tinexta
defence

Retrieval Augmented Generation (RAG) per la consultazione locale e sicura di documenti: miglioramenti e nuove funzionalità

#TinextaDefenceBusiness

AI4Cyber

AI4Cyber studio, è il nuovo spazio di ricerca applicata di Tinexta Defence, dedicato alla ricerca e applicazione dell'**Intelligenza Artificiale in ambito cybersecurity**.

Al suo interno, diversi specialisti con competenze ed esperienze eterogenee collaborano per affrontare le sfide poste dal panorama delle minacce informatiche, accompagnare i clienti nell'adozione dell'AI nei propri processi aziendali e condividere le attività di ricerca con la comunità tecnico-scientifica.

Il team cura l'**intero ciclo di sviluppo dei sistemi basati su AI**: dalla raccolta e dal preprocessing dei dati, all'addestramento e validazione dei modelli, fino al deployment in produzione.

Le soluzioni proposte si fondano su tecnologie avanzate di **Machine Learning**, **Deep Learning** e **Large Language Models**, e possono essere personalizzate in base alle specifiche esigenze del cliente.

L'AI Team è costantemente impegnato in attività di ricerca e sperimentazione, con un'attenzione particolare agli **aspetti etici**, alla **trasparenza** e alla **tutela della privacy**.

L'obiettivo è proporre soluzioni innovative che siano in linea con i principi di **equità, inclusività e rispetto dei diritti fondamentali**.

Summary

Abstract

Il presente studio illustra l'evoluzione del sistema “Retrieval-Augmented Generation (RAG) per la consultazione locale e sicura di documenti” [1], già descritto precedentemente, avente come scopo non solo il miglioramento dell’interazione fra utente e documenti di varia natura, ma anche l’introduzione di nuove funzionalità.

Infatti, rispetto allo studio precedente, si introduce un nuovo approccio all’interrogazione dei documenti e si integrano diversi moduli aggiuntivi, pensati per offrire uno strumento completo e tecnologicamente avanzato.

La componente di Retrieval mantiene l’impostazione già descritta in [1], combinando un sistema di ricerca basato su rappresentazioni numeriche dei contenuti (vettori) con un corpus indicizzato tramite TF-IDF [2]. Tale struttura è supportata da meccanismi di similarità semantica e da tecniche avanzate di normalizzazione e pulizia testuale, già consolidate nella versione iniziale del sistema. Questa base solida ha permesso di puntare sull’affinamento del contesto recuperato, migliorando ulteriormente la precisione e la coerenza delle risposte generate.

Oltre ad una nuova gestione documentale che consente il supporto di contenuti testuali, audio e visivi, il sistema è reso accessibile tramite una dashboard interattiva e user-friendly, anch’essa migliorata rispetto alla versione descritta in [1].

Inoltre, la configurabilità della soluzione in base alle risorse hardware disponibili e alle preferenze dell’utente la rende particolarmente scalabile, flessibile e adatta a scenari applicativi reali, anche in contesti aziendali.

Infine, ogni utente dispone di uno spazio dedicato, che garantisce la separazione e la sicurezza dei dati, consentendo un’esperienza personalizzata e protetta.

Autori:

- Simone Parrella: Analista AI4CYBER
- Simona Sorgente: Team Leader AI4CYBER

1. Introduzione

L'architettura presentata in [1] è stata arricchita con una serie di componenti avanzati che ne estendono significativamente le capacità applicative. La struttura di base, già articolata nei moduli di indicizzazione, recupero e generazione, è stata mantenuta e potenziata attraverso l'integrazione di nuove funzionalità mirate all'elaborazione multimodale, all'interazione personalizzata e al miglioramento della precisione contestuale.

È stato introdotto un modulo di elaborazione audio in grado di trascrivere flussi vocali, tradurli in altre lingue e sintetizzarne i contenuti principali. Questo modulo supporta inoltre la diarizzazione, consentendo di distinguere i singoli speaker all'interno di un audio ed esporta separatamente le rispettive tracce vocali.

