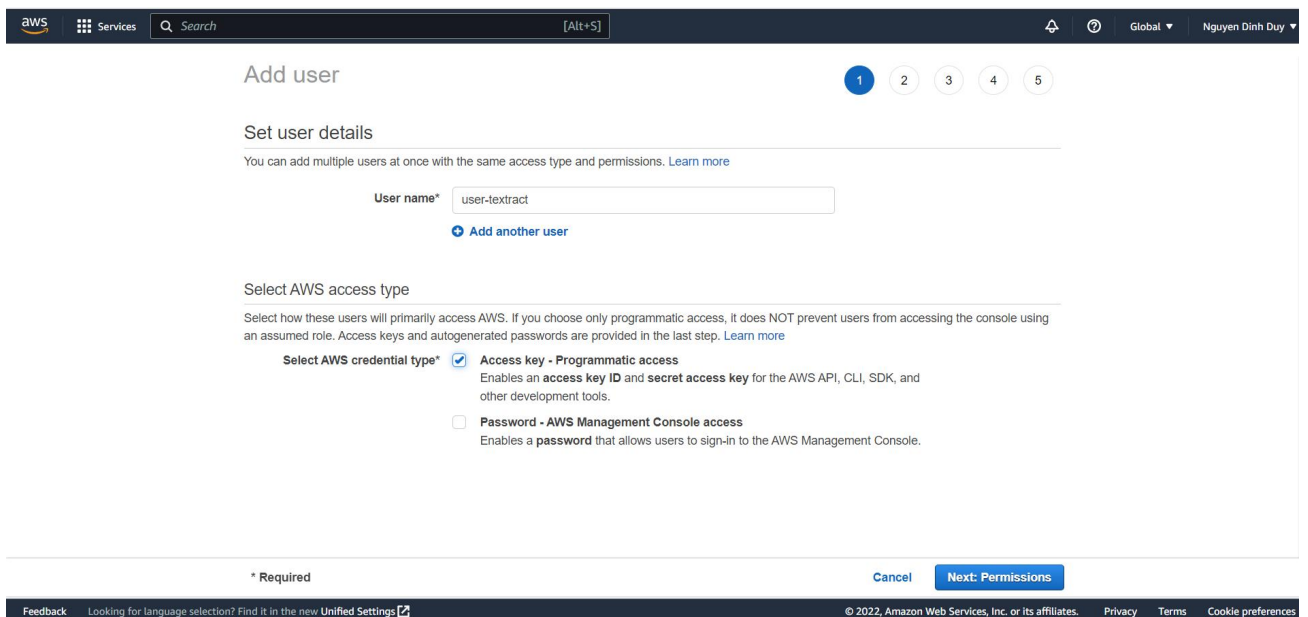
Bước 1: Tìm IAM ở AWS



Bước 2: Tạo user bằng các thực hiện chọn Users → Add users



Bước 3: Đặt tên user ở User name và click vào Access key - Programmatic access



Bước 4: Phân quyền cho user bằng cách chọn Attach existing policies directly → AdministratorAccess

aws

Services

Search

[Alt+S]

Global

Nguyen Dinh Duy

Add user

12345

Set permissions

Add user to group

Copy permissions from existing user









Attach existing policies directly

Create policy

Filter policies

Search

Showing 800 results

	Policy name	Type	Used as
<input checked="" type="checkbox"/>	 AdministratorAccess	Job function	None
<input type="checkbox"/>	 AdministratorAccess-Amplify	AWS managed	None
<input type="checkbox"/>	 AdministratorAccess-AWSElasticBeanstalk	AWS managed	None
<input type="checkbox"/>	 AlexaForBusinessDeviceSetup	AWS managed	None
<input type="checkbox"/>	 AlexaForBusinessFullAccess	AWS managed	None
<input type="checkbox"/>	 AlexaForBusinessGatewayExecution	AWS managed	None
<input type="checkbox"/>	 AlexaForBusinessLifesizeDelegatedAccessPolicy	AWS managed	None
<input type="checkbox"/>	 AlexaForBusinessOnlyDelegatedAccessPolicy	AWS managed	None

CancelPreviousNext: Tags

Feedback

Looking for language selection? Find it in the new Unified Settings

© 2022, Amazon Web Services, Inc. or its affiliates.

