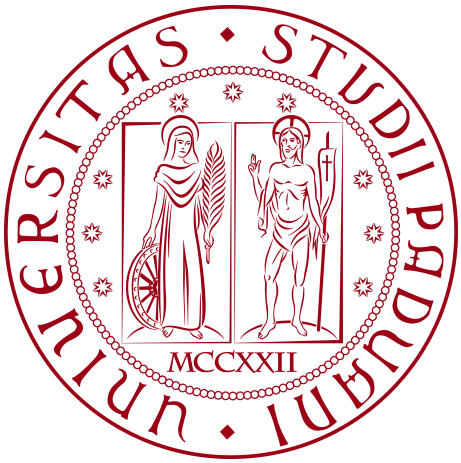
Università degli Studi di Padova

Dipartimento di Matematica "Tullio Levi-Civita"

Corso di Laurea in Informatica



Utilizzo dei Modelli di Machine Learning di AWS per la Classificazione e l'Estrapolazione di Informazioni contenute nelle Mail PEC

Tesi di Laurea Triennale

Relatore

Prof. Lamberto Ballan

Laureando

Riccardo Zaupa

Matricola 2034303

Anno Accademico 2023 - 2024

"I'm stronger, I'm smarter, I'm better"

— Homelander.

"Diventerò il re dei pirati"

— Monkey D. Luffy.

"Non devo fuggire"

— Shinji Ikari.

"Non dire gatto se non ce l'hai nel sacco"

— Nonna.

"GG\MM\AAAAAAA"

— Alex Scantamburlo.

"Ucciderò tutti i giganti"

— Eren Yeager.

"Con il superamento della revisione PB – a fronte di una prestazione estremamente deludente – avete concluso il vostro progetto didattico di IS."

— Tullio Vardanega.

"Non posso aiutarvi"

— Alessandro Staffolani.

"Mi avete fatto perdere mesi della mia vita"

— Riccardo Cardin.

Ringraziamenti

Desidero esprimere la mia gratitudine al professor Lamberto Ballan, mio relatore, per l'aiuto e il sostegno che mi ha dato durante la stesura dell'elaborato. Vorrei anche ringraziare, con affetto, i miei genitori per il loro sostegno, il grande aiuto e la loro presenza in ogni momento durante gli anni di studio. Desidero poi ringraziare i miei amici per i bellissimi anni trascorsi insieme e le mille avventure vissute.

Padova, Settembre 2024 Riccardo Zaupa

Sommario

Il presente documento descrive il lavoro svolto durante il periodo di stage del laureando Riccardo Zaupa presso l'azienda Sanmarco Informatica S.p.A. . Tale periodo, svolto alla conclusione del percorso di studi triennale in Informatica presso l'Università degli Studi di Padova, ha avuto una durata complessiva di trecentoventiore.

Gli obiettivi principali del progetto hanno riguardato l'analisi e l'utilizzo dei servizi AWS per l'addestramento di modelli di Intelligenza Artificiale (AI), finalizzati alla classificazione e all'estrapolazione automatica delle informazioni contenute nelle mail PEC (Poste Elettroniche Certificate). Durante lo stage, è stata eseguita un'analisi dettagliata dei requisiti applicativi e tecnici necessari per implmentare una soluzione efficace e robusta.

L'attività di sviluppo ha incluso l'utilizzo di un modello di apprendimento automatico capace di analizzare il contenuto delle PEC importate, assegnando loro categorie appropriate basate su criteri come mittente, destinatario, data e argomento. In parallelo, è stato esplorato l'utilizzo di algoritmi avanzati di IA in grado di adattarsi e migliorare le prestazioni del modello attraverso l'apprendimento continuo dai dati e dai feedback ricevuti.

Infine, si è considerata l'integrazione con un sistema documentale per l'archiviazione automatica delle PEC, con la creazione dei metadati necessari e il loro posizionamento nella corretta categoria di appartenenza. Questi aspetti desiderabili, sebbene non obbligatori, hanno rappresentato un'opportunità di estendere la funzionalità del sistema, migliorando ulteriormente l'efficienza e l'accuratezza dell'archiviazione delle PEC.

Indice

Elenco delle figure

Elenco delle tabelle

Convenzioni tipografiche

Riguardo la stesura del testo, relativamente al documento sono state adottate le seguenti convenzioni tipografiche:

- gli acronimi, le abbreviazioni e i termini ambigui o di uso non comune menzionati vengono evidenziati in blu alla prima occorrenza nel documenti e definiti nel glossario, situato alla fine del presente documento;
- per la prima occorrenza dei termini riportati nel glossario viene utilizzata la seguente nomenclatura: *parola*[g];
- i termini in lingua straniera o facenti parti del gergo tecnico sono evidenziati con il carattere *corsivo*.

