

# Q&A through Retrieval Augmented Generation



Documentazione Q&A Medico

Antimo Barbato  
M63/1079

## Sommario

1. Introduzione .....	3
2. Fasi del progetto.....	4
2.1 Raccolta dati con Selenium .....	4
2.2 Pre-Processing dei dati .....	4
2.3 Basedati e RAG .....	6
2.3.1 Embedding.....	7
2.3.2 MongoDB.....	7
2.3.3 LLM.....	8
2.4 Visualizzazione e analisi dei dati .....	10
2.4.1 Filtro di ricerca .....	12
2.4.2 Analitiche .....	16
2.5 Esempi di utilizzo della sezione Q&A .....	21
Query 1: Consigli su un alimentazione sana e bilanciata .....	21
Query 2: Come faccio a capire se sto soffrendo di depressione?.....	22
Query 3: Vorrei notizie sulla partita nel napoli .....	23
Query 4: La pasta e patate è consigliata nella dieta? .....	24
Query 5: Palestra e alimentazione sono correlate? .....	25
Query 6: Ho dolore allo stomaco, quali possono essere le cause?.....	26
Query 7: Problemi di pressione cardiaca .....	27
Conclusioni .....	28

# 1. Introduzione

Negli ultimi anni, l'avvento dei modelli di linguaggio di grandi dimensioni (LLM) ha rivoluzionato numerosi ambiti applicativi, tra cui i sistemi di Domanda-Risposta (Q&A). Questi sistemi, grazie alla loro capacità di comprendere e generare testo in linguaggio naturale, rappresentano uno strumento potente per rispondere a domande basate su informazioni specifiche. Tuttavia, le prestazioni dei modelli LLM possono essere ulteriormente migliorate attraverso tecniche avanzate come il *Retrieval-Augmented Generation* (RAG). Il RAG combina il recupero di informazioni da fonti esterne con la generazione testuale, consentendo ai modelli di incorporare conoscenze aggiornate e contestualmente rilevanti nelle loro risposte.

Nel contesto della sanità, l'aumento esponenziale delle informazioni mediche disponibili offre l'opportunità di implementare sistemi Q&A in grado di supportare i professionisti nella diagnosi, nel trattamento e nella gestione dei pazienti. Questi sistemi possono offrire un accesso rapido e affidabile a informazioni essenziali, aiutando i medici a prendere decisioni accurate in tempi ristretti. Allo stesso modo, tali strumenti possono essere utilizzati dai pazienti per comprendere meglio i propri sintomi e ottenere indicazioni preliminari, pur senza sostituire una valutazione professionale.

Questo progetto si propone di sviluppare un sistema di Q&A in ambito medico che sfrutti un modello LLM e un approccio RAG per fornire risposte precise, dettagliate e comprensibili a domande formulate in linguaggio naturale.

## 2. Fasi del progetto

Lo sviluppo del sistema di Q&A si è articolato in diverse fasi:

- **Raccolta dei dati:** estrazione dei dati da fonti esterne tramite web scraping
- **Preprocessing dei dati:** pulizia e filtraggio dei dati.
- **Archiviazione del database:** salvataggio delle informazioni in un database con relativa strutturazione della stessa.
- **Integrazione con il modello di LLM:** utilizzo del modello di LLM con tecniche RAG per migliorare la capacità del sistema.
- **Analisi dei dati:** esecuzione di pipeline di analisi per estrarre insight significativi
- **Visualizzazione:** creazione di una dashboard interattiva per visualizzare le analisi e i report.

### 2.1 Raccolta dati con Selenium

La prima fase del progetto prevede la **raccolta dei dati grezzi**, realizzata mediante l'utilizzo di **Selenium**, una libreria open-source progettata per l'automazione delle attività di web scraping e interazione con le pagine web. Selenium è stata utilizzata su Google Colab, un ambiente di sviluppo collaborativo che consente l'esecuzione di script Python in cloud, semplificando l'accesso alle risorse computazionali e garantendo un flusso di lavoro flessibile e scalabile.

Le fonti selezionate per l'estrazione dei dati sono tre forum medici autorevoli citati nella traccia: *Dica33*, *MioDottore* e *MedicItalia*. Per ognuna di queste piattaforme, sono stati individuati ed estratti topic specifici, considerati rilevanti per il contesto medico e per la costruzione del sistema di Domanda-Risposta. Nello specifico:

- **Dica33:** Cuore e circolazione, malattie del sangue, apparato respiratorio, fegato, alimentazione, naso e gola, mente e cervello.
- **MioDottore:** Ansia, depressione.
- **MedicItalia:** Cefalea, colon irritabile, alimentazione.

La selezione dei topic è stata effettuata con l'obiettivo di coprire una gamma diversificata di argomenti medici.

A causa, inoltre, di problemi legati a come le pagine erano scritte in html in alcuni casi è stato necessario apportare delle trasformazioni, per ottenere alla fine del processo gli stessi attributi per tutti i dataset ottenuti, vedremo nel dettaglio nel paragrafo successivo.

### 2.2 Pre-Processing dei dati

Successivamente, si è proceduto al **filtraggio e alla trasformazione dei dati** per garantire coerenza ed evitare incongruenze nella base dati finale. Di seguito vengono descritte le principali operazioni effettuate:

- **Conversione della data:** Le date estratte dai dati grezzi erano inizialmente nel formato 16 Dicembre 2024. Per garantire compatibilità con una base dati come MongoDB, si è preferito riformatarle nel formato standard ISO-8601, ossia 2024-12-16.  
Si precisa che nel caso di *medicitalia* era presente anche l'ora, che è stata rimossa.
- **Estrazione di informazioni complesse:** A causa della struttura del codice HTML di alcune fonti, non è stato possibile estrarre direttamente i singoli campi di interesse. In particolare, nel caso di *Dica33*, è stato necessario estrarre un blocco di informazioni aggregate dal campo ANSWER, che conteneva i seguenti dati:
  - 1) **Data della risposta** (*Answer Date*);
  - 2) **Testo della risposta** (*Answer*);

- 3) Nome del medico (*Doctor*);
- 4) Specializzazione del medico (*Specialization*);
- 5) Luogo di esercizio (*Location*).
- 6) Lavoro di riferimento (*Job*).

L'output finale a seguito della divisione dei dati è il seguente

	URL	Question Date	Question	Answer Date	Answer	Doctor	Job	Specialization	Location
0	<a href="https://www.dica33.it/esperito-risponde/domanda...">https://www.dica33.it/esperito-risponde/domanda...</a>	2024-10-26	Buonasera ho 28 anni in salute vado in palestr...	2017-05-07	Le consiglio, dopo aver consultato il suo medi...	GIAN PIERO DI BARTOLOMEI	None	None	None
1	<a href="https://www.dica33.it/esperito-risponde/domanda...">https://www.dica33.it/esperito-risponde/domanda...</a>	2024-10-24	Salve vorrei chiedere informazioni riguardo al...	2017-05-07	Direi che il suo medico abbia ragione.	GIAN PIERO DI BARTOLOMEI	None	None	None
2	<a href="https://www.dica33.it/esperito-risponde/domanda...">https://www.dica33.it/esperito-risponde/domanda...</a>	2024-10-26	Buongiorno, ho effettuato ecocardiogramma come...	2017-05-07	Il colesterolo con la mitriale non c'entra rie...	GIAN PIERO DI BARTOLOMEI	None	None	None
3	<a href="https://www.dica33.it/esperito-risponde/domanda...">https://www.dica33.it/esperito-risponde/domanda...</a>	2024-11-01	Buongiorno sono portatore di pacemaker e attua...	2017-05-07	Immagino che dovrà fare una scintigrafia cardi...	GIAN PIERO DI BARTOLOMEI	None	None	None
4	<a href="https://www.dica33.it/esperito-risponde/domanda...">https://www.dica33.it/esperito-risponde/domanda...</a>	2024-10-30	Buonasera, sono un ragazzo di 27 anni(70kg per...	2017-05-07	Direi che il suo quadro lipidico è più che buo...	GIAN PIERO DI BARTOLOMEI	None	None	None
...	...	...	...	...	...	...	...	...	...
895	<a href="https://www.dica33.it/esperito-risponde/domanda...">https://www.dica33.it/esperito-risponde/domanda...</a>	2017-05-08	Buongiorno, sono una donna di 35 anni, mamma, ...	2017-05-07	I dolori che descrive ( poco ) non fanno certo...	ALBERTO TITTOBELLO	Casa di cura privata	Gastroenterologia	Milano (MI)

Figura 1: Output dall'operazione di riformattazione del dato Answer

- **Eliminazione parte del testo:** nel caso di MioDottore e Medicialia si è provveduto ad eliminare il testo a seguito di alcune parole chiavi poiché portava nel dataset dati inutili o ridondanti. Tale processo ha interessato il campo *Answer*.
- **Eliminazione spazi:** in alcuni casi il testo presentava spazi vuoti a causa di un uso improprio del tag <br> generando righe vuote tra linee di testo. Tale processo ha interessato il campo *Question* per quanto concerne i siti MioDottore e Medicialia.
- **Rimozione "Dott" e simili:** nel caso del campo *Dottore*, si è eliminato l'appellativo per ottenere in uscita solo il nome e il cognome.
- **Normalizzazione delle città:** sono stati rimossi spazi, valori dopo le parentesi e la città è stata scritta in minuscolo.

Dopo aver filtrato e trasformato i dati si è provveduto al merge dei dataset relativi ai vari topic in modo da ottenere un unico dataset coerente.

In seguito, si è provveduto ad effettuare un'ulteriore operazione di filtraggio sul dataset unificato, **andando a rimuovere i valori nulli** dai campi *Question* e *Answer*, poiché campi rilevanti per il sistema di Q&A.