Privacy

Terms

Cookie preferences

aws

Services

Search

[Alt+S]

Global

Nguyen Dinh Duy

Add user

12345

Add tags (optional)

IAM tags are key-value pairs you can add to your user. Tags can include user information, such as an email address, or can be descriptive, such as a job title. You can use the tags to organize, track, or control access for this user. [Learn more](#)

Key	Value (optional)	Remove
<input type="text" value="Add new key"/>	<input type="text"/>	

You can add 50 more tags.

CancelPreviousNext: Review

Bước 5: Nhấn Create user để tạo user

12

aws

Services

Search

[Alt+S]

Global

Nguyen Dinh Duy

Add user

12345

Review

Review your choices. After you create the user, you can view and download the autogenerated password and access key.

User details

User name

user-texttract

AWS access type

Programmatic access - with an access key

Permissions boundary

Permissions boundary is not set

Permissions summary

The following policies will be attached to the user shown above.

Type	Name
Managed policy	AdministratorAccess

Tags

No tags were added.

Cancel

Previous

Create user

https://console.aws.amazon.com/console/home

Find it in the new Unified Settings

© 2022, Amazon Web Services, Inc. or its affiliates.

Privacy

Terms

Cookie preferences

Add user

12345

Success

You successfully created the users shown below. You can view and download user security credentials. You can also email users instructions for signing in to the AWS Management Console. This is the last time these credentials will be available to download. However, you can create new credentials at any time.

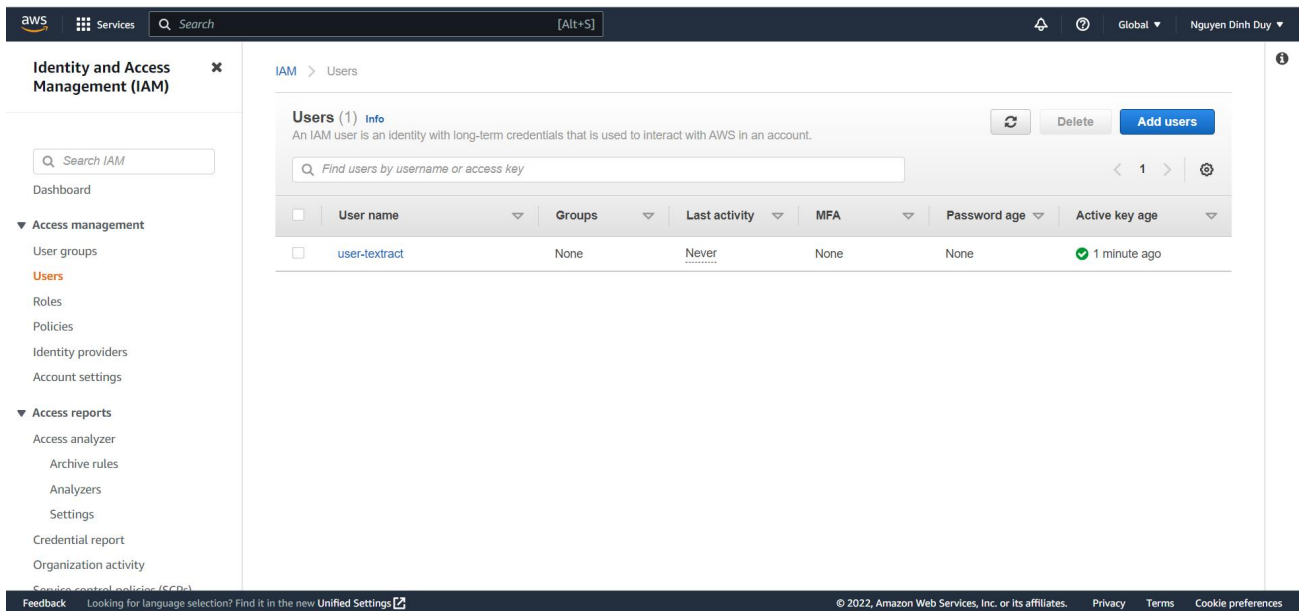
Users with AWS Management Console access can sign-in at: <https://246672979937.signin.aws.amazon.com/console>