Il sistema supporta ora l'indicizzazione di una gamma estesa di formati, inclusi file audio, immagini e documenti sia strutturati che non strutturati. La fase di pre-processing è stata potenziata con meccanismi di normalizzazione automatica, rendendo l'indicizzazione più robusta e uniforme.

Per quanto riguarda i contenuti visivi, è stato integrato un modulo di descrizione automatica basato su modelli multimodali, capace di estrarre rappresentazioni semantiche dalle immagini e generare descrizioni coerenti rispetto a quanto raffigurato.

È stato inoltre sviluppato un sistema di riassunto condizionato, che consente di sintetizzare porzioni specifiche di testo, come paragrafi o intervalli di pagine, sulla base delle esigenze dell'utente.

La componente di question-answering è stata migliorata introducendo la possibilità di limitare le risposte a documenti selezionati, aumentando così il controllo sul contesto informativo.

Infine, è stata implementata una memoria persistente delle interazioni che conserva lo storico delle conversazioni e dei documenti indicizzati, permettendo al sistema di adattarsi dinamicamente al contesto d'uso e di offrire un retrieval progressivamente più coerente e personalizzato.

2. Modulo di indicizzazione

La fase di indicizzazione consiste nella trasformazione dei documenti in una forma che ne consenta un accesso e una ricerca efficaci e immediati. Questo processo prende avvio dall'estrazione del testo dai vari formati di file, operazione fondamentale per garantire la qualità e l'efficacia delle elaborazioni successive. I documenti già indicizzati vengono memorizzati in modo tale da evitare la duplicazione dei contenuti durante indicizzazioni successive.

È possibile scegliere di indicizzare il documento in maniera strutturata (standard) o non strutturata (libera):

Quando si sceglie la modalità strutturata, il documento viene suddiviso in paragrafi che vengono poi frammentati in chunk* di lunghezza controllata tramite RecursiveCharacterTextSplitter* di LangChain [3], così da garantire coerenza negli embedding* generati.

Se si seleziona la modalità libera, il sistema utilizza una funzione dedicata che suddivide il testo in chunk di dimensioni e caratteristiche analoghe a quelle adottate nella modalità standard, ma senza basarsi sulla struttura formale del documento.

Gli embedding generati vengono archiviati in un indice vettoriale implementato con FAISS [4], che permette di effettuare ricerche efficienti basate sulla similarità semantica tra i contenuti. A ogni unità di testo indicizzata vengono inoltre associate informazioni contestuali quali il nome del file, la sezione di origine (per documenti strutturati), la categoria del contenuto e la porzione testuale originale.

**Chunk: Porzione di testo in cui viene suddiviso un documento per facilitarne l'indicizzazione e la ricerca mirata.*

**RecursiveCharacterTextSplitter: Funzione di LangChain che suddivide il testo in chunk in modo ricorsivo, seguendo una gerarchia di delimitatori (paragrafi, frasi, parole), mantenendo la coerenza del contenuto.*

**Embedding: Rappresentazione numerica del significato di un testo. Consente confronti semantici tra contenuti per una ricerca più intelligente.*

2.1 Modalità strutturata

Nel caso di documenti strutturati, come quelli composti da paragrafi ben definiti o da parole chiave che delimitano parti del testo in maniera intuitiva, è stata implementata una strategia di analisi basata su un LLM personalizzabile tramite Docling [5], che converte il testo e lo traduce in MARKDOWN(.md) *.

Docling adotta un approccio avanzato di analisi documentale che va oltre il semplice riconoscimento basato sulle caratteristiche del font. Grazie all'analisi strutturale del documento è in grado di comprendere gerarchie, relazioni tra sezioni, tabelle e altri elementi complessi, garantendo una rappresentazione più accurata e fedele del contenuto originale.

Questo consente di identificare le principali sezioni logiche del testo (ad esempio capitoli, titoli e paragrafi), separandole in modo coerente utilizzando gli indicatori delle sezioni tipiche dei file md (##, **, ###). Ogni sezione così individuata è etichettata come strutturata.