<a href="https://www.dica33.it">https://www.dica33.it</a>	2023-05-01	Gentili dottori, Sono	2011-10-21	Buongiorno. Farei co	Avellino	Psichiatria	avellino
<a href="https://www.dica33.it">https://www.dica33.it</a>	2023-03-23	Preg. Mo dott, la pre	2011-10-21	Mi pare che sia sta s	FABIO TRECATE	Casa di cura conven	milano
<a href="https://www.dica33.it">https://www.dica33.it</a>	2023-03-10	Buongiorno da quasi	2011-10-21	Ha letto bene su inte	FABIO TRECATE	Casa di cura conven	milano
<a href="https://www.dica33.it">https://www.dica33.it</a>	2023-02-04	Salve mio figlio 3 gi	2011-10-21	Attendono che il qu	FABIO TRECATE	Casa di cura conven	milano
<a href="https://www.dica33.it">https://www.dica33.it</a>	2023-02-05	Salve a tutti, spero q	2011-10-21	Suo padre va inuac	FABIO TRECATE	Casa di cura conven	milano
<a href="https://www.dica33.it">https://www.dica33.it</a>	2023-01-23	Quante pillole di ber	2011-10-21	Bisogna vedere di qu	FABIO TRECATE	Casa di cura conven	milano
<a href="https://www.dica33.it">https://www.dica33.it</a>	2023-01-13	Mia madre morì di al	2011-10-21	Non esiste solo la De	FABIO TRECATE	Casa di cura conven	milano
<a href="https://www.dica33.it">https://www.dica33.it</a>	2023-01-06	Dimentico facilment	2011-10-21	Il quadro va sicuram	FABIO TRECATE	Casa di cura conven	milano
<a href="https://www.dica33.it">https://www.dica33.it</a>	2022-12-02	Salve È da quattro g	2011-10-21	La sua vertigine prot	FABIO TRECATE	Casa di cura conven	milano
<a href="https://www.dica33.it">https://www.dica33.it</a>	2022-11-06	Buongiorno premette	2011-10-21	Potrebbe trattarsi di	FABIO TRECATE	Casa di cura conven	milano
<a href="https://www.dica33.it">https://www.dica33.it</a>	2022-10-11	Sono quattro mesi cl	2011-10-21	Andrebbe indagato i	FABIO TRECATE	Casa di cura conven	milano
<a href="https://www.dica33.it">https://www.dica33.it</a>	2022-10-05	Il tritico 150 mg mi	2011-10-21	Certamente, ma se h	FABIO TRECATE	Casa di cura conven	milano
<a href="https://www.dica33.it">https://www.dica33.it</a>	2022-07-06	Ho la mamma di 80	2011-10-21	La sottoponga ad un	FABIO TRECATE	Casa di cura conven	milano
<a href="https://www.dica33.it">https://www.dica33.it</a>	2022-07-06	Egredi dottori, scrivo	2011-10-21	Il Tapentadolo ha ar	FABIO TRECATE	Casa di cura conven	milano
<a href="https://www.dica33.it">https://www.dica33.it</a>	2022-06-02	Salve, domando gen	2011-10-21	Buongiorno, ciò che	EDOARDO GENTILE Medico Ospedaliero	Neurologia	ciriè
<a href="https://www.dica33.it">https://www.dica33.it</a>	2022-06-02	Salve, da tre settima	2011-10-21	Buongiorno, i sintom	EDOARDO GENTILE Medico Ospedaliero	Neurologia	ciriè
<a href="https://www.dica33.it">https://www.dica33.it</a>	2022-06-03	Indagine eseguita co	2011-10-21	Buongiorno, non ho	EDOARDO GENTILE Medico Ospedaliero	Neurologia	ciriè
<a href="https://www.dica33.it">https://www.dica33.it</a>	2022-06-05	Cosa vuol dire il refe	2011-10-21	Buongiorno, il refert	EDOARDO GENTILE Medico Ospedaliero	Neurologia	ciriè
<a href="https://www.dica33.it">https://www.dica33.it</a>	2022-06-05	Per risolvere le vertig	2011-10-21	Buongiorno, come p	EDOARDO GENTILE Medico Ospedaliero	Neurologia	ciriè

Figura 2: Output del dataset finale

*Nota: il codice per Scraping e Pre-Processing è presente nel file 'Scraping\_Q&A.ipynb/.py*

## 2.3 Basedati e RAG

MongoDB è un database NoSQL orientato ai documenti, progettato per gestire in modo efficace dati eterogenei e scalare con facilità. Grazie al suo formato basato su JSON, MongoDB consente di archiviare informazioni strutturate e semi-strutturate in un'unica piattaforma, rendendolo ideale per la gestione di grandi volumi di dati provenienti da fonti diverse. Questa flessibilità lo rende una scelta eccellente come base per un sistema di **Retrieval-Augmented Generation (RAG)**.

MongoDB rappresenta il nucleo del nostro sistema RAG, fornendo un repository scalabile e flessibile per i dati utilizzati per l'arricchimento delle risposte generate dal modello di linguaggio. La RAG (Retrieval-Augmented Generation) è una tecnica avanzata che combina il recupero di dati da una base esterna con le capacità di generazione del linguaggio naturale di un modello di intelligenza artificiale (LLM, Large Language Model). Questo approccio mira a superare alcune delle limitazioni intrinseche dei modelli LLM, tra cui:

### 1. **Cut-off Date:**

- I modelli LLM hanno una "data di cutoff", cioè un limite temporale per la conoscenza acquisita durante il loro addestramento. Ciò significa che non possono includere informazioni più recenti senza un nuovo ciclo di addestramento, che è lungo e costoso.
- Con un sistema RAG, le informazioni aggiornate vengono recuperate dinamicamente dalla base dati, garantendo risposte sempre attuali e pertinenti.

### 2. **Allucinazioni:**

- I modelli LLM possono generare risposte sintatticamente corrette ma concettualmente errate o fuorvianti, spesso a causa di lacune nelle informazioni.
- Un sistema RAG riduce questo rischio attingendo a una base di conoscenza curata e verificata.

### 3. **Citazione delle fonti:**

- Le risposte generate dai modelli LLM non includono sempre informazioni sulle fonti, il che può creare problemi di trasparenza e fiducia.
- Integrando MongoDB come repository per i dati, il sistema può restituire non solo le risposte ma anche riferimenti diretti alle fonti utilizzate.

### 4. **Tempo di addestramento:**

- L'addestramento di un modello LLM è un processo lungo e dispendioso, che non può essere ripetuto frequentemente.
- Con RAG, l'arricchimento delle risposte avviene in tempo reale, senza richiedere il riaddestramento del modello.

Per concludere la parte teorica, MongoDB può essere configurato come un **Vector Database**, una funzionalità essenziale per consentire l'archiviazione degli embedding generati. Gli embedding ci permettono di rappresentare dati complessi (come testo, immagini o audio) in uno spazio numerico, dove le relazioni semantiche tra i dati possono essere facilmente calcolate utilizzando apposite metriche, come la similarità coseno. Pertanto, il nostro obiettivo in questa fase è vettorizzare i dati di interesse e archivarli nel database. Questi dati vettorizzati verranno poi confrontati con la query in ingresso, che sarà anch'essa trasformata in un embedding, utilizzando la similarità coseno per determinare la relazione semantica tra la query e i dati archiviati.

### 2.3.1 Embedding

Per il progetto di riferimento si sarebbe potuto effettuare l'embedding sia per la colonna delle Question che per quella delle Answer ma si è preferito avere un matching diretto con i dati relativi all'answer e la query in ingresso, di conseguenza l'embedding è stata effettuata sui dati relativi all'attributo answer.

Si è usato per l'embedding il modello "thenlper/gte-large" indicato per operazioni di questo tipo, sfruttando la libreria transformers di Hugging Face.

Figura 3: output parziale per l'embedding basato sulla colonna Answer

### 2.3.2 MongoDB

Successivamente, è stato creato un database su **MongoDB Atlas** con la relativa collezione, destinata a contenere i dati insieme ai loro embedding, pronti per essere inseriti e gestiti in modo efficiente.

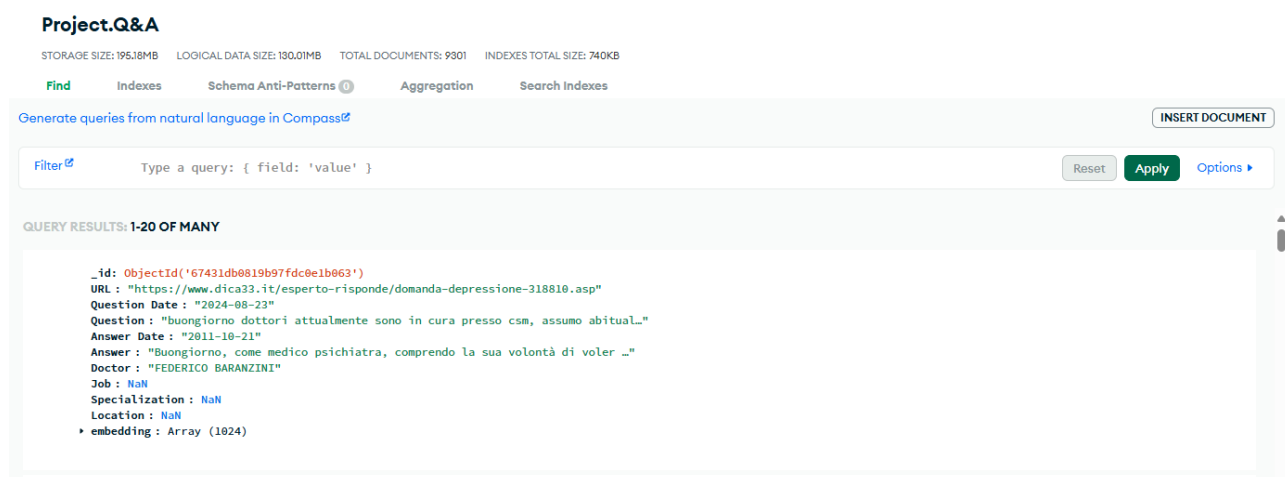


Figura 4: Inserimento dei dati su MongoDB Atlas

Successivamente, è stato creato un **Search Index** di tipo "**vectorSearch**" all'interno di MongoDB Atlas, per abilitare operazioni di **ricerca vettoriale** sulla base di dati. Questo indice consente di eseguire ricerche efficienti basate sulla similarità tra gli embedding.

Database	Collection	Index Name	Type	Index Fields	Status	Queryable	Size	Documents	Actions
Project	Q&A	vector_index	vectorSearch	"embedding"	READY <small>View status details</small>	✓	36,64MB	9,301 (100%) indexed of 9,301	QUERY ...

