Estrazione di dati da documenti strutturati con Amazon Textract, AWS Lambda e Amazon S3

7 Luglio 2023 - 9 min. read

Amazon S3 *Amazon Textract* *AWS Lambda*

Nell'era digitale, elaborare e gestire efficacemente grandi quantità di documenti è una priorità per le aziende di ogni settore. Molte organizzazioni si trovano ad affrontare il compito di digitalizzare grandi volumi di documenti cartacei o di elaborare dati provenienti da documenti strutturati, come fatture o contratti, in modo automatico. In questo contesto, l'Optical Character Recognition (OCR) si è rivelato uno strumento indispensabile per automatizzare i processi e migliorare l'efficienza complessiva.

Tuttavia, riuscire ad estrarre il testo da un documento è solo parte di quello di cui la maggior parte delle applicazioni hanno bisogno. Se vogliamo la si può considerare una funzione primitiva. Spesso l'obiettivo è di estrarre specifiche informazioni, selezionando il testo in base alla struttura del documento.

Per selezionare correttamente le informazioni di valore, diventa quindi importante ottenere informazioni sulla struttura del documento, come ad esempio su come il testo è raggruppato, intabellato o sulla posizione occupata all'interno della pagina.

Trovare risposta a queste domande è esattamente l'area in cui **Amazon Textract** si distingue.

Oltre a fornire la capacità di estrarre testo da documenti, Amazon Textract è in grado di **identificare e restituire informazioni sulla struttura della pagina**, aprendo la strada a una vasta gamma di possibilità di elaborazione dei dati.

A differenza dei tradizionali software OCR, che richiedono configurazioni manuali e aggiornamenti continui per adattarsi ai cambiamenti dei moduli, Amazon Textract utilizza **modelli di machine learning** per elaborare qualsiasi tipo di documento, garantendo un'estrazione accurata di testo, scrittura a mano, tabelle e altri dati senza alcun intervento manuale.

Senza ulteriori preamboli, passiamo quindi alla descrizione di uno use case.

Estrazione delle informazioni da una fattura

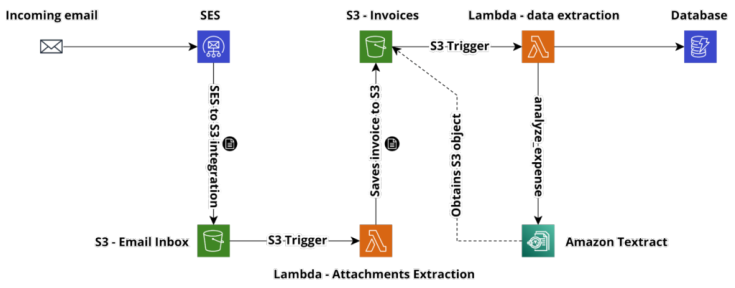
Per esplorare le potenzialità di Amazon Textract ci avvarremo di un caso (neanche troppo) ipotetico, in cui la necessità sia quella di estrarre in maniera automatica alcune informazioni dalle fatture degli acquisti aziendali, in modo da inserire gli importi e la data in un database che viene periodicamente importato nel software gestionale.

Dobbiamo quindi costruire un **sistema automatico** in grado di estrarre l'importo e la data dalle fatture che riceve. Per semplicità, poniamo che le fatture abbiano tutte la medesima struttura perchè provengono dal sito del fornitore da cui la nostra azienda si rifornisce di beni di consumo, anche se Textract può tranquillamente analizzare fatture eterogenee.

Le fatture sono documenti PDF pensati per essere letti da un umano. Contengono intestazioni, l'immagine del logo del fornitore, testo e tabelle in diverse posizioni della pagina.

Il sito invia tramite email una fattura per ogni ordine. Nel nostro scenario, l'indirizzo fa riferimento ad un gruppo mail, pertanto possiamo fare in modo che il sistema automatico ne riceva una copia senza intaccare i processi che coinvolgono i nostri operatori.

In questa situazione potremmo abbozzare la seguente soluzione ad alto livello



Sottoscrivedo un indirizzo email ad-hoc al gruppo email possiamo indirizzare una copia delle email contenenti le fatture al nostro sistema.

