Relazione Homework 2

Petrolito Guido - Matricola 629688

1. URL del progetto

Il codice lo trovate su GitHub:
https://github.com/ArthexTheKing/Ingegneria-dei-Dati/tree/master/Homework%202/Petrolito%20Guido

2. Descrizione del progetto

L'obiettivo del progetto è creare un sistema di indicizzazione e ricerca di file .txt in locale utilizzando Elasticsearch. Il sistema è composto da due script principali:

- **indice.py**: gestisce l'avvio del container Docker con Elasticsearch, crea l'indice e indicizza tutti i file .txt presenti nella cartella data.
- **ricerca.py**: interroga l'indice Elasticsearch e permette ricerche tramite query su nome o contenuto dei file. Include anche il lancio automatico di indice.py all'avvio, così da garantire la presenza dell'indice aggiornato.

3. Analyzer utilizzati

Per l'indicizzazione ho scelto di distinguere due campi:

1. **nome**
   - **Tipo**: text con un sotto-campo raw di tipo keyword.
   - **Analyzer**: italian_analyzer (standard analyzer + stopwords italiane).

   Perché: così posso fare sia ricerche esatte (tipo "cardiologia.txt") sia parziali (tipo "cardio") grazie al campo raw e alle wildcard.
2. **contenuto**
   - **Tipo**: text.
   - **Analyzer**: stesso italian_analyzer.

   Perché: voglio che le ricerche sul testo siano più intelligenti, ignorando parole comuni in italiano ("di", "il", "la", ecc.) e ottenendo risultati più rilevanti.

4. Numero di file indicizzati e tempi

- Ho indicizzato 6 file di test nella cartella data.
- Il tempo totale per indicizzarli è stato di circa 0,03 secondi.

Nome File	**Contenuto File**
Anatomia.txt	L'anatomia umana studia la struttura degli organi e dei tessuti. Include sistemi come scheletrico, muscolare, nervoso e circolatorio.
Cardiologia.txt	La cardiologia studia organo cuore e le malattie cardiovascolari, come ipertensione, infarto e insufficienza cardiaca.
Epidemiologia.txt	L'epidemiologia si occupa della distribuzione e dei determinanti delle malattie nella popolazione, studiando anche prevenzione e controllo.
Farmacologia.txt	La farmacologia si occupa dello studio dei farmaci, dei loro effetti sul corpo umano e delle interazioni tra molecole.
Neurologia.txt	La neurologia analizza il sistema nervoso centrale e periferico, inclusi organo cervello, midollo spinale e nervi.
NeurologiaCardiologia.txt	La neurologia analizza il sistema nervoso centrale e periferico, inclusi cervello, midollo spinale e nervi. La cardiologia studia organo cuore e le malattie cardiovascolari, come ipertensione, infarto e insufficienza cardiaca.

5. Query di test

Tipo di ricerca	Query	Risultati ottenuti
Nome	nome anatomia	anatomia.txt
Nome	nome cardiologia.txt	cardiologia.txt
Nome	nome neuro	neurologiacardiologia.txt, neurologia.txt
Nome	nome neurologia.txt	neurologia.txt
Nome	nome neurologiacardiologia.txt	neurologiacardiologia.txt
Contenuto	contenuto "sistema nervoso"	neurologiacardiologia.txt, neurologia.txt
Contenuto	contenuto cuore	cardiologia.txt, neurologiacardiologia.txt
Contenuto	contenuto pressione sanguigna	Nessun risultato
Contenuto	contenuto "organo cuore"	cardiologia.txt, neurologiacardiologia.txt
Contenuto	contenuto "organo"	cardiologia.txt, neurologia.txt, neurologiacardiologia.txt