Figura 5: Indice vettoriale su MongoDB Atlas

Nello specifico l'indice vettoriale è stato definito come segue:

```
1  {
2    "fields": [
3      {
4        "numDimensions": 1024,
5        "path": "embedding",
6        "similarity": "cosine",
7        "type": "vector"
8      }
9    ]
10 }
```

Figura 6: Vector Search index

In particolare i campi di riferimento saranno i seguenti:

- **numDimensions**: indica il numero di dimensioni del vettore, che considerando il modello usato è stato impostato a 1024
- **path**: indica il percorso nel documento dove è presente la colonna con l'embedding
- **similarity**: indica la similarità usata per effettuare il confronto ed è appunto quella coseno
- **type**: specifica che il tipo di dato è un vettore

### 2.3.3 LLM

Per la selezione dell'**LLM** (Large Language Model) più adatto al progetto, si è optato per il modello "**mistralai/Mixtral-8x7B-Instruct-v0.1**" disponibile su Hugging Face. Questo modello, con **7 miliardi di parametri**, è particolarmente adatto per applicazioni di **Text Generation** grazie alle sue dimensioni e capacità di comprensione e generazione del linguaggio naturale. Inoltre, è un modello **multimodale**, che supporta diversi tipi di input (come testo, immagini e altri dati), ma per il nostro caso d'uso, è stato utilizzato principalmente per la generazione di risposte a partire da input testuali, rendendolo perfetto per un sistema di domande e risposte in ambito medico. Per ottimizzare le prestazioni del modello e garantire risposte appropriate e pertinenti, è stato creato un **prompt strutturato** utilizzando i tag **[INST]** e **[/INST]**. Questi tag sono stati scelti per indirizzare il modello a seguire istruzioni specifiche, migliorando la qualità delle risposte.

Nel dettaglio, il prompt inviato al modello contiene:

- **La query dell'utente**: la domanda o richiesta dell'utente, che è il punto di partenza per la generazione della risposta.
- **Il contesto estratto dal sistema RAG**: le informazioni rilevanti recuperate da fonti esterne, che aiutano il modello a generare una risposta più informata e accurata.
- **Le istruzioni specifiche**: indicazioni su come il modello deve rispondere, incluse le seguenti direttive:



- **Rispondere in italiano:** la lingua preferita per le risposte è l'italiano, in modo da allinearsi con il contesto linguistico dell'utente.
- **Utilizzare le informazioni di contesto:** il modello è stato istruito a sfruttare le informazioni di contesto fornite dal sistema RAG per arricchire e rendere più precise le risposte.
- **Non produrre una risposta se il contesto non è sufficiente:** nel caso in cui le informazioni disponibili nel contesto non siano adeguate per rispondere alla domanda, il modello è stato istruito a non fornire una risposta generica o errata, evitando così le allucinazioni.
- **Formattare la risposta in modo coerente e leggibile:** è stato richiesto al modello di formattare la risposta in modo chiaro e strutturato, in modo che sia facilmente comprensibile dall'utente.
- **Mantenersi nell'ambito del contesto medico:** per garantire la pertinenza delle risposte, il modello è stato indirizzato a concentrarsi esclusivamente su informazioni mediche e a non deviare verso altri ambiti non rilevanti per il caso d'uso.

Di seguito uno schema generale del sistema che si è realizzato

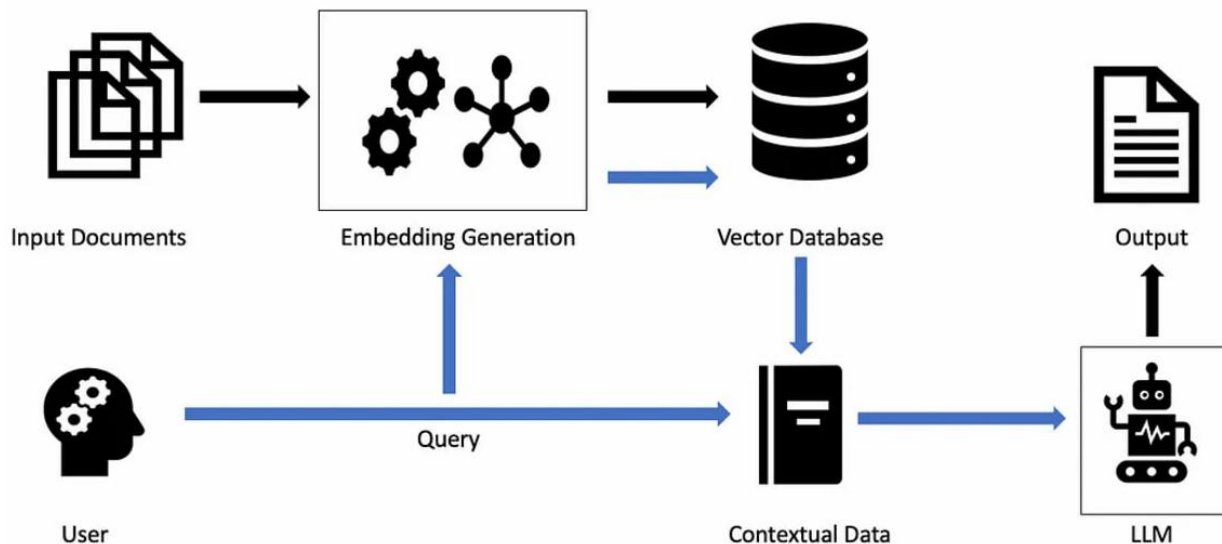


Figura 7: Schema del progetto (Fonte web)

## 2.4 Visualizzazione e analisi dei dati

Per quanto riguarda la visualizzazione dei dati si è usato Streamlit che è una libreria python dedicata allo sviluppo di applicazioni web interattive per la visualizzazione e la manipolazione dei dati.

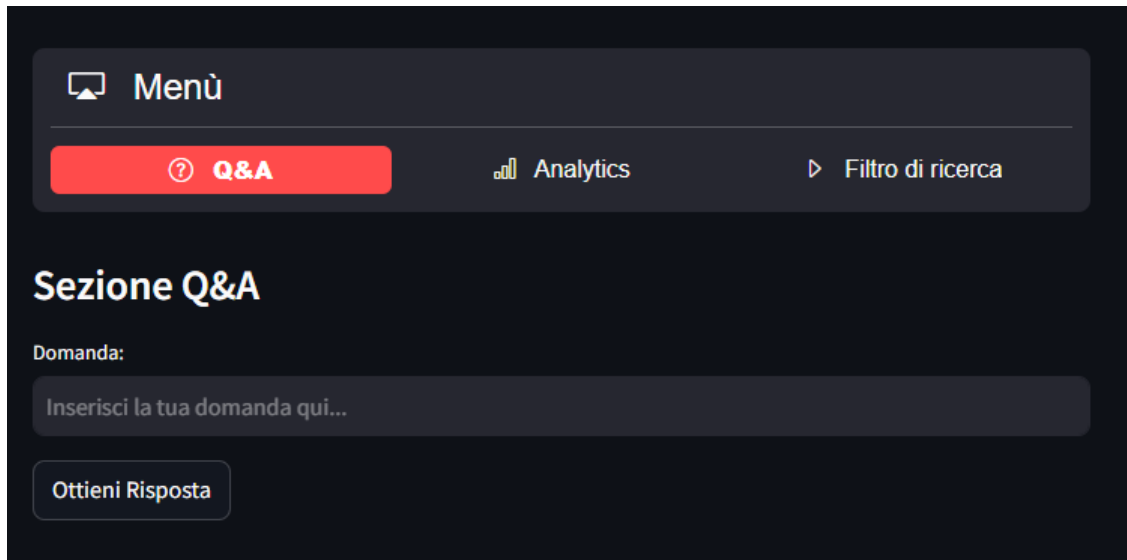
La navigazione avviene tramite l'uso di variabili di stato (`st.session_state`), che ci permette di determinare la pagina corrente e l'opzione selezionata dall'utente.

La prima pagina che incontreremo all'avvio sarà dunque la **Home**.



Figura 8: Home sistema di Q&A (logo AI Generated)

Successivamente, premendo il tasto “Vai al menu” incontreremo il menù di navigazione realizzato mediante la libreria “*streamlit\_option\_menu*”.



*Figura 9: Menù con le tre opzioni*

Potremo selezionare a partire dal menù 3 opzioni:

- **Q&A**
- **Filtro di ricerca**
- **Analytics**

Per quanto concerne la sezione **Q&A**, permette di interagire con il sistema di domande e risposte (RAG) in ambito medico. Gli utenti potranno porre domande specifiche relative alla medicina e ricevere risposte accurate generate dal sistema, basato su una raccolta di informazioni pertinenti. *Gli esempi di utilizzo saranno riportati alla fine del documento.*

### 2.4.1 Filtro di ricerca

Nella sezione **Filtro di ricerca**, siamo in grado di selezionare due funzioni di ricerca:

- **Ricerca dottori per specializzazioni**
- **Analisi delle domande per anno e mese**