Questa tecnica è efficace su ogni tipo di file, compresi quelli scansionati, permettendo una strutturazione intelligente anche in presenza di layout complessi. In precedenza, l'OCR utilizzato era Tesseract [6], ma è stato sostituito da EasyOCR [7], che garantisce maggiore accuratezza e robustezza soprattutto su immagini complesse, senza la necessità di rielaborarle per rendere il testo leggibile.

**Markdown (.md) è un linguaggio di markup leggero che permette di scrivere testo formattato in modo semplice e leggibile, usato soprattutto per documenti, blog e codice.*

2.2 Modalità non strutturata

Se si utilizza la modalità non strutturata, il documento viene trattato come un unico blocco testuale. Il contenuto viene quindi sottoposto a un processo di pulizia, che utilizza espressioni regolari per individuare e rimuovere caratteri non significativi, elementi di disturbo o "rumore" (come spazi bianchi e simboli anomali). Successivamente, il documento viene classificato come non strutturato. Questa modalità è utile se si desidera ricercare informazioni puntuali senza necessità di un contesto approfondito e particolari spiegazioni.

2.3 Indicizzazione audio

Nel caso vengano indicizzate tracce audio, il sistema le elabora come qualsiasi altro documento: prima si effettuata la trascrizione del testo, successivamente, se è stata selezionata la modalità strutturata, il file sarà gestito come un paragrafo distinto, suddiviso in chunk per facilitare la ricerca; se invece si procede adottando la modalità libera, ogni chunk dell'audio sarà trattato e recuperato singolarmente.

3. Modulo di ricerca per pagine

Il sistema non si limita a rispondere in base ai contenuti indicizzati, ma supporta anche query focalizzate su singole pagine o intervalli di pagine. In questo modo, l'utente può ricevere risposte mirate, limitate alle specifiche sezioni del documento, aumentando così la precisione e la rilevanza delle informazioni fornite.

Per implementare questa funzionalità è stato sviluppato un modulo che, tramite la dashboard, consente di selezionare specifiche pagine di un documento e utilizzarne il relativo contenuto come contesto per porre domande al modello.

A differenza della versione precedente, il modulo consente ora un'estrazione mirata delle pagine selezionate, circoscrivendo con maggiore precisione il contesto di riferimento. Questo approccio permette di operare direttamente sulle pagine indicate, senza limitarsi ai soli paragrafi in esse contenuti.

4. Modulo di sintesi documentale

Il modulo permette di sintetizzare in modo mirato specifiche porzioni di un documento, offrendo due modalità operative.

Nella modalità che sfrutta le pagine, il contenuto testuale viene estratto direttamente dalle pagine indicate e fornito al modello per la sintesi.

Nella modalità focalizzata sui paragrafi, il sistema utilizza le unità testuali individuate durante l'indicizzazione, consentendo all'utente di selezionare tramite la dashboard il paragrafo da elaborare.

5. Modulo di traduzione

Per evitare che i dati vengano trasmessi a terze parti, compromettendo la riservatezza delle informazioni trattate, è stato implementato un modulo di traduzione locale basato su modelli Transformers [8], specializzato nella traduzione bidirezionale tra italiano e inglese.

I modelli adottati, Helsinki-NLP/opus-mt-it-en [9] e Helsinki-NLP/opus-mt-en-it [10], operano interamente in locale, garantendo una traduzione accurata senza dipendere da provider esterni come Google Translate o DeepL.

6. Modulo di elaborazione audio

Tramite l'utilizzo di Whisper [11] è possibile convertire file audio in testo. Il sistema mette a disposizione diverse modalità di trascrizione, che si differenziano per accuratezza, velocità e requisiti computazionali:

- **tiny:** modalità estremamente veloce e leggera, con un livello di accuratezza limitato; indicata per dispositivi con risorse ridotte o per trascrizioni rapide.
- **base:** un buon compromesso tra velocità ed accuratezza, più preciso di tiny ma ancora poco efficiente.
- **small:** adatto per scenari generici; bilancia meglio precisione e tempi di elaborazione.
- **medium:** più accurato, ma anche più lento; consigliato per trascrizioni di qualità o contenuti complessi.
- **large:** è la modalità più precisa in quanto supporta più lingue e gestisce meglio accenti e rumori di fondo, ma richiede risorse elevate.
- **turbo:** una versione ottimizzata per velocità e scalabilità, offre prestazioni simili a large ma con una latenza inferiore, ideale per applicazioni in tempo reale o su larga scala.