Per ricevere le email e processarle possiamo sfuttare Amazon SES; sebbene sia comunemente noto per l'invio delle email in ambiente AWS, può anche essere utilizzato per la ricezione e la conseguente integrazione con i servizi AWS utili all'elaborazione delle email.

La ricezione di email supporta diverse integrazioni, ma per il nostro caso quella più pratica è certamente quella con Amazon S3.

Mediante l'integrazione, Amazon SES salva un oggetto contenente i dati grezzi della mail, in formato MIME, sul bucket S3 designato. Questo ci consente quindi di mantenere lato AWS uno storico dei body raw di ogni messaggio ricevuto, utile sia per archiviazione sia per indagare eventuali malfunzionamenti.

L'utilizzo dello spazio di storage comporta costi minimi e, in caso di grandissime quantità di messaggi ricevuti, è possibile ottimizzare il billing sfruttando tutte le funzioni di Amazon S3 come ad esempio le **Lifecycle policy** e le **classi di storage** a basso costo.

A questo punto un S3 trigger entra in funzione, avviando una Lambda Function che si occupa di fare il parsing del corpo della mail ed estrarre l'allegato. Il file può essere quindi salvato in un bucket S3 dedicato.

In questo articolo non esploreremo il codice necessario all'estrazione degli allegati perchè non è il nocciolo di quanto intendiamo trattare. Tuttavia, esistono librerie per la maggior parte dei linguaggi più diffusi che si occupano di semplifare il parsing e la manipolazione di dati in formato MIME. Per esempio, questa è una libreria da cui partire per sviluppare la funzione usando NodeJs.

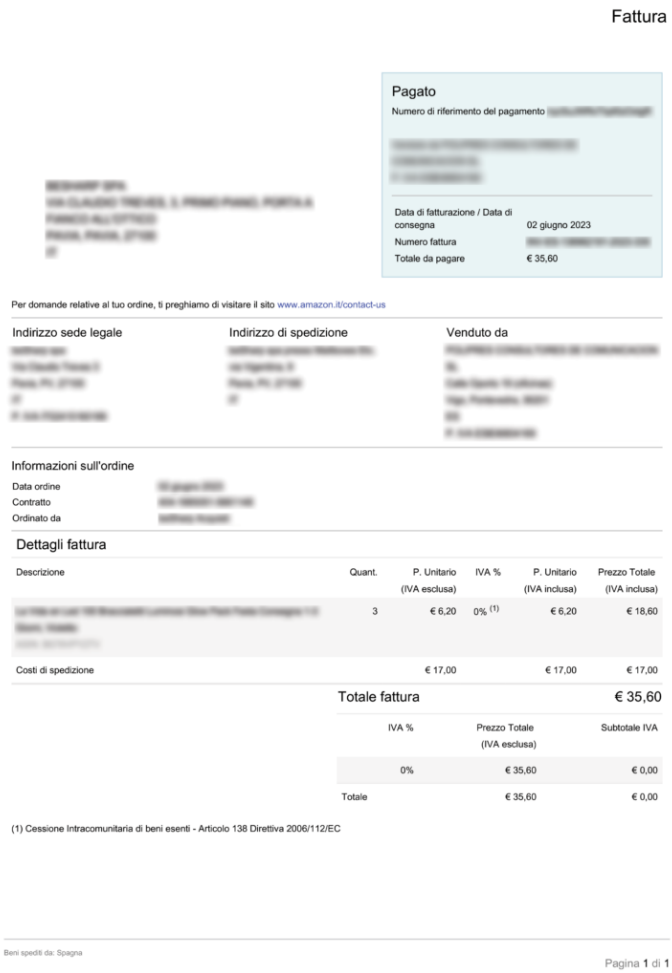
A questo punto, se la fattura è salvata in uno dei formati supportati da Amazon Textract, è possibile procedere all'estrazione delle informazioni. In caso contrario, è possibile aggiungere una ulteriore funzione, oppure estendere quella che manipola la mail, per effettuare una conversione verso un formato universale, come ad esempio PDF.