Download .csv

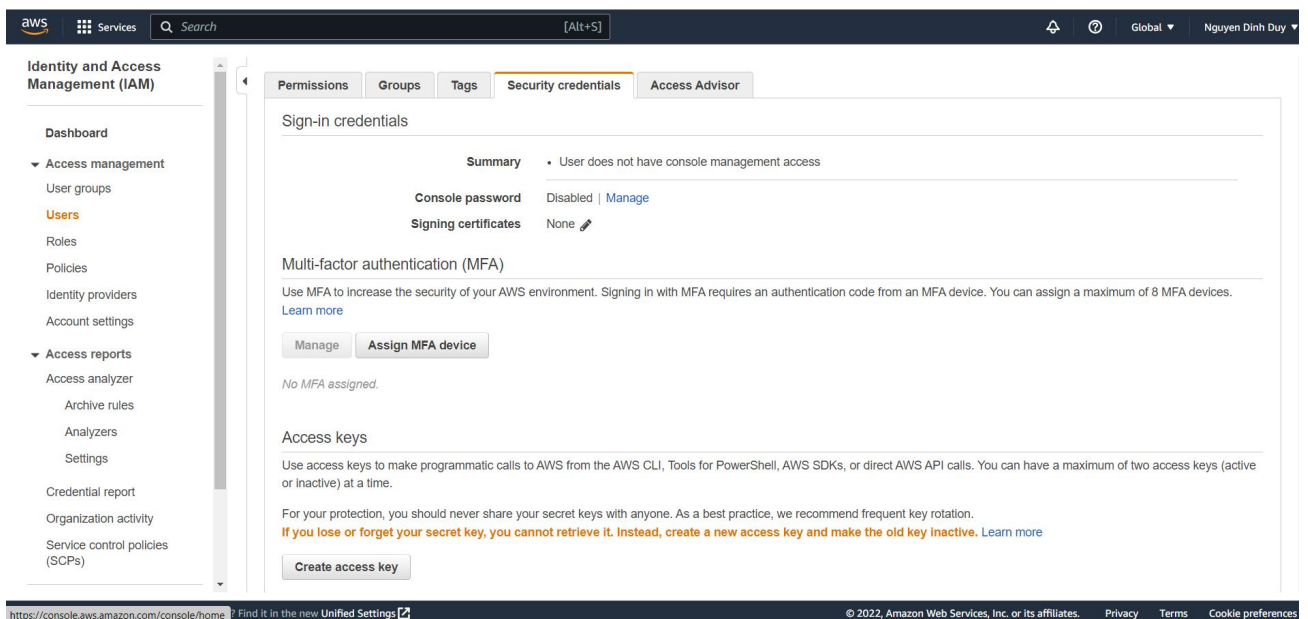
User	Access key ID	Secret access key
<div>user-texttract</div>	AKIATS3W267Q5VELTK7E	***** Show