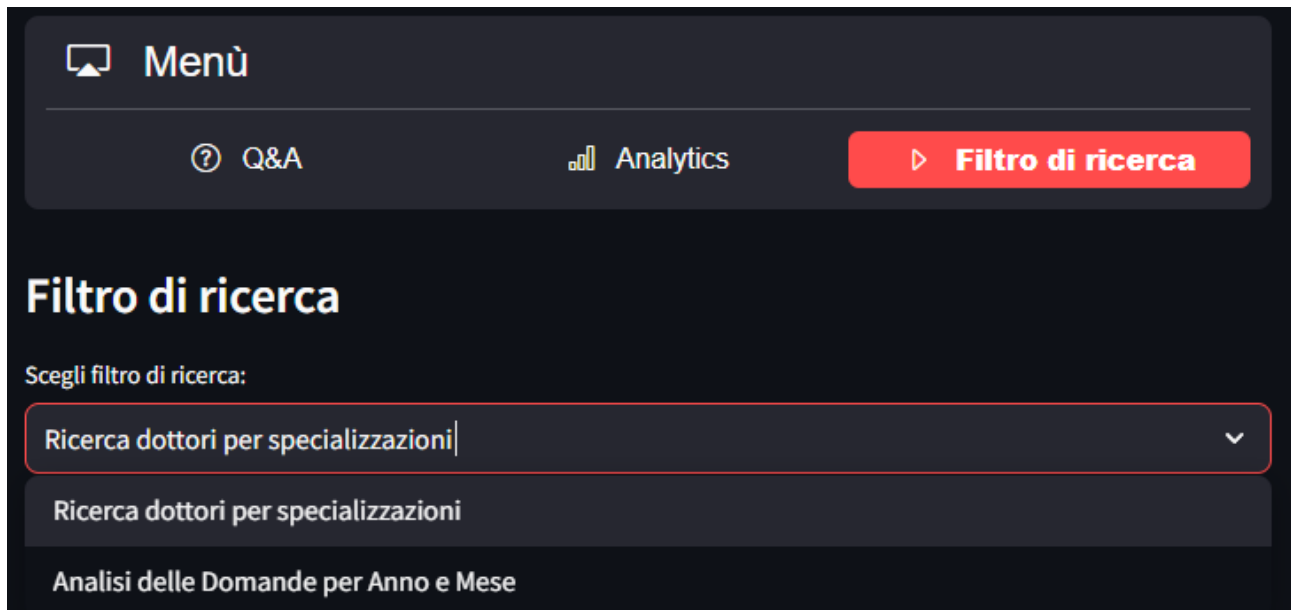


Figura 10: Menù per il filtro di ricerca

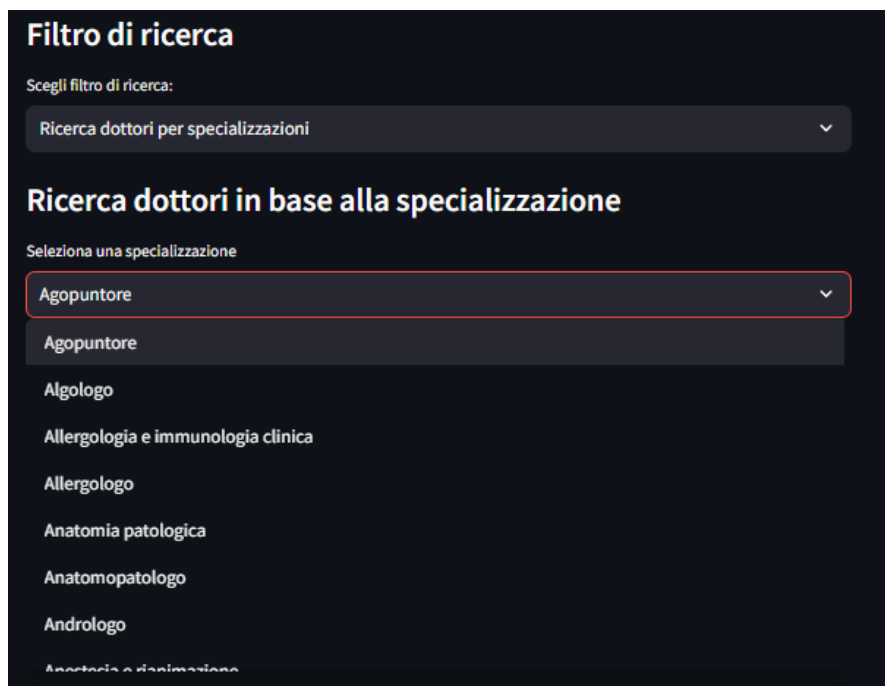
### Ricerca dottori per specializzazioni

Siamo in grado selezionando un'apposita specializzazione e di ottenere una lista dei dottori associati a tale specializzazione.

La pipeline esegue le seguenti operazioni:

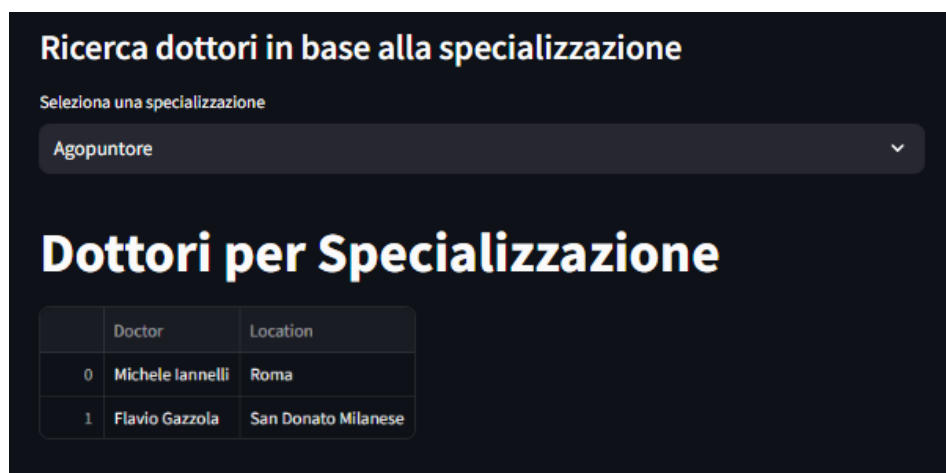
- **\$match**: Filtra i documenti che hanno Specialization e Doctor come stringhe valide.
- **\$project**: Divide le specializzazioni multiple in singole voci utilizzando \$split.
- **\$unwind**: Espande le specializzazioni in singole voci.
- **\$group**: Raggruppa i documenti per specializzazione unica e raccoglie i dottori e le loro location utilizzando \$addToSet per evitare i duplicati.
- **\$sort**: Ordina le specializzazioni in ordine alfabetico.

Il menù è stato creato mediante *"st.selectbox"*



The screenshot shows a dark-themed web interface. At the top, there's a section titled 'Filtro di ricerca'. Below it, a label 'Scegli filtro di ricerca:' is followed by a dropdown menu currently showing 'Ricerca dottori per specializzazioni'. Below this is another section titled 'Ricerca dottori in base alla specializzazione'. Underneath, a label 'Seleziona una specializzazione' is followed by a dropdown menu. This menu is open, showing a list of medical specializations: 'Agopuntore', 'Algologo', 'Allergologia e immunologia clinica', 'Allergologo', 'Anatomia patologica', 'Anatomopatologo', 'Andrologo', and 'Anatropia e fisiologia'. The 'Agopuntore' option is highlighted with a red border.

Figura 11: Elenco specializzazioni mediche



The screenshot shows the same dark-themed web interface. The dropdown menu for 'Seleziona una specializzazione' is now closed and shows 'Agopuntore'. Below this, the section title has changed to 'Dottori per Specializzazione'. Underneath, there is a table with two columns: 'Doctor' and 'Location'. The table contains two rows of data.

	Doctor	Location
0	Michele Iannelli	Roma
1	Flavio Gazzola	San Donato Milanese

Figura 12: Filtro di ricerca

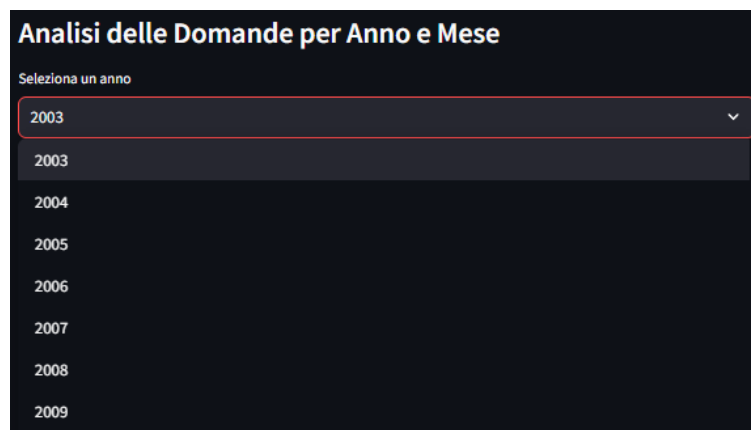
## Analisi delle domande per Anno e Mese

Questa funzione serve per analizzare il numero di domande fatte in un sistema, raggruppandole per anno e mese. Fornisce una panoramica delle tendenze temporali nel numero di domande inviate, permettendo agli utenti di visualizzare il conteggio delle domande per ciascun mese di un anno selezionato.

La pipeline esegue le seguenti operazioni:

- **\$match**: Filtra i documenti per selezionare solo quelli che hanno il campo "Question Date" come una stringa. Questo assicura che vengano considerati solo i documenti che contengono una data di domanda valida.
- **\$project**: Utilizza il comando \$substr per estrarre due componenti dalla data della domanda:
  1. **year**: Estrae i primi 4 caratteri della data (l'anno).
  2. **month**: Estrae i caratteri dal 5° al 6° della data (il mese). Questo passo serve a separare l'anno e il mese dalla data completa per consentire un successivo raggruppamento.
- **\$group**: Raggruppa i documenti in base all'anno e al mese estratti nel passo precedente. Per ogni gruppo, conta il numero di domande utilizzando l'operatore \$sum: 1, che somma per ogni documento (domanda) all'interno di ogni gruppo.
- **\$sort**: Ordina i risultati prima per anno (\_id.year) e poi per mese (\_id.month) in ordine crescente. Questo garantisce che i risultati siano visualizzati in ordine cronologico, dal primo anno/mese al più recente.

Di seguito il box per la selezione dell'anno



The image shows a dark-themed web interface. At the top, the title "Analisi delle Domande per Anno e Mese" is displayed in white. Below the title, there is a label "Seleziona un anno" followed by a dropdown menu. The dropdown menu is currently showing "2003" and has a small downward arrow on the right. Below the dropdown, a list of years from 2003 to 2009 is visible, with 2003 being the selected year.

Figura 13: Box di selezione dell'anno

Esempio di estrazione per l'anno 2021 con tabella e grafico annessi.