Capitolo 1

Introduzione

In questo capitolo andremo ad enunciare la struttura del documento ed analizzeremo l'azienda ospitante stage curricolare e l'offerta proposta.

1.1 L'azienda

Sanmarco Informatica S.p.A. (logo in figura 1.1) è un'azienda italiana di sviluppo software e consulenza informatica. Da oltre quarant'anni si dedica alla riorganizzazione dei processi aziendali in tutti i settori, progettando e implementando soluzioni digitali integrate.
L'azienda, che ad oggi conta più di 600 dipendenti e oltre 2500 aziende seguite ha come sede principale Villa Ramanelli a Grisignano di Zocco, in provincia di Vicenza, poco distante dai Centri di Ricerca e Sviluppo (CRS) e dal Centro per la Formazione di Vicenza. Conta anche diverse filiati in Trentino-Alto Adige, Friuli-Venezia Giulia, Lombardia, Piemonte, Emilia-Romagna, Toscana, Campania e Puglia.
L'obiettivo principale è l'innovazione e il progresso tecnologico, con l'obiettivo di creare soluzioni software che siano in grado di rispondere alle esigenze dei clienti, garantendo la massima qualità e sicurezza.
L'azienda è organizzata in *Business Unit*, dei centri di competenza specifici e autonomi ma in relazione costante. Ognuna delle quali è specializzata in un settore specifico. La *Business Unit* interessata dallo stage è XC situata nel Centro per la Formazione di Vicenza. Tale team composto da 10 persone, si oocupa di sviluppare e manutenere i servizi di XC.
La metodologia di lavoro, indipendentemente dalla Business Unit, è basata su un approccio [g]Agile implementata con il framework [g]Scrum. Agile è un approccio alla gestione dei progetti che si basa su principi di collaborazione, auto-organizzazione e flessibilità. Scrum è un framework Agile che permette di gestire progetti complessi, garantendo la massima trasparenza e la massima flessibilità e suddividendo il progetto in sprint ovvero periodi di tempo relativamente brevi in cui vengono fissati determinati obbiettivi ed attività.
Eventuali ulteriori informazioni sono disponibili sul sito web dell'azienda[1].

[1] https://www.sanmarcoinformatica.com/

Figura 1.1: Logo di Sanmarco Informatica

1.2 L'offerta di stage

L'obiettivo dello stage consiste nella catalogazione delle Poste Elettroniche Certificate ([g]PEC), integrando tecnologie di Intelligenza Artificiale ([g]AI) per l'analisi e l'efficienza del processo.
Il modello di apprendimento automatico analizza il contenuto delle PEC e le classifica in base al contenuto.
Il progetto è stato proposto dall'azienda in occasione dell'evento Stage IT 2024, organizzato dall'Università degli Studi di Padova e promosso da Confindustria Veneto Est. Tale evento mira ad agevolare l'incontro tra studenti e aziende, offrendo la possibilità di svolgere uno stage formativo con specifico riferimento al settore ICT (Information and Communication Technology). Tale settore si riferisce all'insieme delle tecnologie utilizzate per la gestione e la comunicazione delle informazioni, incluse quelle legate all'informatica e alle telecomunicazioni.

1.3 Organizzazione del testo

Il secondo capitolo descrive ...

Il terzo capitolo approfondisce ...

Il quarto capitolo approfondisce ...

Il quinto capitolo approfondisce ...

Il sesto capitolo approfondisce ...

Nel settimo capitolo descrive ...

Capitolo 2

Descrizione dello stage

In questo capitolo

2.1 Introduzione al progetto

L'elaborazione intelligente dei documenti ([g]IDP) è una tecnologia che automatizza il processo di immissione manuale dei dati da documenti cartacei o immagini digitali, integrandoli con altri processi aziendali digitali. Ad esempio, in un flusso di lavoro aziendale automatizzato, come l'invio di ordini ai fornitori al momento del calo delle scorte, l'IDP può sostituire l'immissione manuale dei dati da parte del team contabile. Invece di inserire manualmente i dati di una fattura ricevuta via e-mail, i sistemi di IDP estraggono automaticamente queste informazioni e le integrano direttamente nel sistema contabile, eliminando ostacoli e riducendo gli errori.

L'IDP offre numerosi vantaggi alle aziende. In termini di [g]Scalabilità, permette di elaborare documenti su larga scala con precisione, evitando errori umani e aumentando l'efficienza operativa. Promuove una **cultura dell'efficienza dei costi**, automatizzando attività ripetitive e riducendo i costi associati all'elaborazione manuale dei dati. Migliora anche la **soddisfazione dei clienti** grazie alla gestione più rapida e automatizzata dei documenti, come l'onboarding, le prenotazioni e i pagamenti, consentendo di fornire risposte personalizzate e veloci ai clienti.