Close

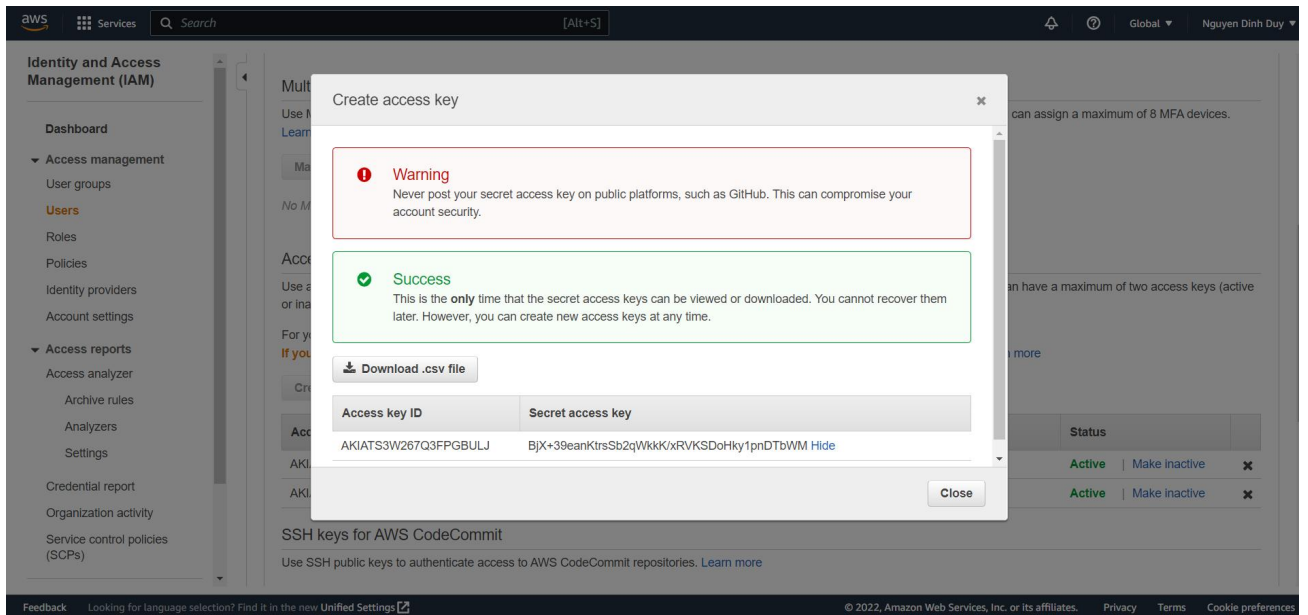
Bước 6: Click vào user vừa tạo



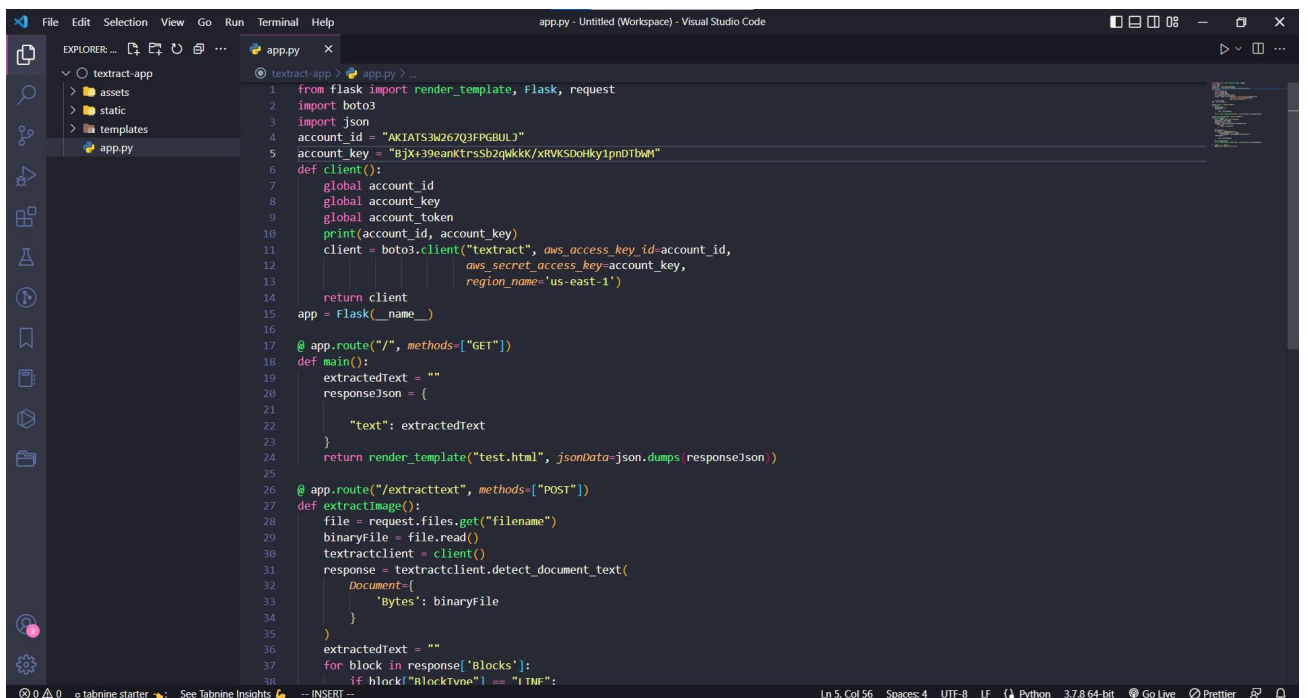
Bước 7: Chọn Security credential → Create access key