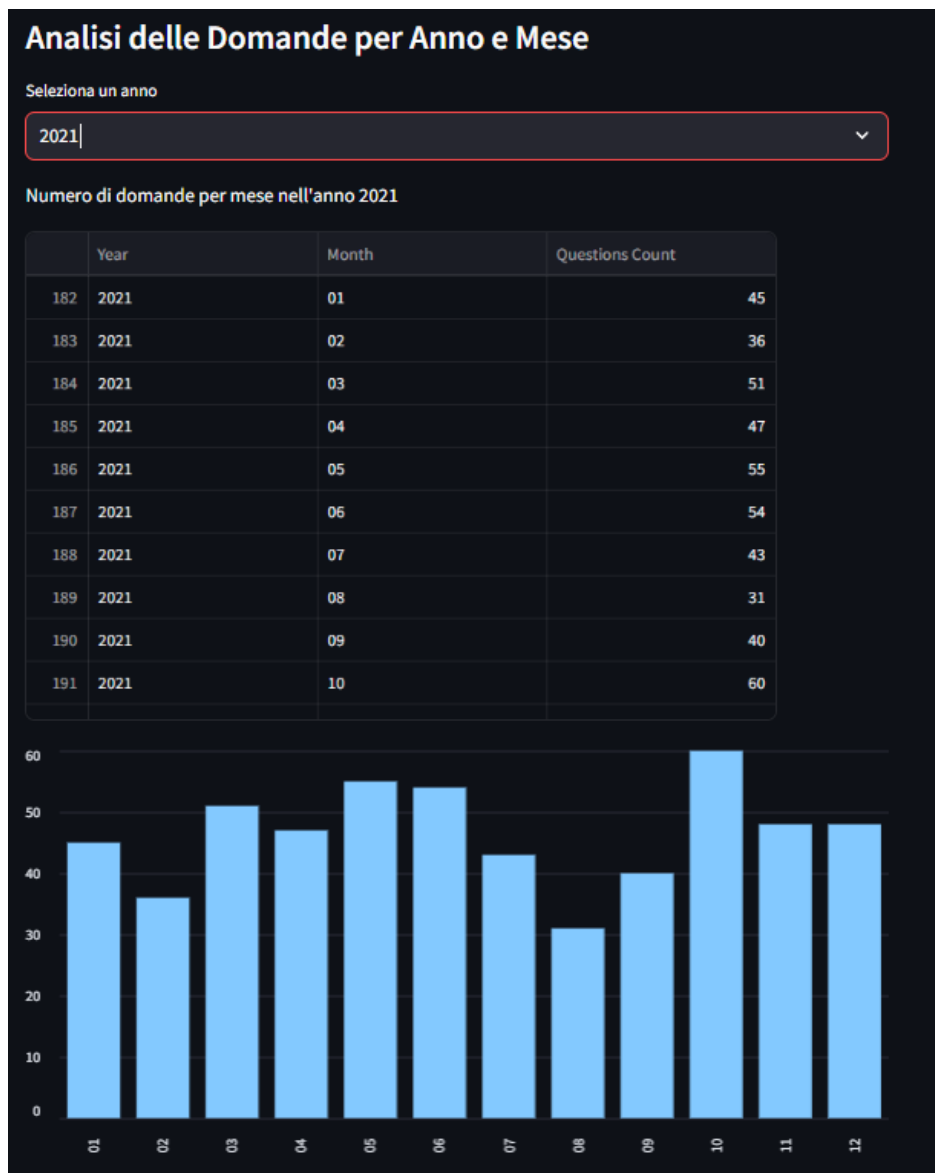


Figura 14: Selezione dell'anno 2021

## 2.4.2 Analitiche

Infine, abbiamo la sezione **Analytics** che comprenderà 3 analitiche esplorabili sempre mediante un apposito menù.

In questa sezione avremo 3 analitiche di riferimento:

- **Distribuzione per specializzazione**
- **Distribuzione dottori per locazione con mappa**
- **Distribuzione per piattaforma**

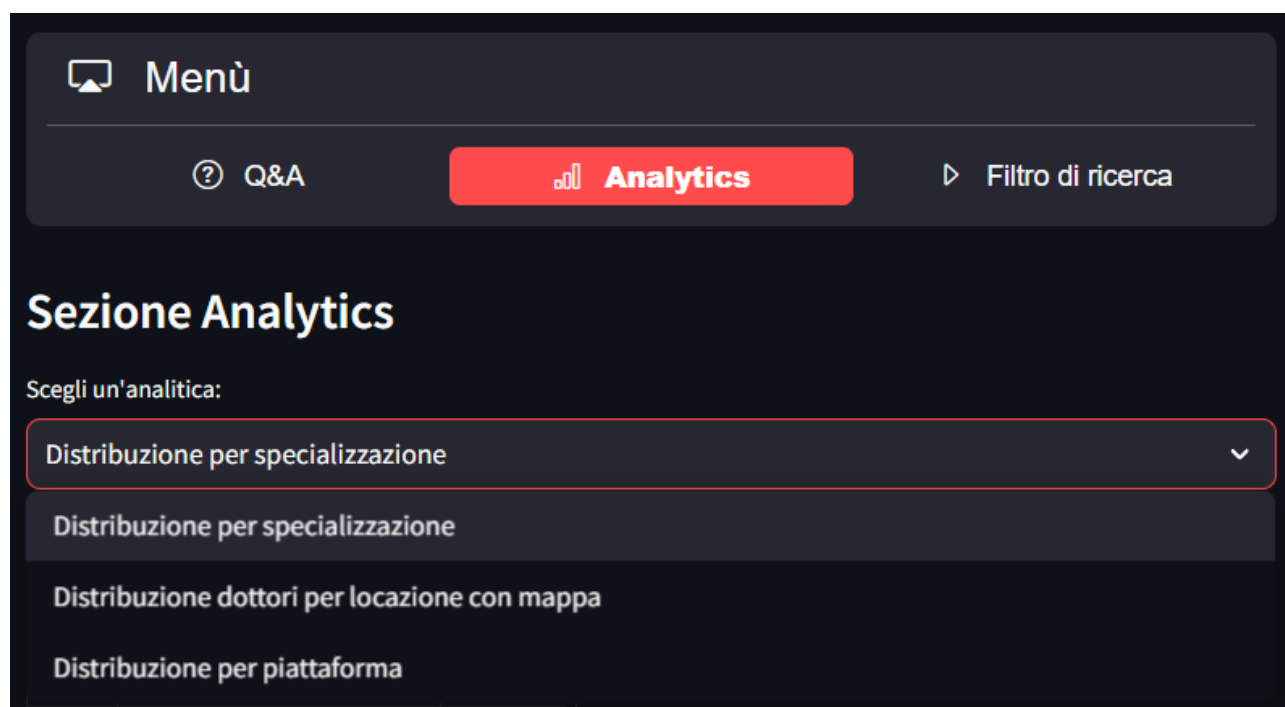


Figura 15: Menù per le analitiche

Di seguito le tre analitiche riportate.



## Distribuzione per Specializzazione

La prima analitica **“Distribuzione per specializzazione”** ci fornisce il numero di dottori per specializzazione.

La pipeline esegue le seguenti operazioni:

- **\$match**: Filtra i documenti che hanno Specialization come stringa valida.
- **\$project**: Crea un nuovo campo chiamato Specialization basato sul valore del campo Specialization esistente. Utilizza l'operatore \$split per dividere il campo Specialization in una lista di specializzazioni separate da virgole.
- **\$unwind**: Espande la lista di specializzazioni in documenti separati, in modo che ogni specializzazione venga trattata come un documento individuale.
- **\$group**: Raggruppa i documenti per Specialization unica e conta il numero di documenti per ciascuna specializzazione utilizzando \$sum: 1.
- **\$sort**: Ordina i risultati in ordine decrescente in base al conteggio dei documenti.

	Specializzazione	Conteggio
0	Gastroenterologo	1,144
1	Psichiatria	1,108
2	Colonproctologo	824
3	Medicina dello sport	775
4	Neurologo	762
5	Chirurgo generale	745
6	Gastroenterologia	724
7	Neurologia	650
8	Chirurgo apparato digerente	624
9	Medicina legale	580

Figura 16: Tabella di output

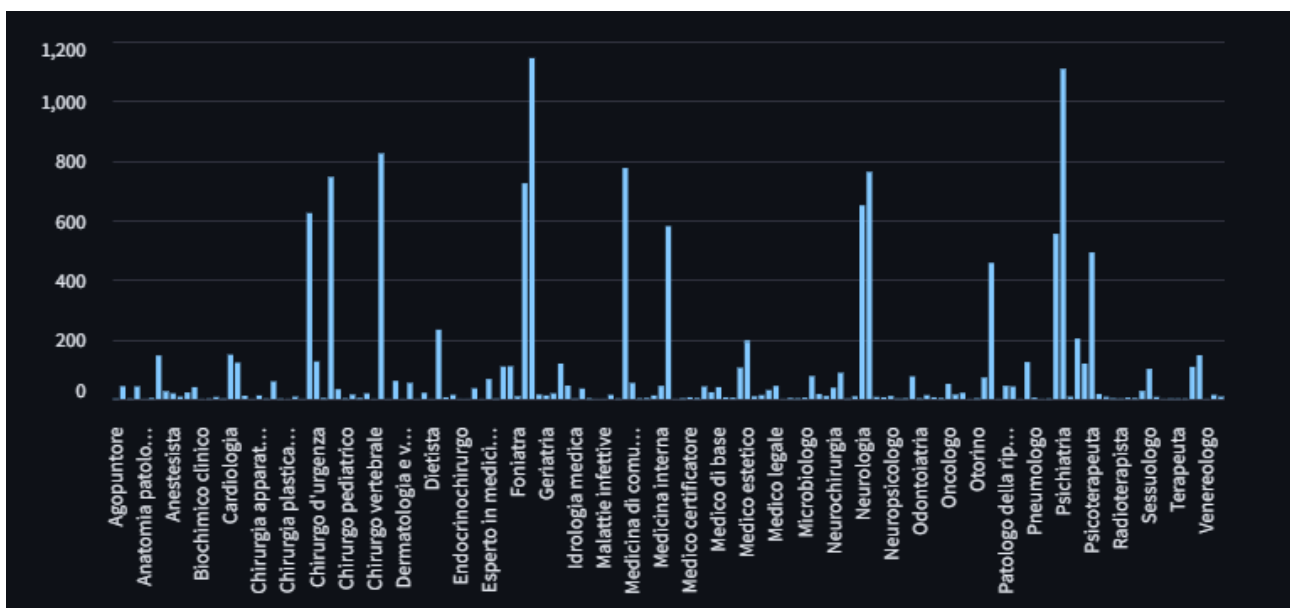


Figura 17: Distribuzione per specializzazione

## Distribuzione dottori per locazione

La seconda analitica **“Distribuzione dottori per locazione con mappa”** ci fornisce il numero di dottori per luogo di riferimento.