Diversi settori traggono beneficio dall'IDP. Nel **settore sanitario**, facilita la gestione delle cartelle cliniche, migliorando l'estrazione e l'organizzazione dei dati dai documenti medici. Le **aziende finanziarie** lo utilizzano per automatizzare la gestione delle spese e l'elaborazione delle fatture, semplificando la gestione dei pagamenti. Nel **settore legale**, l'IDP analizza contratti e documenti legali, utilizzando tecnologie di elaborazione del linguaggio naturale ([g]NLP) per estrarre informazioni chiave. Le aziende della **logistica** lo impiegano per tracciare spedizioni e documenti di transito, riducendo gli errori umani. Infine, nel settore delle **risorse umane**, l'IDP semplifica la selezione del personale, gestisce le buste paga e automatizza altre funzioni HR.

Le tecnologie alla base dell'IDP comprendono il **riconoscimento ottico dei caratteri ([g]OCR)**, che converte immagini di testo in dati leggibili dalle macchine, e l'**elaborazione del linguaggio naturale

(NLP)**, che analizza e comprende il linguaggio umano. L'**automazione robotica dei processi (RPA)** consente invece di automatizzare i flussi di lavoro aziendali ripetendo azioni umane predefinite.

Il processo di IDP si articola in diverse fasi: acquisizione e **classificazione dei documenti**, **estrazione dei dati** rilevanti tramite OCR e NLP, **convalida e successiva elaborazione dei dati** nei sistemi aziendali, e **apprendimento continuo** attraverso algoritmi di machine learning per migliorare l'accuratezza nel tempo. Inoltre, i sistemi di IDP offrono **report e analisi** per ottimizzare ulteriormente i flussi di lavoro aziendali.

[g]AWS (AWS) supporta l'implementazione dell'IDP attraverso servizi come **Amazon Textract**, che utilizza il machine learning per estrarre informazioni dai documenti senza interazioni manuali, e **Amazon Comprehend**, che sfrutta l'NLP per scoprire informazioni preziose nei testi. Entrambi i servizi consentono alle aziende di automatizzare la gestione dei documenti in modo efficiente e sicuro, integrandosi con altre piattaforme aziendali per un flusso di lavoro senza interruzioni.

2.2 Requisiti e obiettivi

Gli obiettivi sono stati definiti in accordo con il tutor aziendale e si identificano nel seguente modo:

[Priorità][Id]

- Priorità: indica la priorità dell'obiettivo, può essere obbligatorio o desiderabile;
- Id: composto da due cifre, identifica l'obiettivo in modo univoco rispetto alla priorità.

ID	Categoria	Descrizione
O01	Obbligatorio	Analisi dei servizi AWS per l'addestramento dei modelli AI
O02	Obbligatorio	Addestramento di un modello di apprendimento AI utilizzando i servizi AWS
O03	Obbligatorio	Analisi requisiti applicativi e tecnici per implementare la soluzione richiesta
O04	Obbligatorio	Implementare un modello di apprendimento automatico che analizzi il contenuto delle PEC importate e assegni loro categorie appropriate in base al contenuto (mittente, destinatario, data e argomento)
D01	Desiderabile	Implementare algoritmi di AI in grado di adattarsi e apprendere continuamente dai dati per migliorare le prestazioni del sistema nel tempo. Ciò include l'ottimizzazione dei modelli di apprendimento automatico in base all'esperienza e ai feedback degli utenti

ID	Categoria	Descrizione
D02	Desiderabile	Integrazione con un sistema documentale per l'archiviazione delle PEC creando i metadati necessari con le informazioni estratte e collocandole nella corretta categoria di appartenenza

2.3 Pianificazione

2.3.1 Pianificazione settimanale

Il periodo di stage è stato suddiviso in 8 settimane, durante le quali sono previste le seguenti attività:

Settimana	Dal	Al	Attività
1	24-06-2024	28-06-2024	- Incontro con persone coinvolte nel progetto per discutere i requisiti e le richieste di implementazione - Ricerca, studio e documentazione per inquadramento progetto - Introduzione ai linguaggi di sviluppo - Introduzione agli ambienti di sviluppo - Introduzione dei servizi AWS
2	01-07-2024	05-07-2024	- Analisi dei servizi AWS per l'addestramento di un modello di apprendimento - Addestramento di un modello di apprendimento utilizzando i servizi di AWS **Milestone:** Utilizzo dei servizi AWS per l'addestramento di un modello di apprendimento
3	08-07-2024	12-07-2024	- Studio della soluzione per definire i requisiti necessari per l'implementazione **Milestone:** Analisi dei requisiti applicativi e tecnici per implementare la soluzione
4	15-07-2024	19-07-2024	- Addestramento modello di apprendimento per catalogare le PEC in base al loro contenuto
5	22-07-2024	26-07-2024	- Implementazioni per interfacciarsi con il modello di apprendimento addestrato e per poter catalogare le PEC importate **Milestone:** Completamento obiettivi minimi
6	29-07-2024	02-08-2024	- Implementazione algoritmo di AI per l'autoapprendimento