È possibile, inoltre, scegliere tra diverse modalità d'azione:

- Trascrizione,
- Traduzione e trascrizione bidirezionale Italiano Inglese,
- Riassunto del contenuto.

6.1 Modulo di separazione audio

Questo modulo permette di separare le voci che intervengono in un file audio utilizzando il modello “pyannote/speaker-diarization” [12]. Il processo, detto “speaker diarization”, segmenta l’audio in intervalli associati a ciascun parlante, permettendo così di:

- suddividere la trascrizione in base a chi sta parlando,
- migliorare la leggibilità e l’analisi dei contenuti,
- facilitare l’associazione tra interventi vocali e utenti nei dialoghi.

Il modulo può essere utile, ad esempio, per trascrivere riunioni, interviste o conversazioni multi-parlante, mantenendo la struttura del dialogo e preservandone il contesto delle interazioni. Inoltre, è possibile scaricare i diversi intervalli in file audio separati.

7. Modulo di elaborazione visiva

Il modulo di elaborazione visiva consente di estrarre le immagini contenute all'interno di un file per analizzarle e restituirne una descrizione testuale.

Una volta estratte, le immagini possono essere processate da un modello di visione artificiale in grado di generare didascalie automatiche o riassunti dello scenario, rendendo le informazioni grafiche fruibili anche in forma testuale.

Dal menu è possibile selezionare con quale modello effettuare l’operazione, in modo da adattare il sistema a diversi livelli di prestazioni hardware.

8. Modulo di memoria conversazionale

Il modulo di memoria conversazionale, se attivato, consente di richiamare domande e risposte precedentemente scambiate con il modello. Ogni coppia domanda-risposta viene archiviata in un indice consultabile: quando una nuova domanda presenta un alto grado di similarità con quelle memorizzate, il modulo amplia il contesto a disposizione del modello, migliorando la coerenza e la continuità delle risposte.

9. Configurazione del sistema e parametri operativi

La dashboard consente la configurazione dinamica di numerosi parametri di sistema. In particolare, è possibile:

- Selezionare il modello linguistico da utilizzare per la generazione delle risposte;
- Selezionare i modelli per l'elaborazione di contenuti audio e visivi;
- Modificare i parametri di indicizzazione, come la dimensione dei chunk e l'overlap* tra segmenti testuali;
- Regolare la profondità del contesto fornito al modello durante l'interrogazione, selezionando il numero di chunk da richiedere al modulo di Retrieval;
- Attivare o disattivare il processo avanzato di elaborazione delle query, descritto nell'articolo precedente [1];
- Abilitare o meno l'utilizzo della memoria conversazionale;

- Scegliere la modalità di indicizzazione (strutturata o libera) per ciascun documento;
- Gestire la memoria del sistema, ad esempio ripulendola o resettandola;
- Creare e gestire profili utente distinti per mantenere separati i corpus documentali;
- Limitare le ricerche a un sottoinsieme selezionato del corpus totale di documenti;
- Definire le modalità operative dei moduli di sintesi testuale, elaborazione audio e traduzione.

10. Funzionalità avanzate del sistema

Oltre a gestire PDF con contenuto testuale estraibile e PDF basati su immagini o scansioni, il sistema è in grado di trattare anche:

- L'elaborazione di documenti in diversi formati.
- La trascrizione e l'analisi di contenuti audio.
- L'isolamento di speaker diversi nelle tracce audio.
- La ricerca per pagine singole anziché intervalli che delimitano i paragrafi.
- Il riassunto di pagine o paragrafi specifici.
- L'interrogazione mirata su documenti specifici.
- L'analisi delle immagini contenute nei file.
- La gestione della memoria dei documenti già indicizzati.
- Il mantenimento del contesto conversazionale durante l'interazione.
- L'elaborazione di richieste da parte di più utenti in simultanea.
- La configurazione dinamica dei parametri in base alle risorse hardware disponibili e alle preferenze dell'utente.