La pipeline esegue le seguenti operazioni:

- **\$match**: Filtra i documenti per assicurarsi che il campo Location non sia null e sia di tipo stringa. Questo esclude i valori nulli e considera solo i valori di tipo stringa, escludendo eventuali valori NaN.
- **\$group**: Raggruppa i documenti per Location unica e conta il numero di documenti per ciascuna location utilizzando \$sum: 1. Il campo \_id rappresenta il valore del campo Location, e il campo doctorCount contiene il conteggio dei documenti per ciascuna location.
- **\$sort**: Ordina i risultati in ordine decrescente in base al campo doctorCount, che rappresenta il numero di documenti per ciascuna location.

	Città	Numero Medici
0	torino	1,263
1	roma	1,007
2	milano	930
3	arsizio	784
4	saronno	119
5	firenze	80
6	bologna	79
7	asti	75
8	verona	65
9	padova	61

Figura 18: Tabella di output

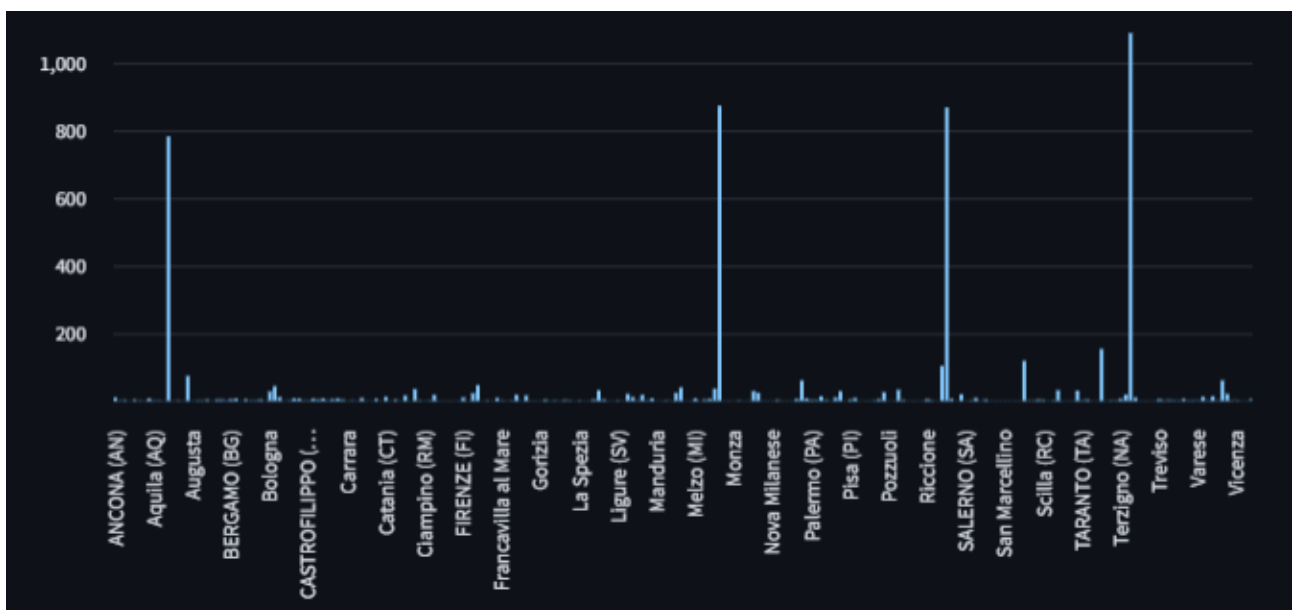


Figura 19: Distribuzione dei dottori per locazione

Si è utilizzato il servizio **API Ninja Geocoder** per convertire i nomi delle città in coordinate geografiche (latitudine e longitudine). Per ottimizzare il processo ed evitare richieste ripetitive al server, i dati ottenuti sono stati salvati in un file .xlsx, garantendo un accesso rapido e riducendo la necessità di ulteriori chiamate API. È importante sottolineare che l'API Ninja impone un limite di utilizzo gratuito, oltre il quale è necessario sottoscrivere un piano a pagamento per continuare ad effettuare richieste.

Per migliorare l'esperienza utente è stato aggiunto un selettore per aumentare o diminuire il numero di marker a schermo.

Cliccando sui marker è possibile vedere il numero di dottori associati alla località indicata.



Figura 20: Vista della mappa con selettore a 186

## Distribuzione per piattaforma

Infine, la terza analitica **“Distribuzione per piattaforma”** fornisce il numero di piattaforme presenti nel database, quindi in sintesi possiamo avere un’idea di come sono state estratte le informazioni dalla fase di *scraping* notando che buona parte sono date da Dica33 e Medicitalia.

La pipeline esegue le seguenti operazioni:

- **\$project**: crea un nuovo campo chiamato platform basato sul valore del campo URL. Utilizza una struttura condizionale (if-then-else) per determinare il valore del campo platform. Se l'URL contiene "dica33", il campo platform sarà "Dica33". Se l'URL contiene "medicitalia", il campo platform sarà "Medicitalia". Se l'URL contiene "miodottore", il campo platform sarà "MioDottore". Se l'URL non corrisponde a nessuno dei precedenti, il campo platform sarà "Altri".
- **\$group**: Raggruppa i documenti in base al campo platform creato nella fase precedente. Il campo \_id rappresenta il valore del campo platform. Conta il numero di documenti per ciascun valore di platform utilizzando \$sum: 1.
- **\$sort**: Ordina i risultati in ordine decrescente in base al campo count.

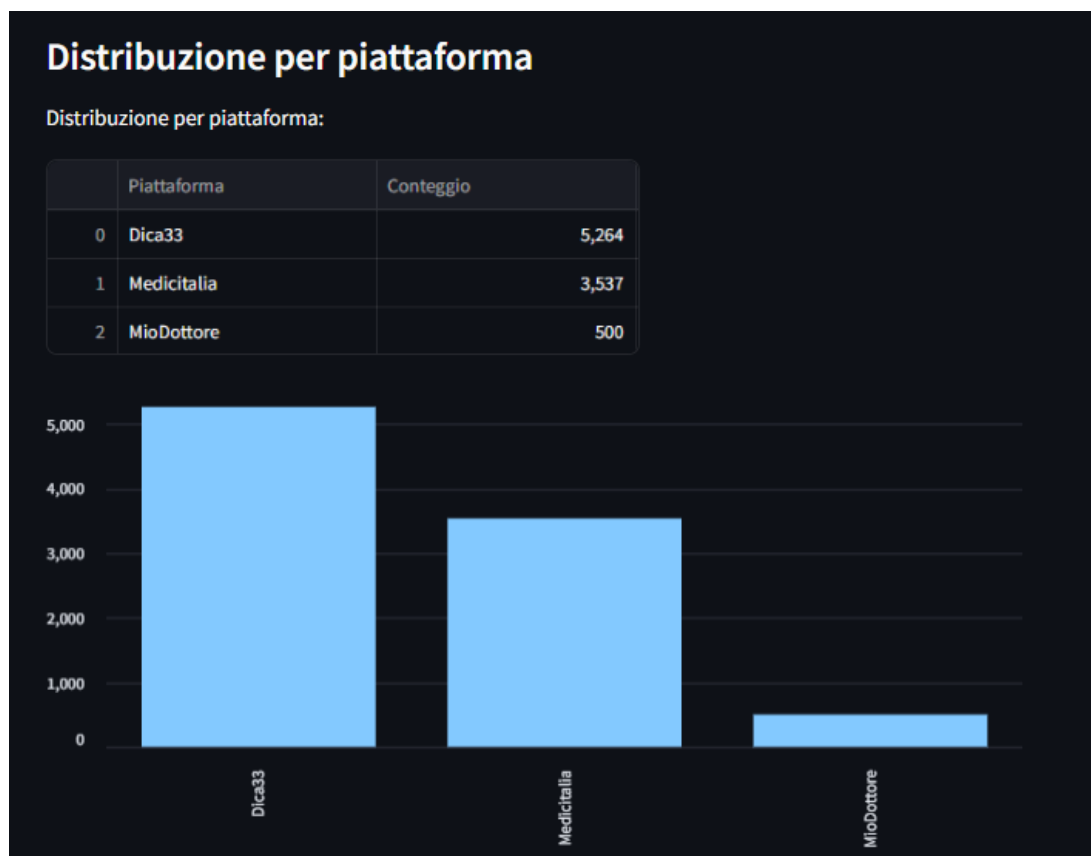
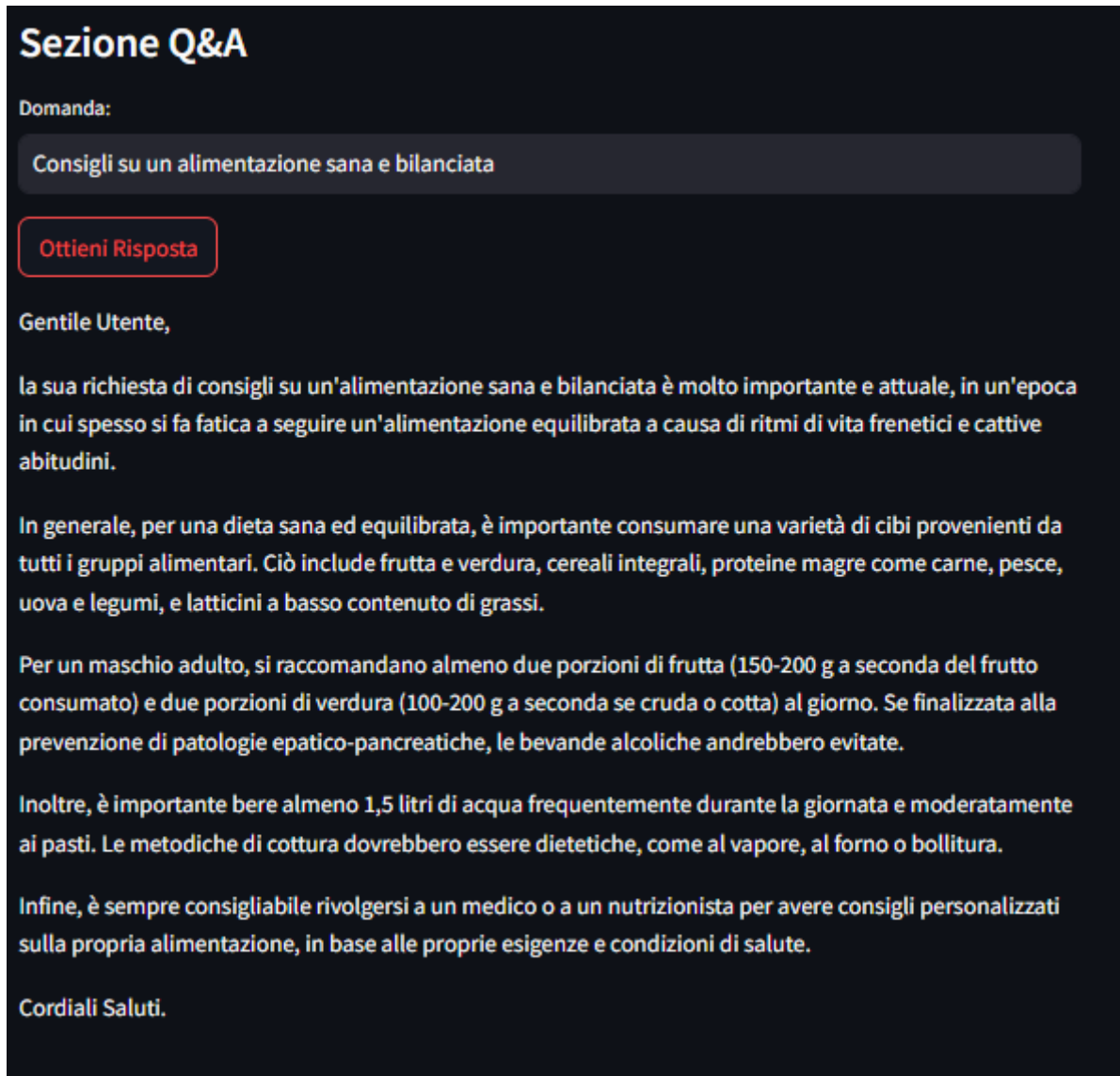