Settimana	Dal	Al	Attività
7	05-08-2024	09-08-2024	- Studio e documentazione sulle [g]Application Program Interface messe a disposizione dal documentale per poter catalogare le mail PEC - Implementazione dell'integrazione con il documentale producendo i metadati necessari per catalogare le PEC
8	12-08-2024	16-08-2024	- Verifica e test archiviazione PEC nel documentale **Milestone:** Completamento obiettivi massimi
9	19-08-2024	23-08-2024	- Recupero eventuali ritardi
10	26-08-2024	30-08-2024	- Recupero eventuali ritardi

Capitolo 3

Tecnologie e strumenti di interesse

In questo capitolo verranno descritti i servizi e le tecnologie analizzate e pertinenti per il problema descritto, in quale modo possono essere impiegate e una panoramica finalizzata a chiarirne il constesto e il caso d'uso.

3.1 Amazon Web Services

Amazon Web Services (AWS) è una piattaforma di servizi cloud che offre potenza di calcolo, storage di database, distribuzione di contenuti e altre funzionalità per aiutare le aziende a scalare e crescere. AWS offre una vasta gamma di servizi che possono essere utilizzati per implementare soluzioni di Artificial Intelligence (AI) e [g]Machine Learning (ML). Per la realizzazione dell'applicazione sono stati individuati diversi servizi che hanno permesso di realizzare un'architettura scalabile e serverless.

3.1.1 Amazon Comprehend

Amazon Comprehend (logo in 3.1) è un servizio di analisi del linguaggio naturale NLP che utilizza l'apprendimento automatico per identificare informazioni e connessioni utili nel testo. Tale servizio offre due funzionalità principali: Custom Classifier e Entity Classifier. Entrambi i servizi permettono di generare dei modelli dopo la fase di training creata mediante dei file csv (label e testo).
Un servizio di Comprehend utile in questo contesto è Flywheel che è il punto di riferimento principale per eseguire MLOps (Machine Learning Operations) per i modelli di Comprehend.
Amazon Comprehend è stato utilizzato per classificare i documenti nelle categorie selezionate.

Figura 3.1: Logo di Amazon Comprehend

3.1.2 Amazon Textract

Amazon Textract (logo in 3.2) è un servizio di Optical Character Recognition (OCR) che utilizza l'apprendimento automatico per riconoscere e analizzare il testo e i dati da immagini o documenti.
Tale servizio oltre ad effettuare l'identificazione ottica dei caratteri permette di identificare il contenuto del documento estraendone il testo, le tabelle, campi e relazioni. Oltre al contenuto rilevato, Amazon Textract, fornisce punteggi di confidenza e bounded box (box di confine) per ogni parola e ogni riga di testo. Supporta i file pdf, txt, doc, docx, jpg, png.
I casi d'uso che Textract ricopre sono i seguenti:

- Estrazione del testo non strutturato (DetectDocumentText). Questa funzionalità permette di estrarre i dati in forma di WORDS e LINES perdendo la formattazione struturale del documento.
- Estrazione ed elaborazione di moduli e tabelle (AnalyzeDocument, feature=TABLES)
- Estrazione di form (chiave/valore) (AnalyzeDocument, feature=FORMS)
- Estrazione tramite query. Possono essere degli strumenti potenti quando solo pochi pezzi o informazioni critiche sono desiderate. (AnalyzeDocument, feature=QUERIES).
- Rilevamento della firma. (AnalyzeDocument, feature=SIGNATURES). La risposta prevede la confidenza della rilevazione e il testo del documento (WORDS e LINES).
- Estrazione informazioni da fatture/ricevute (AnalyzeExpense)
- Estrazione informazioni dai documenti di identità (CALL_TEXTRACT_ANALYZE_ID)
- Rilevamento multicolonna

Per personalizzare le query si può usare uno strumento chiamato Custom Queries. Tramite questo strumento si può riconoscere:

- termini univoci
- strutture
- informazioni specifiche

Tale soluzione rispetto alle query non personalizzate offre maggiore precisione e un intervento umano minore. Amazon Textract è stato utilizzato per estrarre il testo dai documenti sia come input al classificatore di Comprehend sia per estrarre informazioni utili.