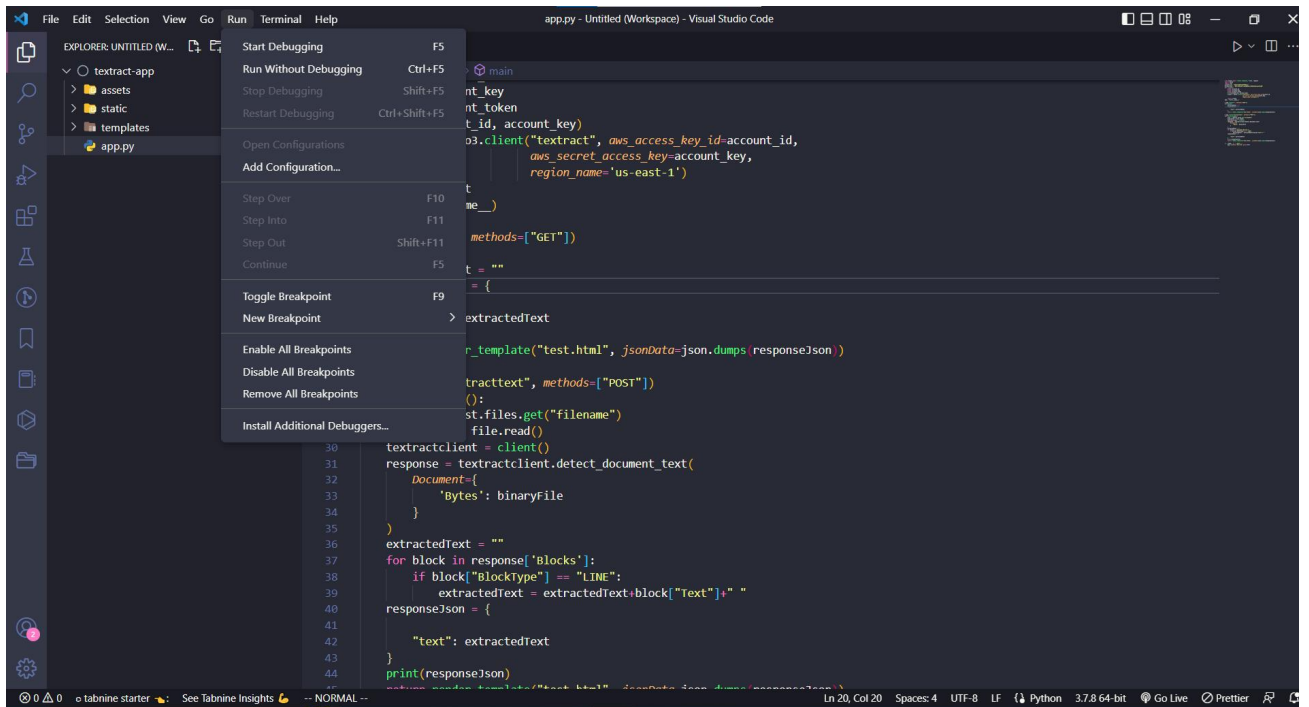
Bước 8: Copy Access key ID và Secret access key



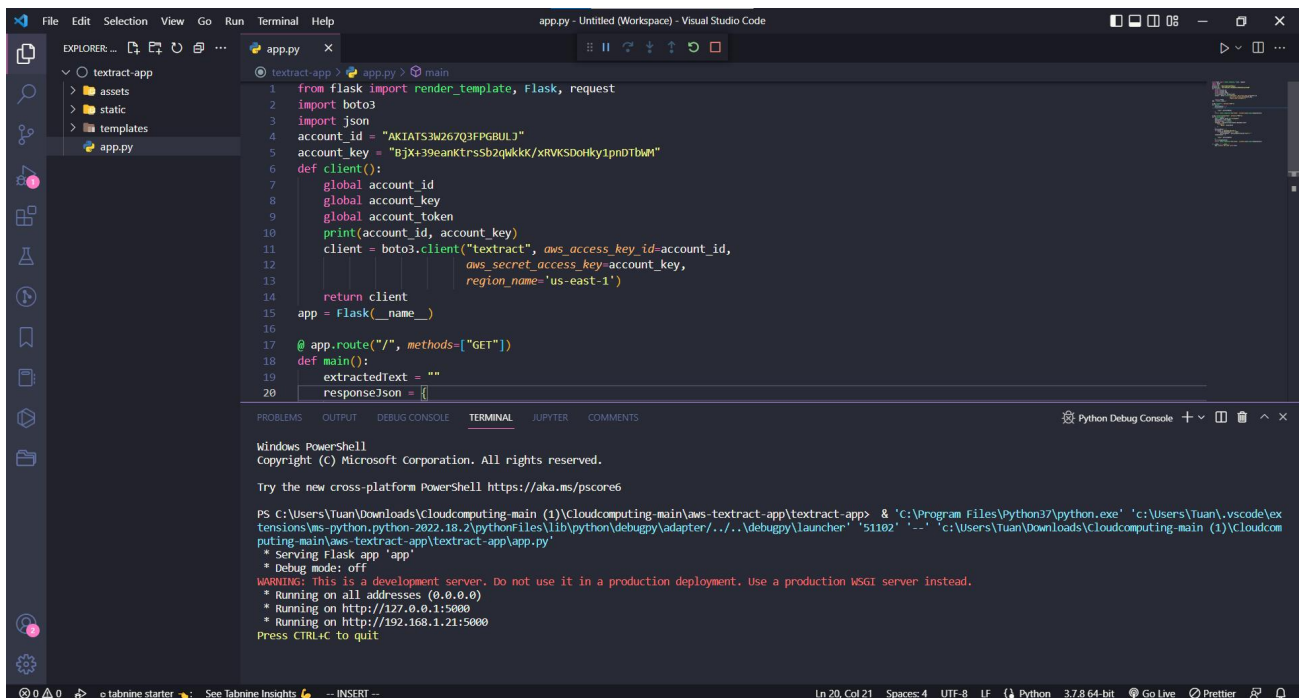
Bước 9: Dán Access key ID và Secret access key lần lượt vào `account_id` và `account_key` ở file `app.py`