Figura 21: Distribuzione per piattaforma

## 2.5 Esempi di utilizzo della sezione Q&A

Query 1: Consigli su un'alimentazione sana e bilanciata



**Sezione Q&A**

Domanda:

Consigli su un'alimentazione sana e bilanciata

Ottieni Risposta

Gentile Utente,

la sua richiesta di consigli su un'alimentazione sana e bilanciata è molto importante e attuale, in un'epoca in cui spesso si fa fatica a seguire un'alimentazione equilibrata a causa di ritmi di vita frenetici e cattive abitudini.

In generale, per una dieta sana ed equilibrata, è importante consumare una varietà di cibi provenienti da tutti i gruppi alimentari. Ciò include frutta e verdura, cereali integrali, proteine magre come carne, pesce, uova e legumi, e latticini a basso contenuto di grassi.

Per un maschio adulto, si raccomandano almeno due porzioni di frutta (150-200 g a seconda del frutto consumato) e due porzioni di verdura (100-200 g a seconda se cruda o cotta) al giorno. Se finalizzata alla prevenzione di patologie epatico-pancreatiche, le bevande alcoliche andrebbero evitate.

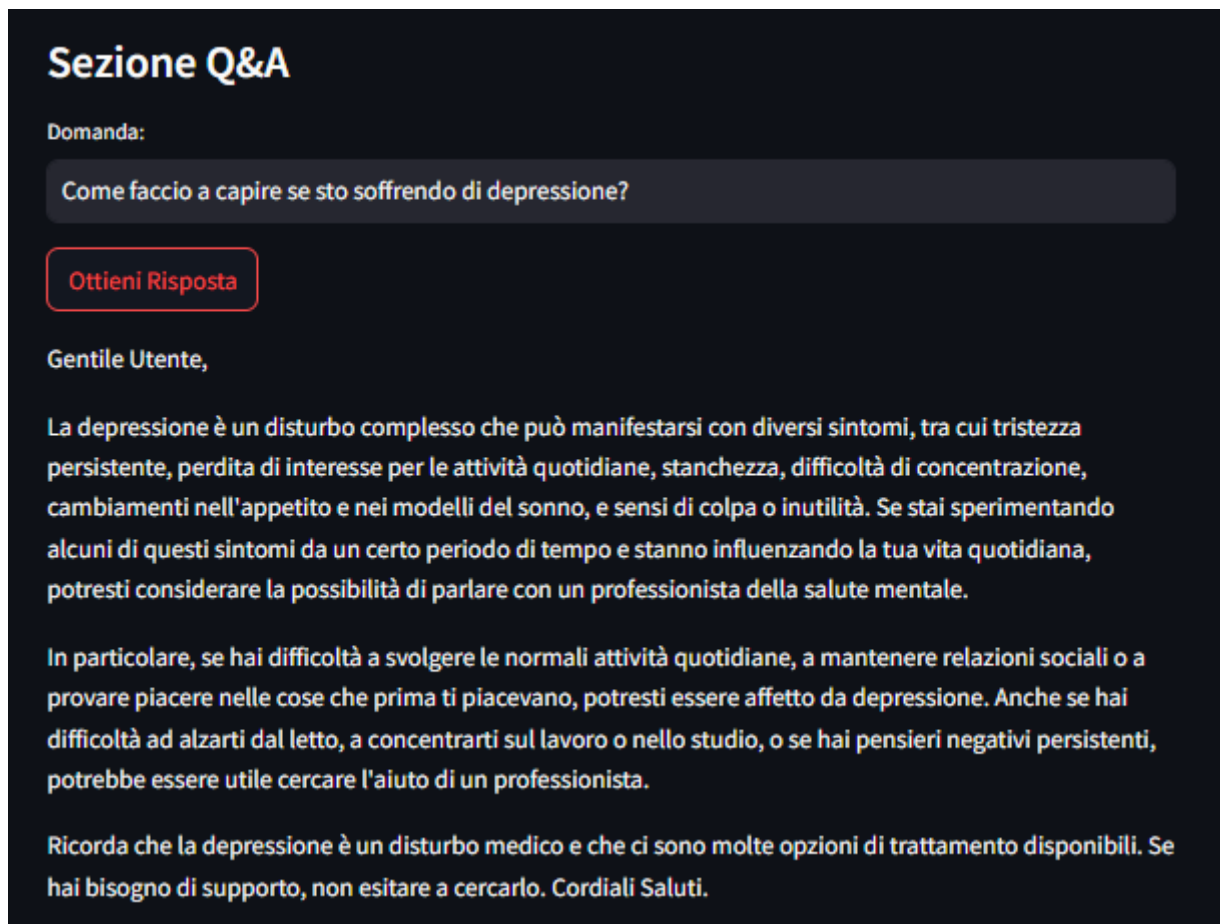
Inoltre, è importante bere almeno 1,5 litri di acqua frequentemente durante la giornata e moderatamente ai pasti. Le metodiche di cottura dovrebbero essere dietetiche, come al vapore, al forno o bollitura.

Infine, è sempre consigliabile rivolgersi a un medico o a un nutrizionista per avere consigli personalizzati sulla propria alimentazione, in base alle proprie esigenze e condizioni di salute.

Cordiali Saluti.

Figura 22: risposta alla prima query

Query 2: Come faccio a capire se sto soffrendo di depressione?



## Sezione Q&A

Domanda:

Come faccio a capire se sto soffrendo di depressione?

Ottieni Risposta

Gentile Utente,

La depressione è un disturbo complesso che può manifestarsi con diversi sintomi, tra cui tristezza persistente, perdita di interesse per le attività quotidiane, stanchezza, difficoltà di concentrazione, cambiamenti nell'appetito e nei modelli del sonno, e sensi di colpa o inutilità. Se stai sperimentando alcuni di questi sintomi da un certo periodo di tempo e stanno influenzando la tua vita quotidiana, potresti considerare la possibilità di parlare con un professionista della salute mentale.

In particolare, se hai difficoltà a svolgere le normali attività quotidiane, a mantenere relazioni sociali o a provare piacere nelle cose che prima ti piacevano, potresti essere affetto da depressione. Anche se hai difficoltà ad alzarti dal letto, a concentrarti sul lavoro o nello studio, o se hai pensieri negativi persistenti, potrebbe essere utile cercare l'aiuto di un professionista.

Ricorda che la depressione è un disturbo medico e che ci sono molte opzioni di trattamento disponibili. Se hai bisogno di supporto, non esitare a cercarlo. Cordiali Saluti.

Figura 23: risposta alla seconda query



Query 3: Vorrei notizie sulla partita nel napoli



**Sezione Q&A**

Domanda:

Vorrei notizie sulla partita del napoli

Ottieni Risposta

Gentile Utente,

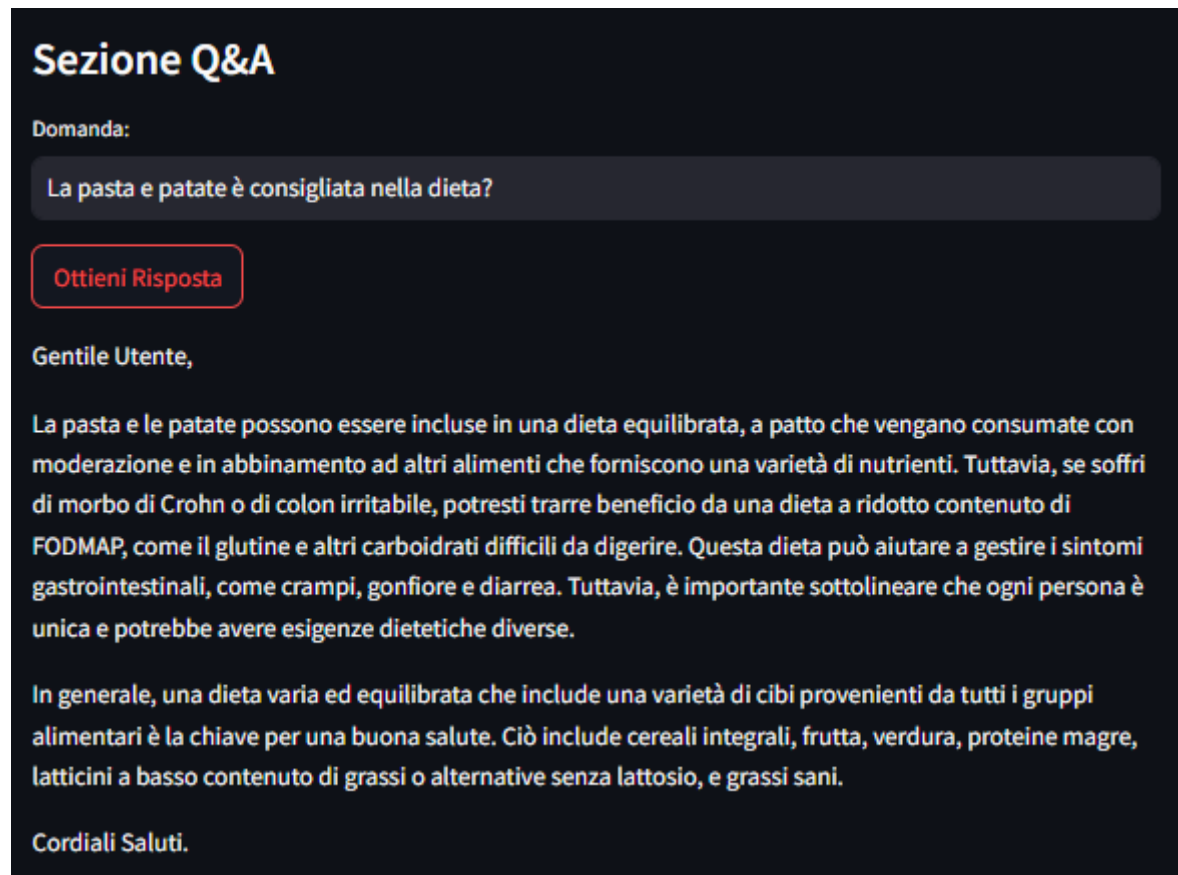
Sono spiacente, ma non posso fornirti notizie sulla partita del Napoli poiché non è un argomento medico. Tuttavia, ti capisco se sei un appassionato di calcio e desideri essere aggiornato sulle ultime partite. Potresti provare a consultare siti web o applicazioni sportive che forniscono risultati e aggiornamenti in tempo reale sulle partite di calcio.

Cordiali Saluti.

*Figura 24: risposta alla terza query*

In questo caso, si è mostrato cosa accade quando la query non riguarda l'ambito medico.

Query 4: La pasta e patate è consigliata nella dieta?



**Sezione Q&A**

Domanda:

La pasta e patate è consigliata nella dieta?

Ottieni Risposta

Gentile Utente,

La pasta e le patate possono essere incluse in una dieta equilibrata, a patto che vengano consumate con moderazione e in abbinamento ad altri alimenti che forniscono una varietà di nutrienti. Tuttavia, se soffri di morbo di Crohn o di colon irritabile, potresti trarre beneficio da una dieta a ridotto contenuto di FODMAP, come il glutine e altri carboidrati difficili da digerire. Questa dieta può aiutare a gestire i sintomi gastrointestinali, come crampi, gonfiore e diarrea. Tuttavia, è importante sottolineare che ogni persona è unica e potrebbe avere esigenze dietetiche diverse.

In generale, una dieta varia ed equilibrata che include una varietà di cibi provenienti da tutti i gruppi alimentari è la chiave per una buona salute. Ciò include cereali integrali, frutta, verdura, proteine magre, latticini a basso contenuto di grassi o alternative senza lattosio, e grassi sani.

Cordiali Saluti.

Figura 25: risposta alla quarta query



## Query 5: Palestra e alimentazione sono correlate?

### Sezione Q&A

Domanda:

Palestra e alimentazione sono correlate?

Ottieni Risposta

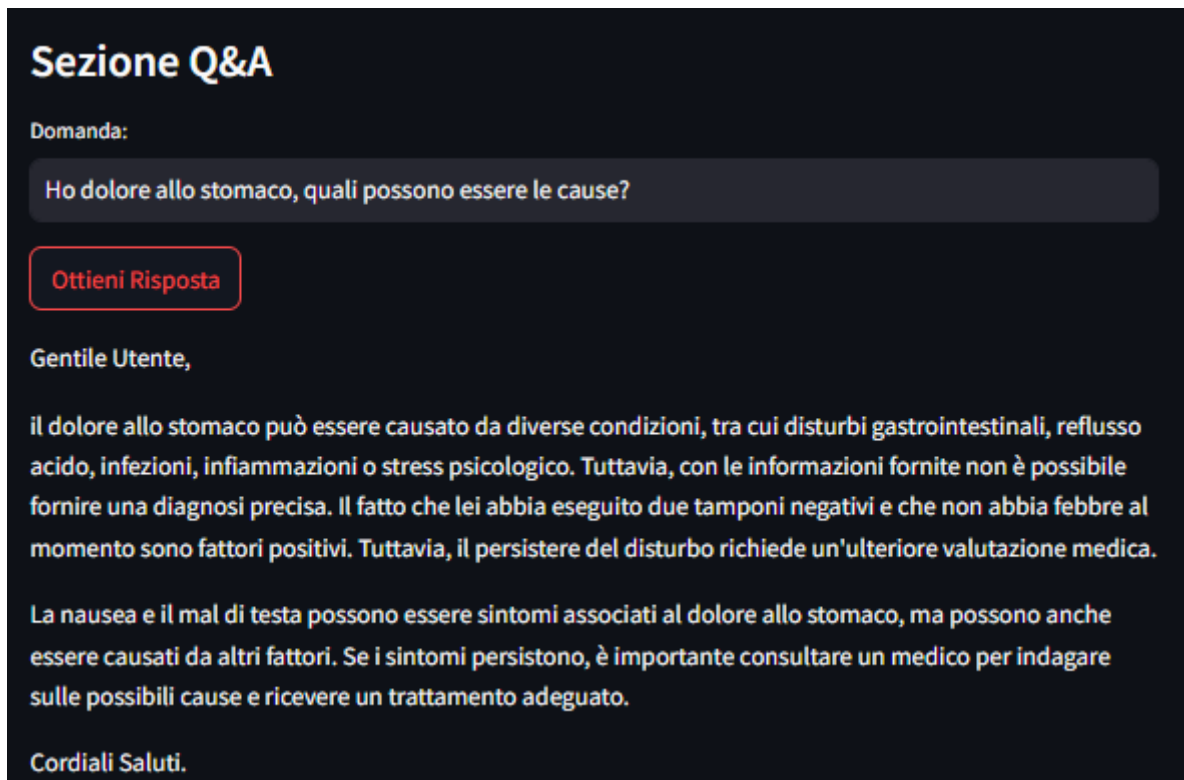
Gentile Utente,

Sì, Palestra e alimentazione sono strettamente correlate. Una corretta alimentazione può influenzare le tue prestazioni in palestra e aiutarti a raggiungere i tuoi obiettivi di fitness. Se ti alleni per aumentare la massa muscolare, ad esempio, avrai bisogno di un apporto adeguato di proteine, mentre se ti stai allenando per la resistenza, dovrai assicurarti di avere abbastanza carboidrati per fornire energia al tuo corpo. Inoltre, una dieta ben bilanciata può aiutarti a recuperare più rapidamente dopo l'allenamento e a prevenire lesioni. Tuttavia, se ti senti male durante l'allenamento, potresti aver esagerato con l'intensità o la durata dell'esercizio. In questo caso, potrebbe essere utile procedere gradualmente e fare attenzione ai segnali del tuo corpo. In alternativa, potresti considerare di svolgere un'altra attività fisica, come camminare o andare in bicicletta, che potrebbe essere meno stressante per il tuo corpo.

Cordiali Saluti.

Figura 26: risposta alla quinta query

Query 6: Ho dolore allo stomaco, quali possono essere le cause?



**Sezione Q&A**

Domanda:

Ho dolore allo stomaco, quali possono essere le cause?

Ottieni Risposta

Gentile Utente,

il dolore allo stomaco può essere causato da diverse condizioni, tra cui disturbi gastrointestinali, reflusso acido, infezioni, infiammazioni o stress psicologico. Tuttavia, con le informazioni fornite non è possibile fornire una diagnosi precisa. Il fatto che lei abbia eseguito due tamponi negativi e che non abbia febbre al momento sono fattori positivi. Tuttavia, il persistere del disturbo richiede un'ulteriore valutazione medica.

La nausea e il mal di testa possono essere sintomi associati al dolore allo stomaco, ma possono anche essere causati da altri fattori. Se i sintomi persistono, è importante consultare un medico per indagare sulle possibili cause e ricevere un trattamento adeguato.

Cordiali Saluti.

Figura 27: risposta alla sesta query

## Query 7: Problemi di pressione cardiaca

### Sezione Q&A

Domanda:

Problemi di pressione cardiaca

Ottieni Risposta

Gentile Utente,

La pressione cardiaca è un valore importante per la salute del nostro cuore. In base alle informazioni che hai fornito, sembra che tu stia attualmente seguendo un trattamento per la pressione alta con il farmaco Ramipril. È importante continuare a monitorare la pressione e il battito cardiaco e mantenere un dialogo costruttivo con il tuo medico riguardo a eventuali cambiamenti o preoccupazioni.

In un caso specifico, hai menzionato di aver avvertito un battito cardiaco accelerato dopo aver misurato la pressione a riposo di 110/70 con una frequenza di 98 battiti al minuto. Non c'è bisogno di preoccuparsi, dato che la pressione e la frequenza cardiaca sono entro i limiti normali. Tuttavia, se dovessi continuare a sentire il cuore che batte velocemente o avvertire altri sintomi preoccupanti, non esitare a consultare il tuo medico.

Ricorda che la pressione cardiaca può essere influenzata da diversi fattori, come l'età, lo stile di vita, l'alimentazione e l'attività fisica. Pertanto, è fondamentale adottare uno stile di vita sano, che includa una dieta equilibrata, esercizio fisico regolare e la riduzione dello stress.

Cordiali Saluti,

Figura 28: risposta alla settima query

## Conclusioni

In conclusione, il sistema di Q&A sviluppato ha mostrato un buon potenziale nel rispondere alle domande mediche utilizzando il framework RAG e un modello di linguaggio. Tuttavia, ci sono ancora delle limitazioni, come la presenza di allucinazioni nelle risposte. Questo potrebbe essere causato da un dataset non troppo esteso e dalla formulazione del prompt, che potrebbe essere migliorata per ottimizzare la qualità delle risposte.