Un secondo trigger S3 avvia una Lambda function adibita ad invocare Amazon Textract passandogli in input l'oggetto di cui effettuare l'analisi. La lambda andrà poi a navigare il risultato dell'analisi per prelevare importo e data di emissione della fattura, e a salvare le informazioni interessate all'interno del database.

L'integrazione con Amazon Textract è piuttosto assistita ed esistono i metodi in tutti gli SDK di AWS per i principali linguaggi di programmazione, come boto3 per Python e l'AWS SDK per JavaScript.

La fattura di riferimento

Nel nostro scenario, useremo delle fatture generate da Amazon Business, che sono fatte più o meno così:



Fattura

Pagato

Numero di riferimento del pagamento

Data di fatturazione / Data di consegna 02 giugno 2023

Numero fattura

Totale da pagare € 35,60

Per domande relative al tuo ordine, ti preghiamo di visitare il sito www.amazon.it/contact-us

Indirizzo sede legale | Indirizzo di spedizione | Venduto da

Informazioni sull'ordine

Data ordine
Contratto
Ordinato da

Dettagli fattura

Descrizione	Quant.	P. Unitario (IVA esclusa)	IVA %	P. Unitario (IVA inclusa)	Prezzo Totale (IVA inclusa)
	3	€ 6,20	0% (1)	€ 6,20	€ 18,60
Costi di spedizione		€ 17,00		€ 17,00	€ 17,00

Totale fattura € 35,60

IVA %	Prezzo Totale (IVA esclusa)	Subtotale IVA
0%	€ 35,60	€ 0,00
Totale	€ 35,60	€ 0,00

(1) Cessione Intracomunitaria di beni esenti - Articolo 138 Direttiva 2006/112/EC

Beni spediti da: Spagna

Pagina 1 di 1

Il documento presenta il totale sia fuori dalla tabella con il dettaglio, sia in fondo alla tabella.

Si tratta di un elemento da tenere in considerazione, perchè potremo selezionare il totale più facile da individuare con Textract, o costruire logiche per cercare il totale sia in tabella che fuori, e selezionare quello presente o se presenti entrambi e di diverso valore, quello con la confidenza maggiore.

L’API di Textract

Il servizio mette a disposizione diverse chiamate API, sia di tipo sincrono che di tipo asincrono.

Le versioni sincrone permettono di inviare un documento per un'analisi immediata e la risposta alla chiamata API è il risultato stesso dell’analisi. Queste chiamate API presentano importanti limitazioni sui formati accettati, e sul fatto che il chiamante deve necessariamente attendere il termine dell’analisi, il che può richiedere diversi secondi.

Le versioni asincrone, invece, restituiscono immediatamente un JOBID mediante il quale è possibile richiedere il risultato dell’analisi in un secondo momento. Esiste anche un meccanismo per ottenere notifica dell’avvenuta analisi mediante SNS; in sintesi è possibile specificare il topic SNS quando si avvia l’analisi asincrona.

Occorre considerare accuratamente l’utilizzo della versione asincrona, soprattutto se la computazione è basata su Lambda. Questo consente di disaccoppiare meglio le componenti infrastrutturali, di costruire meccanismi di retry robusti, aumentando l’alta affidabilità complessiva della soluzione e riducendo i costi eliminando il pagamento del tempo di computazione Lambda occupato dall’attesa sincrona del risultato dell’analisi.

Al momento in cui sto scrivendo questo articolo, le chiamate API disponibili sono le seguenti:

- analyze-document
- analyze-expense
- analyze-id
- detect-document-text
- get-document-analysis
- get-document-text-detection
- get-expense-analysis
- get-lending-analysis
- get-lending-analysis-summary
- start-document-analysis

- start-document-text-detection
- start-expense-analysis
- start-lending-analysis

Per il nostro caso possiamo sfruttare la funzione **analyze-expenses**, che utilizza un modello appositamente addestrato per riconoscere ed estrarre dati finanziari da documenti come ad esempio le fatture.

Navigare l'esito delle analisi di Textract

La risposta dell'analisi di Amazon Textract può essere estremamente verbosa e contiene tantissime informazioni sulle parti di documento scoperte.

Come è possibile vedere dall'immagine precedente, una fattura media può produrre una risposta di oltre 34.000 righe.