3.1.3 Amazon S3

Amazon Simple Storage Service (Amazon S3) è un servizio di storage di oggetti che offre scalabilità, disponibilità dei dati, sicurezza e prestazioni. Amazon S3 è progettato per memorizzare grandi quantità di dati a un costo molto basso. Amazon S3 è stato utilizzato per memorizzare i file inerenti alle diversi fasi del progetto.

Figura 3.2: Logo di Amazon Textract

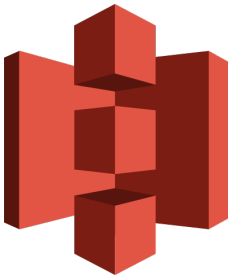


Figura 3.3: Logo di Amazon S3

3.1.4 AWS Lambda

AWS Lambda è un servizio di calcolo serverless che esegue il codice in risposta a eventi e gestisce automaticamente le risorse di calcolo richieste dal codice. AWS Lambda è stato utilizzato per implementare le funzioni di backend dell'applicazione.

Figura 3.4: Logo di AWS Lambda

3.1.5 Amazon DynamoDB

Amazon DynamoDB è un servizio di database NoSQL completamente gestito che offre prestazioni di singolo millisecondo a qualsiasi scala. Amazon DynamoDB è stato utilizzato per memorizzare i dati relativi ai vari utenti dell'applicazione.

3.1.6 Amazon Step Functions

AWS Step Functions è un servizio di orchestrazione di serverless che consente di coordinare facilmente i componenti di applicazioni distribuite e microservizi utilizzando logica visuale.

Figura 3.5: Logo di Amazon DynamoDB

Figura 3.6: Logo di Amazon Step Functions

3.1.7 Amazon Sagemaker

3.1.8 Amazon Bedrock

3.2 Strumenti di sviluppo

3.2.1 Visual Studio Code

Visual Studio Code è un editor di codice sorgente sviluppato da Microsoft per Windows, Linux e macOS. Visual Studio Code è stato utilizzato per scrivere il codice dell'applicazione.

Figura 3.7: Logo di Visual Studio Code

3.2.2 Git

[g]Git è un sistema di controllo di versione distribuito utilizzato per tenere traccia delle modifiche al codice sorgente durante lo sviluppo del software. Git è stato utilizzato per tenere traccia delle modifiche al codice sorgente dell'applicazione.

Figura 3.8: Logo di Git

3.2.3 Bitbucket

Bitbucket è un servizio di hosting di [g]repository Git basato su cloud. Bitbucket è stato utilizzato per memorizzare il codice sorgente dell'applicazione.

Figura 3.9: Logo di Bitbucket

3.3 Linguaggi di programmazione

3.3.1 Python

Python è un linguaggio di programmazione ad alto livello, interpretato, adatto per lo sviluppo di applicazioni web, desktop e mobile. Python è stato utilizzato per la realizzazione del backend dell'applicazione.

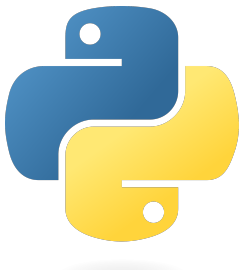


Figura 3.10: Logo di Python

Capitolo 4

Progettazione e codifica

Breve introduzione al capitolo

4.1 Introduzione

Partendo da quella che è la richiesta dell'azienda ospitante (catalogazione ed elaborazione delle mail e dei documenti allegati), ho individuato diverse fasi per il processo di elaborazione dei documenti dalle email:

- Estrazione degli allegati presenti in un'email
- Classificazione dei documenti
- Estrazione delle informazioni importanti dai documenti
- Revisione e valutazione umana delle informazioni estratte
- Persistenza dei dati

4.2 Estrazione degli allegati

In questa fase ,le indicazioni iniziali dell'azienda consistevano nell'analizzare il contenuto delle eventuali email ,per poi effettuarne una classificazione basata sull'elaborazione del linguaggio naturale e dei metadati contenuti. Tuttavia, con il chiarimnento delle categorie prese in analisi durante lo stage (ordini, fatture e contratti) , ho scelto di concentrarmi sull'estrazione degli allegati presenti nelle stessa piuttosto che sul contenuto della email. Tale visione è supportata dal fatto che i documenti di interesse per l'azienda sono spesso allegati delle mail, quindi tale estrazione determina un flusso di lavoro più chiaro e diretto, oltre al fatto che spesso il contenuto della mail non è rilevante o lo è solo in parte. Dunque, in questa fase il file di input è un file .eml , mentre l'output sono gli allegati presenti nella mail. Per progettare tutto ciò, ho pensato ai seguenti servizi:

- Un [g]bucket contenente i file .eml da analizzare

- Un bucket contente gli allegati estratti dalle mail
- Una funzione lambda che viene attivata dall'inserimento di un file .eml all'interno del bucket di input e che elabora tale file per ottenere degli allegati di output