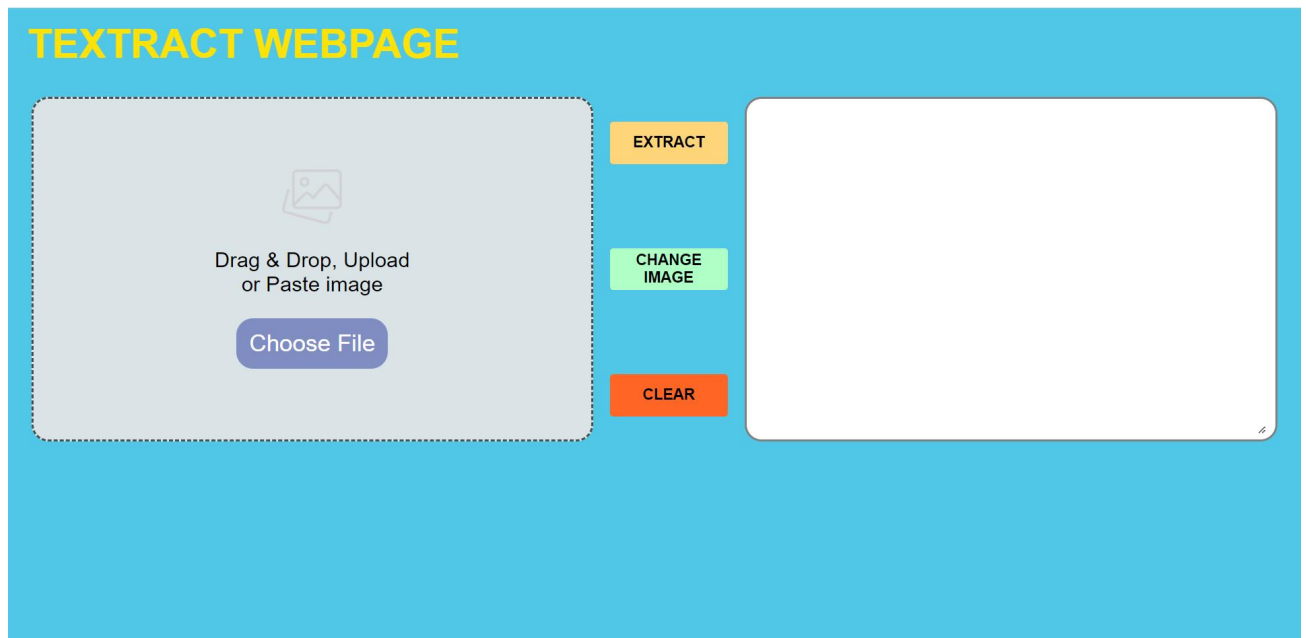
Bước 10: Khởi chạy trang web bằng cách chọn Run → Run Without Debugging