La risposta della funzione **analyze-expenses** si compone di 3 macroblocchi:

1. SummaryFields
2. LineItemGroups
3. Blocks

Nella sezione **SummaryFields** vengono raccolte ed etichettate tutte le informazioni che Amazon Textract trova al di fuori di una tabella.

Nella sezione **LineItemGroups** vengono invece raccolti, separati per riga, tutte le informazioni presenti all'interno delle tabelle.

In **Block** è possibile trovare tutti i blocchi di testo trovati da textract, con informazioni sui bounding box e la posizione all'interno della pagina. Questa sezione è quella che contiene i dati di più basso livello, ed è la sezione che sarebbe stata restituita eseguendo la generica funzione di analyze-document.

Per le fatture che stiamo prendendo in esame, useremo i dati estrapolati da SummaryFields, dove possiamo trovare già etichettati per tipo i riferimenti agli importi ed alle date.

Questo è un frammento del primo elemento del vettore SummaryFields, la chiamata è relativa all'analisi di una fattura Amazon Business.


```
{
    "ExpenseIndex": 1,
    "SummaryFields": [
        {
            "Type": {
                "Text": "ADDRESS",
                "Confidence": 59.766048431396484
            },
            "LabelDetection": {
                "Text": "Venduto da",
                "Geometry": {
                    "BoundingBox": {
                        "Width": 0.08296574652194977,
                        "Height": 0.008674301207065582,
                        "Left": 0.6447910070419312,
                        "Top": 0.3040856122970581
                    },
                    "Polygon": [
                        {
                            "X": 0.6447910070419312,
                            "Y": 0.30408763885498047
                        } … ,
```

```
                    ]
                },
                "Confidence": 95.342376708984375
            },
            "ValueDetection": {
                "Text": "Amazon EU S. r.l. Succursale Italiana\nViale
Monte Grappa 3/5\n20124 Milano\nItalia",
                "Geometry": {
                    "BoundingBox": {
                        "Width": 0.21900421380996704,
                        "Height": 0.055659566074609756,
                        "Left": 0.6449917554855347,
                        "Top": 0.3213878870010376
                    },
                    "Polygon": [
                        {
                            "X": 0.6449917554855347,
                            "Y": 0.32139283418655396
                        } …,                    ]
                },
                "Confidence": 56.650238037109375
            },
            "PageNumber": 1,
            "GroupProperties": [
                {
                    "Types": [
                        "RECEIVER"
                    ],
                    "Id": "42d41651-ba4a-4fc5-97ce-f9b34d1d987d"
                }
            ]
        }
```


Alcuni dettagli sulle geometrie sono stati tagliati dal JSON per brevità e perchè irrilevanti al fine della nostra trattazione.

Come è possibile notare, è disponibile una grande quantità di informazioni sul frammento di testo estratto. Per esempio, nel campo Type è possibile conoscere che il frammento catturato è testo, e che contiene un indirizzo.

Poco sotto, l'oggetto LabelDetection ci dice che la label associata all'indirizzo è "Venduto da". Continuando a scorrere saltando tutte le informazioni sulla posizione e sui bounding box possiamo arrivare alla sezione ValueDetection, che contiene il valore corretto dell'indirizzo dell'esercente.

Per trovare le informazioni che ci servono non dobbiamo fare altro che scorrere il vettore di oggetti SummaryFields e cercare nel campo Type la stringa "TOTAL".

