UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI **INGEGNERIA INFORMATICA, MODELLISTICA, ELETTRONICA E SISTEMISTICA**
DIMES
ALTILIA
intelligent automation

RAG INDEXING

Corso di metodi e strumenti per lo sviluppo di progetti

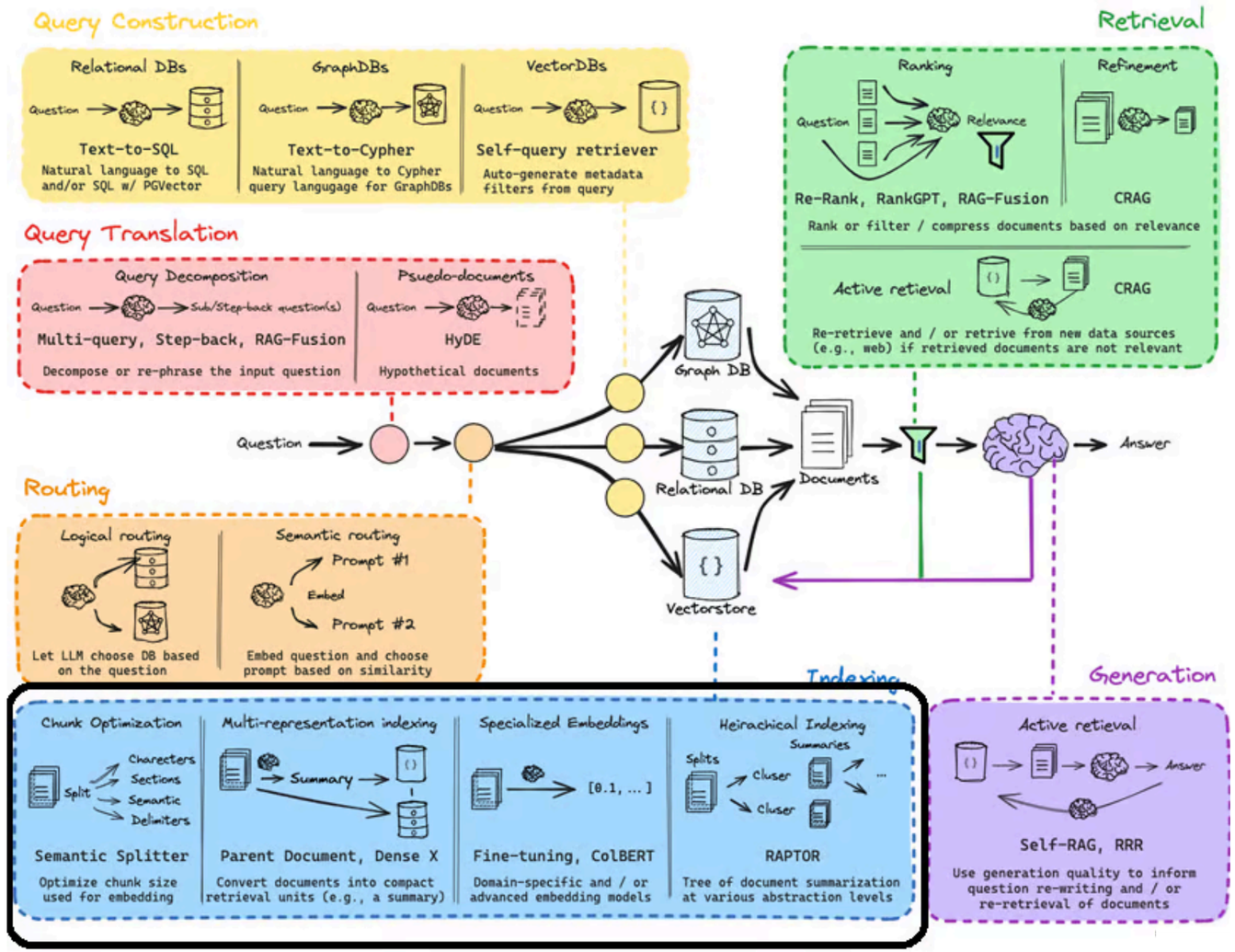
Corso di laurea magistrale in Ingegneria Informatica - Indirizzo Artificial Intelligence & Machine Learning

