Workshop AI (day 1) Microsoft-AGIC-UniPadova

Funzionamento dell'infrastruttura

Il chatbot è integrato in un'architettura che consente l'interazione con l'utente su diversi canali, l'elaborazione delle domande tramite un LLM e la restituzione della risposta, sfruttando anche funzionalità avanzate di ricerca semantica e vettoriale.

Canali di comunicazione

Il bot è gestito tramite **Azure Bot Service**, che consente l'utilizzo diretto tramite API o l'integrazione con diversi canali predefiniti (come Microsoft Teams, WhatsApp, Facebook, ecc.).

Flusso di funzionamento del sistema

1. L'utente invia una domanda tramite il **front-end,** che nel caso del workshop è una webapp integrata con il **bot framework**. Quindi Azure Bot Service inoltra la richiesta al backend, includendo il **contesto della conversazione** fino a quel momento.
2. Il **backend** effettua una chiamata al **servizio Azure OpenAI**, integrando il contesto e le informazioni per accedere a Azure AI Search.
3. **Azure AI Search** esegue in autonomia la ricerca negli indici, **recuperando i contenuti rilevanti dalla** knowledge base.
4. **La LLM genera una risposta basata sia sul contesto sia sui contenuti trovati**, e la invia al backend.
5. **Il backend restituisce la risposta al front-end**, che la presenta all'utente.

Azure Open AI fornisce metodi per interagire con i modelli e ottenere le risposte. Se il modello LLM viene modificato è necessario aggiornare manualmente la configurazione del backend. È possibile però utilizzare una

soluzione di API unification: **Model Inference API parametrica**. Con questo approccio, l'endpoint e le credenziali possono essere passati come parametri, permettendo di riutilizzare lo stesso codice per testare modelli diversi con pochissime modifiche.

Esplorazione della KB

La fase iniziale del progetto RAG prevede sempre una **esplorazione della knowledge base**.

- Se i dati includono immagini o audio, possono essere sfruttati tramite skillset appositi.
- Non esiste un metodo unico per esplorare i dati, spesso si analizzano a campione.

Azure AI Search

Datasource

Sono le fonti da cui Azure AI Search può estrarre i dati da indicizzare (ad esempio Blob Storage per file non strutturati come PDF).

Indexer

Sono processi automatizzati che leggono i dati dai **data source**, li trasformano e li inseriscono nell'indice seguendo un preciso schema. Preleva i dati dalle sorgenti e li mappa nei campi dell'indice. Possono essere schedulati per aggiornare periodicamente l'indice. Se collegati a uno **skillset**, possono anche applicare elaborazioni avanzate prima di scrivere i dati nell'index.

Skillset

Sono gli strumenti che l'**indexe**r utilizza per arricchire o modificare i dati da inserire nell'indice, es. OCR, speech-to-text, embedding, ecc.

Index

Rappresenta il contenuto ricercabile su cui operano funzionalità come la full-

text search, la ricerca vettoriale e la ricerca ibrida. L'indice è definito da uno **schema** ed è separato dalle sorgenti dati originali. Permette una efficiente ed elaborata ricerca delle informazioni.

Ha diversi attributi configurabili (retrievable, filterable, sortable, facetable, searchable, ecc.). **La configurazione dei campi dell'indice influenza direttamente la precisione dei risultati**.

Tutti questi oggetti sono gestibili via interfaccia Azure o tramite JSON.

Evaluation delle risposte

Metodi principali

1. **Confronto con la Ground Truth:**

   Le risposte del chatbot vengono confrontate con delle risposte ideali (fornite da un operatore umano).

2. **Confronto con il contesto**

   Le risposte vengono confrontate con i chunk di informazioni restituiti da Azure AI Search. Si valuta se la risposta è pertinente e precisa rispetto al contesto fornito.