Textract troverà più totali, almeno sul documento di esempio, perchè la cifra è riportata più volte e ci sono totali parziali. Nel caso specifico, il campo più sicuro da estrarre è quello individuato con il tipo TOTAL, che ha anche la confidenza maggiore. Ulteriore riprova si può avere analizzando il campo LabelDetection, che conterrà il testo a descrizione del campo, nel nostro caso "Totale fattura"


```
"Type": {
        "Text": "TOTAL",
        "Confidence": 96.169677734375
    },
    "LabelDetection": {
        "Text": "Totale fattura",
        "Geometry": {
            "BoundingBox": {
                "Width": 0.10979866981506348,
                "Height": 0.010328351520001888,
                "Left": 0.49856579303741455,
                "Top": 0.6540638208389282
            },
            "Polygon": [
                {
                    "X": 0.49856579303741455,
                    "Y": 0.6540638208389282
                },...,
            ]
```

```
        },
        "Confidence": 95.28824615478516
    },
    "ValueDetection": {
        "Text": "64,99",
        "Geometry": {
            "BoundingBox": {
                "Width": 0.06178450211882591,
                "Height": 0.011797532439231873,
                "Left": 0.8738263249397278,
                "Top": 0.6542784571647644
            },
            "Polygon": [
                {
                    "X": 0.8738263249397278,
                    "Y": 0.6542784571647644
                },...,
            ]
        },
        "Confidence": 64.54907989501953
    },
    "PageNumber": 1,
    "Currency": {
        "Code": "EUR"
    }
}
```


Il valore individuato nella sezione ValueDetection è il totale corretto di 64,99 Euro.

Allo stesso modo possiamo trovare anche le date di emissione e, se disponibili, quelle di scadenza.

Alcuni tipi utili che Amazon Textract è in grado di identificare sono:

- TOTAL
- AMOUNT_DUE

- VENDOR_ADDRESS
- VENDOR_NAME
- RECEIVER_NAME
- SUBTOTAL
- OTHER

Utilizzando OTHER e il testo individuato da textract come label è possibile estrarre campi personalizzati.

Certamente, utilizzare il campo Type per trovare i valori notevoli rende la soluzione in grado di interpretare fatture con struttura eterogenea e sfrutta a pieno il modello di ML specializzato che Amazon ha realizzato. Tuttavia, in caso il labeling di Amazon Textract non fosse sufficientemente preciso per gli scopi specifici, è possibile ignorare il campo type e procedere scorrendo le coppie Label value individuate, cercando in label delle parole chiave per identificare i valori di interesse.

È sufficiente una singola chiamata API per ottenere tutti i dati presenti nel documento.

La parte più tediosa è fare qualche prova per capire quali informazioni vengono correttamente classificate e per quali invece occorre implementare un riconoscimento basato sulle label o su una combinazione di label e tipo riconosciuto.

Occorre ovviamente implementare una solida logica di **fallback**, in caso in cui non si trovino tutte le informazioni o la confidenza risulti sotto una soglia definibile per marcare il documento per un'analisi manuale.

Conclusioni

In questo articolo abbiamo visto come Amazon Textract possa semplificare il processo di estrazione dei dati rilevanti da un documento di fattura, senza richiedere una configurazione manuale o una conoscenza approfondita del formato del documento.

Abbiamo anche illustrato le principali caratteristiche dell'API di Textract, che permette di analizzare le fatture e i documenti finanziari in modo rapido ed accurato, restituendo i dati in un formato strutturato e facilmente utilizzabile.

Questa soluzione può essere applicata a diversi scenari, non solo a quello ipotetico presentato, ed è possibile personalizzare le integrazioni per gestire scenari variegati.

Qual è secondo voi l'impiego perfetto per Amazon Textract? Fatecelo sapere!

A presto con un nuovo articolo dal mondo AI/ML su AWS!

About Proud2beCloud

Proud2beCloud è il blog di beSharp, APN Premier Consulting Partner italiano esperto nella progettazione, implementazione e gestione di infrastrutture Cloud complesse e servizi AWS avanzati. Prima di essere scrittori, siamo Solutions Architect che, dal 2007, lavorano quotidianamente con i servizi AWS. Siamo innovatori alla costante ricerca della soluzione più all'avanguardia per noi e per i nostri clienti. Su Proud2beCloud condividiamo regolarmente i nostri migliori spunti con chi come noi, per lavoro o per passione, lavora con il Cloud di AWS. Partecipa alla discussione!

Alessio Gandini

Cloud-native Development Line Manager @ beSharp, DevOps Engineer e AWS expert.Computer geek da quando avevo 6 anni, appassionato di informatica ed elettronica a tutto tondo. Ultimamente sto esplorando l'esperienza utente vocale e il mondo dell'IoT.Appassionato di cinema e grande consumatore di serie TV, videogiocatore della domenica.