Gaia Assunta Bertolino
Mat. 242590

Mario Saccomanno
Mat. 248124

Gianluca Ferrari
Mat. 248004

- QUERY CONSTRUCTION
- RETRIEVAL
- QUERY TRANSLATION
- ROUTING
- GENERATION
- **INDEXING**

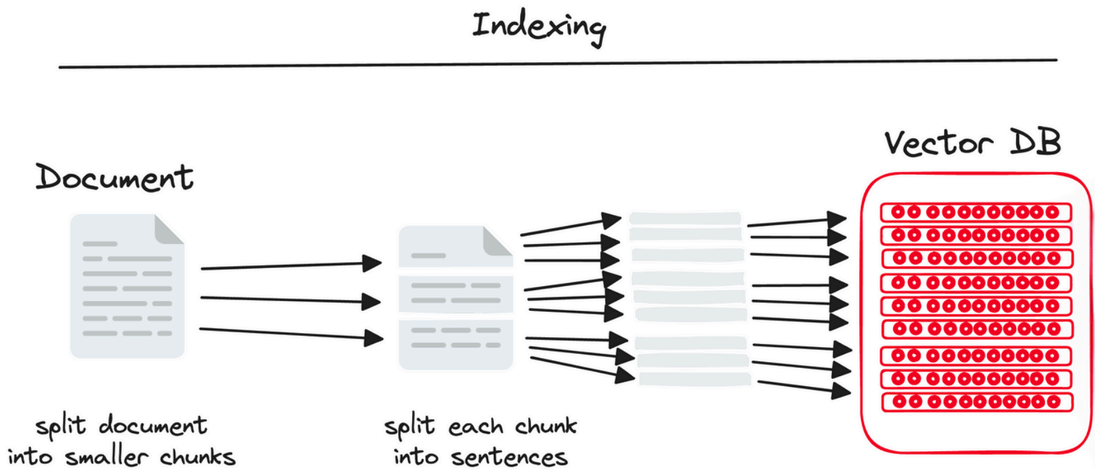


Gaia Assunta Bertolino Mario Saccomanno Gianluca Ferrari

RAG Indexing

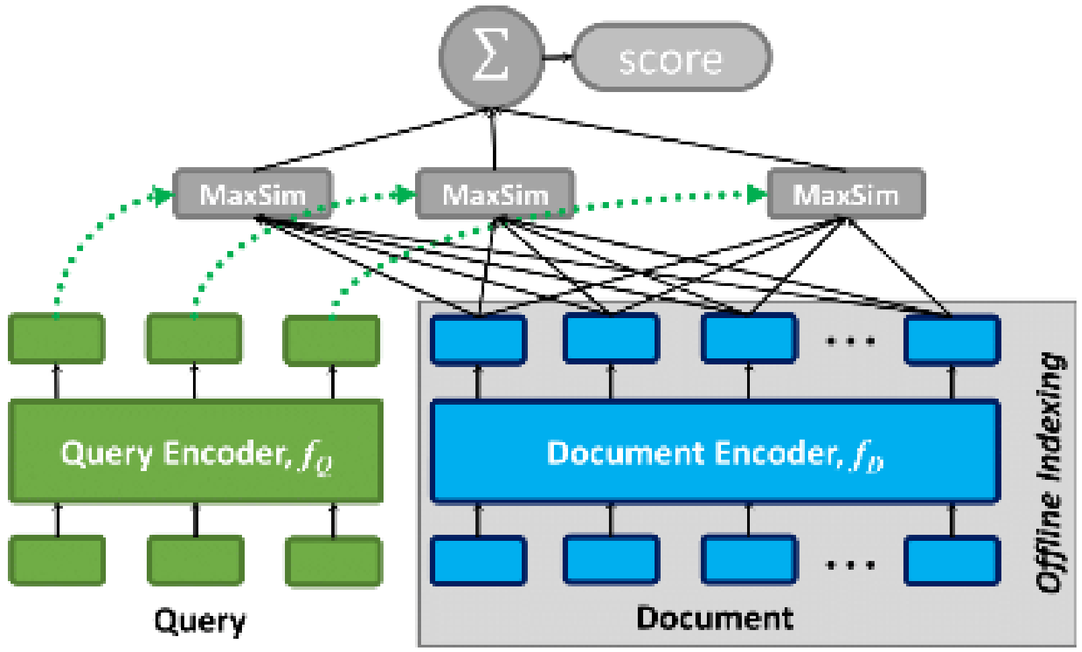
- L'indexing consiste nell'**indicizzare opportunamente** i dati. Tale fase è importante per le operazioni di retrieval