4.3 Classificazione dei documenti

Dovendo classificare le email in base al loro contenuto, ho analizzato diverse strade possibili , come quella di utilizzare modelli di ML proposti da Sagemaker per classificare le email. Tuttavia, con il chiarimento delle categorie prese in analisi durante lo stage (ordini, fatture e contratti) , ho scelto di concentrarmi sull'estrazione degli allegati presenti nelle stessa piuttosto che sul contenuto della email. Questo cambio di prospettiva ha portato a una semplificazione del processo di classificazione, in quanto i metadati (denominati anche features nel campo del machine learning) si riducono al semplice testo estratto dagli allegati. Dunque, per poter classificare un documento in una delle categorie di interesse, ho pensato di utilizzare un modello di machine learning che prendesse in input il testo estratto e restituisse la categoria di appartenenza. In questo senso ,l'utilizzo di strumenti come Textract e Comprehend di AWS si è rivelato molto utile, in quanto permette di estrarre il testo dai documenti e di analizzarlo per ottenere informazioni utili. Tuttavia, c'è anche da sottolineare come in una fase iniziale si sia dibattuto riguardo l'utilizzo al loro posto del servizio Bedrock con claude-3. Tale servizio è stato poi scartato a favore di un modello customizzato e maggiormente adatto alla soluzione. In questa fase è necessario distinguire due tipi di flusso di lavoro:

- Flusso di lavoro per il training del modello
- Flusso di lavoro per la classificazione

4.3.1 Flusso di lavoro per il training del modello

In questa fase per poter addestrare un modello di Comprehend è necessario disporre di un dataset ampio ,significativo e bilanciato per poter distingure le categorie di interesse. Inoltre, è necessario disporre di un dataset etichettato, in cui ogni documento è associato alla sua categoria di appartenenza. Durante la fase di etichettatatura, sono emerse delle considerazioni importanti. Inanzitutto, i file da analizzare sono per lo piu file pdf (di documenti scansionati o meno), per tale motivo per la fase di training sono stati utilizzati unicamente file con tale estensione. Inoltre, si è scelto, in accordo con il tutor aziendale ,di utilizzare per il training unicamente le prime pagine di tali documenti per diversi motivi : spesso le prime pagine contengono le informazioni più importanti e rilevanti per la classificazione, inoltre, il costo di analizzare un documento è proporzionale al numero di pagine, quindi riducendo il numero di pagine si riducono i costi, assumendo anche il fatto che rispetto ai documenti incontrati il numero di pagine variava fino a 100 pagine. Queste due scelte (utilizzo di file pdf e utilizzo delle prime pagine) hanno portato a una riduzione della varietà dei dati e quindi a una riduzione della capacità del modello di generalizzare su nuovi dati ,creando dei potenziali [g]bias nel modello. Di seguito vengono riportati i dati e il loro numero per trainare la prima versione del modello

- 100 documenti per la categoria ordini
- 100 documenti per la categoria fatture
- 100 documenti per la categoria contratti
- 100 documenti per la categoria non classificato

Per il processo chiamato [g]active learning è stato scelto l'utilizzo di un servizio incluso in Comprehend introdotto recentemente da aws chiamato flywheel. Tale processo segue i seguenti passi:

- Viene creato un dataset flywheel
- Viene inizializzata un'iterazione flywheel
- In base ai risultati dell'iterazione viene scelto se attivare il nuovo modello formatosi in base a parametri scelti in precedenza

4.3.1.1 Analisi del dataset

Percentuale di oridine, fattura, contratti e non classificato

4.3.1.2 Preprocessing

- Estrazione del testo tramite Amazon Textract
- Creazione del file csv
- Caricamento del file di training csv trmite flywheel

4.3.1.3 Training

- Creazione di una versione del classificatore su Custom Classifier

4.3.1.4 Valutazione

4.3.1.5 Test del modello

4.3.2 Flusso di lavoro per la classificazione

Per tale fase, ho scelto di utilizzare i seguenti servizi:

- una funzione lambda che riceve in input un'allegato di qualsiasi tipo e (in base se è un pdf o meno) fa partire il processo di classificazione mediate il modello attivo di comprehend
- un modello attivo di comprehend che utilizza le funzionalità di textract per estrarre il testo in chiaro dal pdf ricevuto in input e restituisce la categoria di appartenenza con un certa confidenza
- tramite un'ulteriore funzione lambda viene analizzata la confidenza e in base a questo viene scelto se salvare l'allegato in un bucket che contiene gli allegati non classificati oppure in un bucket con alto grado di confidenza