11. Limiti

Durante la fase di testing sono emerse alcune criticità tipiche dei sistemi basati su RAG, molte delle quali sono state efficacemente mitigate con gli ultimi aggiornamenti, grazie a:

- Riconoscimento più accurato della struttura documentale: grazie a nuovi algoritmi di parsing e segmentazione, il sistema interpreta con maggior precisione titoli, paragrafi e tabelle, riducendo la necessità di intervento manuale da parte dell'utente.

- Migliore elaborazione dei documenti scansionati: l'integrazione potenziata con EasyOCR [7] ha aumentato la qualità dell'estrazione anche in presenza di layout complessi e rumore visivo.

- Riduzione della ridondanza documentale tramite filtri intelligenti: sono stati implementati metodi avanzati per la selezione dei documenti rilevanti, che limitano la duplicazione delle informazioni prima dell'interrogazione.

12. Conclusioni e sviluppi futuri

Il sistema sviluppato si è dimostrato una soluzione solida ed efficace per l'analisi e la consultazione di documenti aziendali, migliorando significativamente l'efficienza delle attività di recupero e comprensione delle informazioni. Grazie all'integrazione con Docling, è ora possibile gestire con successo una vasta gamma di tipologie di documenti, superando uno dei principali limiti delle versioni precedenti.

Uno degli aspetti più significativi del sistema è la possibilità di disporre di un assistente AI locale, in grado di operare in completa autonomia e nel rispetto dei principi della sicurezza informatica (Confidenzialità, Integrità, Disponibilità – CIA). In un contesto in cui l'uso di soluzioni cloud comporta potenziali rischi per la protezione dei dati, l'esecuzione locale rappresenta un vantaggio strategico, soprattutto in ambiti aziendali e ad alta riservatezza.

Gli sviluppi futuri si concentreranno principalmente su:

- la sperimentazione e l'adozione di modelli di linguaggio più avanzati e precisi, al fine di migliorare ulteriormente la qualità delle risposte e delle elaborazioni;
- l'integrazione con infrastrutture aziendali esistenti (come CRM, ERP o gestionali documentali), favorendo l'inserimento del sistema all'interno di ecosistemi operativi già consolidati.

In sintesi, il sistema si pone come un framework completo, scalabile e sicuro, in grado di rispondere in modo concreto alle esigenze moderne di accesso intelligente all'informazione.

Bibliografia

1. Articolo precedente: https://www.linkedin.com/feed/update/urn:li:activity:7346447176404733952/

2. TF-IDF: Introduction to Information Retrieval. Cambridge University Press, 2008.ISBN: 978-0521865715.

3. LangChain: https://python.langchain.com/docs/get_started/introductionMeta

4. FAISS: https://python.langchain.com/docs/integrations/vectorstores/faiss/

5. Docling: https://docling-project.github.io/docling/

6. Tesseract: https://github.com/tesseract-ocr/tesseract

7. EasyOCR: https://github.com/JaidedAI/EasyOCR

8. Transformers: https://github.com/huggingface/transformers

9. Helsinki-NLP/opus-mt-it-en: https://huggingface.co/Helsinki-NLP/opus-mt-it-en

10. Helsinki-NLP/opus-mt-en-it: https://huggingface.co/Helsinki-NLP/opus-mt-en-it

11. Whisper: https://github.com/openai/whisper

12. Speaker diarization: https://huggingface.co/pyannote/speaker-diarization

tinexta
defence

Defence Tech | Next |
Donexit | Foramil | Innodesi

Via Giacomo Peroni, 452 – 00131 Roma
tel. 06.45752720 – info@defencetech.it
www.tinextadefence.it

#TinextaDefenceBusiness