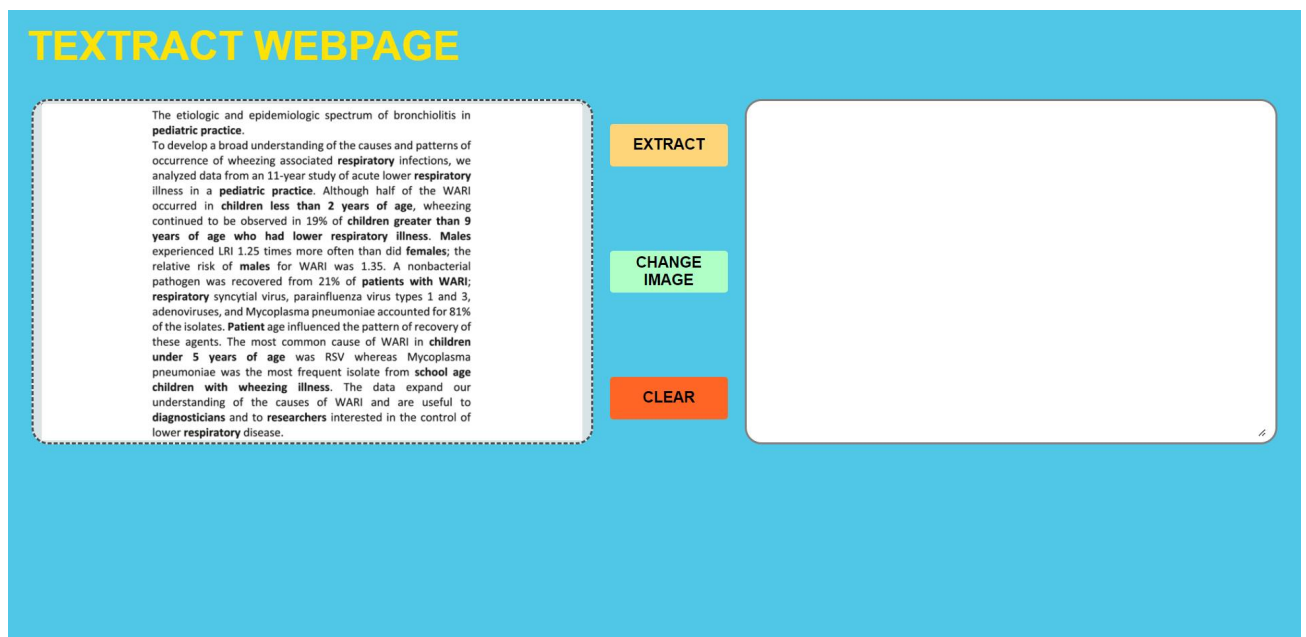
Bước 11: Ctrl + Click vào URL



Bước 12: Nhấn Choose File chọn file văn bản jpg



Bước 13: Nhấn Extract tiến hành trích xuất dữ liệu của văn bản



Kết quả:

TEXTTRACT WEBPAGE



Drag & Drop, Upload
or Paste image

Choose File

EXTRACT

CHANGE
IMAGE

CLEAR

The etiologic and epidemiologic spectrum of bronchiolitis in pediatric practice. To develop a broad understanding of the causes and patterns of occurrence of wheezing associated respiratory infections, we analyzed data from an 11-year study of acute lower respiratory illness in a pediatric practice. Although half of the WARI occurred in children less than 2 years of age, wheezing continued to be observed in 19% of children greater than 9 years of age who had lower respiratory illness. Males experienced LRI 1.25 times more often than did females; the relative risk of males for WARI was 1.35. A nonbacterial pathogen was recovered from 21% of patients with WARI; respiratory syncytial virus, parainfluenza virus types 1 and 3, adenoviruses, and Mycoplasma pneumoniae accounted for 81% of the isolates. Patient age influenced the pattern of recovery of these agents. The most common cause of WARI in children under 5 years of age was RSV whereas Mycoplasma pneumoniae was the most frequent isolate from school age children with wheezing illness. The data expand our understanding of the causes of WARI and are useful to diagnosticians and to researchers interested in the control of lower respiratory disease.

4

TÀI LIỆU THAM KHẢO :

<https://docs.aws.amazon.com/textract/latest/dg/what-is.html>

<https://aws.amazon.com/vi/textract/features/>

<https://docs.aws.amazon.com/textract/latest/dg/how-it-works.html>

<https://github.com/aniketwattamwar/AWS-Flask/tree/master/textract>

<https://dev.to/aws-builders/amazon-textract-with-expense-analyzing-516b>