4.4 Estrazione delle informazioni

In questa fase l'obiettivo è l'estrazione delle informazioni associate a ciascuna categoria escludendo la categoria non classificato. A partire dai risultati di classificazione della fase precedente si è analizzato il metodo migliore per poter estrarre le informazioni ricercate dalle categorie di contratti, ordini e fatture. Fondamentalmente sono stati analizzati diversi metodi utilizzando differenti servizi per aderire a tale scopo:

- Comprehend custom entities
- Amazon Bedrock
- features di textract

Digressione sui vantaggi e svantaggi ... Alla fine si è optato per le seguenti opzioni:

- Custom queries per le fatture
- Custom queries per gli ordini
- Analisi delle tabelle e dei form per i contratti

C'è da sottolineare che per ogni informazione estratta viene anche riportata la percentuale di confidenza. Il flusso per ogni categoria è il seguente:

- Quando un file viene caricato nel bucket relativo ai documenti classificati tale azione scatena l'esecuzione di una lambda apposita per il tipo di documento
- Al termine dell'esecuzione tali informazioni estratte vengono passate alla fase successiva

4.4.1 Estrazioni delle informazioni dai contratti

Per tale fase essendo i contratti della stessa forma, (una tabella con le seguenti informazioni ...) si è optato per un'opzione poco costosa ma comunque efficiace. Tale soluzione consiste nell'identificare tale tabella ed estrarne i campi in base alla conoscenze note.

4.4.2 Estrazione delle informazioni dalle fatture e degli ordini

Per tale fase si è pensato all'utilizzo di custom queries (adapter) di textract dato che tali documenti possiedono una struttura variabile. L'utilizzo di analisi delle fatture tramite la funzione apposita di textract è stata considerata ma poi scartata. Per gli ordini invece si è pensato di utilizzarla per ricavarne gli articoli in modo piu' diretto e sicuro.

4.5 Persistenza dei dati

In questa fase l'obiettivo è far persistere i dati. La scelta è ricaduta su Amazon DynamoDB. Il flusso è il seguente:

- Per ogni categoria (contratti, ordini, fatture) è creata una lambda, tale lambda salva i risultati delle informazioni estratte in DynamoDB nelle tabelle Ordini, Contratti, Fattura, Articoli_Fatture, Articoli_Ordini

4.6 Analisi dei costi

Capitolo 5

Sviluppi futuri

5.1 Analisi del contenuto della mail

Per poter analizzare il contenuto della mail ed estrarre le informazioni associate si può modificare la funzione lambda *processEmail* in modo tale da estrarre il testo della mail e non solo gli allegati.
Inoltre, si può implementare un modello di classificazione di Comprehend per classificare il testo della mail in base al contenuto analogamente a quanto fatto per gli allegati e successivamente estrarre le informazioni associate.

5.2 Aggiunta di nuove categorie

Si possono aggiungere nuove categorie di classificazione se necessario andando a modificare il modello di classificazione di Comprehend e in particolare il dataset fornito. Inoltre si possono aggiungere nuove funzioni lambda per l'estrazioni delle informazioni associate a ciascuna categorie.

5.3 Completamento delle informazioni

Si possono completare le informazioni mancanti non estratte interrogando il database DynamoDB. Questo lavoro si può fare tra lo step di 2 e lo step 3.

5.4 Sviluppo di un'interfaccia grafica

Capitolo 6

Conclusioni

Lorem [g]SDK

Lorem Application Program Interface

6.1 Consuntivo finale

Ipsum

6.2 Raggiungimento degli obiettivi

Sit amet

6.3 Conoscenze acquisite

6.4 Valutazione personale

Acronimi e abbreviazioni

AI Artificial Intelligence. i, 7, 20

API Application Programming Interface. i, 20

AWS Amazon Web Services. i, 7, 20

IDP Intelligence Document Processing. i, 20

ML Machine Learning. i, 7, 20

NLP Natural Language Processing. i, 20

OCR Optical Character Recognition. i, 8, 20

PEC Posta Elettronica Certificata. i, 20

SDK Software Development Kit. i, 21

UML Unified Modeling Language. i, 21

Glossario

Active learning Nell'ambito del machine learning per active learning si intende .... i, 14, 20

Agile Nell'ambito dell'ingegneria del software con il termine Agile si intende .... i, 1, 20

AI Per aritificial Intelligence (AI) si intende .... i, 2, 4, 5, 19

API In informatica con il termine *API* si indica ogni insieme di procedure disponibili al programmatore, di solito raggruppate a formare un set di strumenti specifici per l'espletamento di un determinato compito all'interno di un certo programma. La finalità è ottenere un'astrazione, di solito tra l'hardware e il programmatore o tra software a basso e quello ad alto livello semplificando così il lavoro di programmazione. i, 6, 18, 19

AWS Amazon Web Services (AWS) è una piattaforma di servizi cloud che offre potenza di calcolo, storage di database, distribuzione di contenuti e altre funzionalità per aiutare le imprese a scalare e crescere. i, 4, 5, 13, 19

Bias Nell'ambito del machine learning per bias si intende .... i, 13, 20

Bucket Nel contesto di AWS, per bucket si intende .... i, 12–14, 20

Git Git è un sistema di controllo di versione distribuito gratuito e open source progettato per gestire tutto, dai piccoli ai grandi progetti, con velocità ed efficienza. i, 11, 20

IDP Con il termine Intelligence document processing (IDP) si intende l'insieme di tecnologie che permettono di estrarre informazioni da documenti cartacei o digitali, elaborarle e trasformarle in dati strutturati. i, 3, 4, 19

ML Per Machine Learning (ML) si intende .... i, 13, 19

NLP Natural Language Processing (NLP) è ... i, 3, 4, 7, 19

OCR Optical Character Recognition (OCR) è .... i, 3, 4, 19

PEC La *Posta Elettronica Certificata* (PEC) è un servizio di posta elettronica che garantisce l'invio e la ricezione di messaggi di posta elettronica con valore legale equivalente a quello della raccomandata con avviso di ricevimento.. i, 2, 4–6, 19

Repository Con il termine repository si intende .. . i, 11, 21

Scalabilità In informatica, la scalabilità è la capacità di un sistema di crescere in dimensioni e complessità in modo lineare o sub-lineare rispetto all'aumento del carico di lavoro.. i, 3, 21

Scrum In ingegneria del software, per Scrum si intende .... i, 1, 21

SDK A software development kit (SDK) is a collection of software development tools in one installable package. They facilitate the creation of applications by having a compiler, debugger and sometimes a software framework. They are normally specific to a hardware platform and operating system combination. To create applications with advanced functionalities such as advertisements, push notifications, etc; most application software developers use specific software development kits. i, 18, 19

Serverless Per serverless si intende .... i, 7, 9, 21

UML In ingegneria del software *Unified Modeling Language* (ing. linguaggio di modellazione unificato) è un linguaggio di modellazione e specifica basato sul paradigma object-oriented. L'*UML* svolge un'importantissima funzione di "lingua franca" nella comunità della progettazione e programmazione a oggetti. Gran parte della letteratura di settore usa tale linguaggio per descrivere soluzioni analitiche e progettuali in modo sintetico e comprensibile a un vasto pubblico. i, 19

Bibliografia

Books

James P. Womack, Daniel T. Jones. *Lean Thinking, Second Editon*. Simon & Schuster, Inc., 2010.

Articles

Einstein, Albert, Boris Podolsky e Nathan Rosen. «Can Quantum-Mechanical Description of Physical Reality be Considered Complete?» In: *Physical Review* 47.10 (1935), pp. 777–780. DOI: 10.1103/PhysRev.47.777.

Siti web consultati

Active learning worflow for Amazon Comprehend. URL: https://aws.amazon.com/it/blogs/machine-learning/active-learning-workflow-for-amazon-comprehend-custom-classification-part-1/.

Amazon Comprehend. URL: https://aws.amazon.com/it/blogs/machine-learning/amazon-comprehend-document-classifier-adds-layout-support-for-higher-accuracy/.

AWS. URL: https://aws.amazon.com/.

aws samples. URL: https://github.com/aws-samples/aws-ai-intelligent-document-processing.

Comprehend idp. URL: https://aws.amazon.com/it/blogs/machine-learning/introducing-one-step-classification-and-entity-recognition-with-amazon-comprehend-for-intelligent-document-processing/.

comprehend samples. URL: https://github.com/aws-samples/amazon-comprehend-examples/blob/master/building-custom-classifier/BuildingCustomClassifier.ipynb.

flywheel. URL: https://aws.amazon.com/it/blogs/machine-learning/introducing-the-amazon-comprehend-flywheel-for-mlops/.

Intelligent document processing parte 1. URL: https://aws.amazon.com/it/blogs/machine-learning/part-1-intelligent-document-processing-with-aws-ai-services/.

invoice textract. URL: https://aws.amazon.com/it/blogs/machine-learning/announcing-expanded-support-for-extracting-data-from-invoices-and-receipts-using-amazon-textract/.

Manifesto Agile. URL: http://agilemanifesto.org/iso/it/.

sdk samples. URL: https://github.com/awsdocs/aws-doc-sdk-examples.

Textract. URL: https://aws.amazon.com/it/blogs/machine-learning/automatically-extract-text-and-structured-data-from-documents-with-amazon-textract/.

Textract bedrock. URL: https://aws.amazon.com/it/blogs/machine-learning/intelligent-document-processing-with-amazon-textract-amazon-bedrock-and-langchain/.