- Parti cruciali dell'indexing sono il chuncking e l'embedding
  - Il **chuncking** serve a trasformare una mole di dati in "pezzi" più piccoli, ponendo attenzione alla qualità e al contesto
-
  - L'**embedding** converte i dati in un insieme di rappresentazioni (vettoriali) che catturano l'essenza dei dati incorporati. Ciò consente il recupero delle informazioni in base alla somiglianza semantica



Specialized Embeddings

- **ColBERT** (Contextual Late Interaction over BERT) è un modello che consente una ricerca scalabile basata su BERT(Bidirection Encoder Representations from Transformers).
- Codifica i dati in una matrice di embeddings. Quando viene eseguita una ricerca, codifica la query dell'utente in un'altra matrice, quindi abbina la query ai pezzi in base al contesto, utilizzando **operatori di similarità vettoriale scalabile** (MaxSim)
- Si parla di **late-stage interaction**: ciò significa che l'interazione tra la query e i documenti avviene in una fase successiva, dopo che i documenti sono stati precomputati



ColBERT Score

- In ColBERT, a differenza della codifica in singolo vettore, ogni token estratto dal testo della query e del documento è **codificato indipendentemente**. Lo score di rilevanza è calcolato come la somma delle massime similarità del coseno tra ogni vettore della query e tutti i vettori nel documento

- Il documento che ottiene lo score di rilevanza più alto per una query ottiene il **ranking** più basso e viceversa

- **CONTRO**: grande mole di dati indicizzati da memorizzare

$$S_{q,d} = \sum_{i=1}^{N} \max_{j=1}^{M} Q_i \cdot D_j^T$$

- Q codifica la query con N vettori
- D codifica il testo con M vettori

L'intuizione di questa architettura è allineare ciascun token della query con il token del **passaggio più rilevante** in modo contestuale, quantificare questi abbinamenti e combinare i punteggi parziali attraverso la query

Santhanam, K., Khattab, O., Saad-Falcon, J., Potts, C., & Zaharia, M. (2021, December 2). ColBERTV2: Effective and efficient retrieval via lightweight late interaction. arXiv.org. https://arxiv.org/abs/2112.01488

Khattab, O., & Zaharia, M. (2020, April 27). ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. arXiv.org. https://arxiv.org/abs/2004.12832

Gaia Assunta Bertolino Mario Saccomanno Gianluca Ferrari

Dataset - document.json


```
[
    {
        "id": "0025577043f5090cd603c6aea60f26e236195594",
        "kind": "movie",
        "text": "<html><title>Pump Up The Volume Transcript</title><pre>\nHappy Harry Hardon - Did you ever get the feeling that everything in America is '
        "summary": " Mark Hunter (Slater), a high school student in a sleepy suburb of Phoenix, Arizona, starts an FM pirate radio station that broadcasts
        "word_count": 11499
    },
    {
        "id": "014de1a8802c05ff64efa047e9290fb7fccea2b4",
        "kind": "gutenberg",
        "text": "\u00ef\u00bb\u00bfThe Project Gutenberg EBook of A Voyage to Arcturus, by David Lindsay\n\nThis eBook is for the use of anyone anywhere at
        "summary": " Maskull, a man longing for adventures, accepts an invitation from Krag, an acquaintance of his friend Nightspore, to travel to Tormanc
        "word_count": 113790
    },
    {
        "id": "019a9611dd8e1b822bd0a58f075cc4a30bdd0797",
        "kind": "gutenberg",
        "text": "\u00ef\u00bb\u00bfThe Project Gutenberg EBook of Lisbeth Longfrock, by Hans Aanrud\n\nThis eBook is for the use of anyone anywhere at no
        "summary": " The story follows its title heroine, from childhood to confirmation. After her mother's death, Lisbeth (given the nickname Longskirt,
        "word_count": 41580
    },
```


- File JSON contente i testi e i relativi summary e metadata
- Gli ids corrispondono alle domande relative al testo contenute nel file *subsampled_golden_pairs.json*

Dataset - subsampled_golden_pairs.json


```
[
    {
        "id": "f7bb9eb9306b79cad4b6466f2ac3dcbd0e5fa63a",
        "question": "Why did Reverend Mother chastise Deloris?",
        "answers": [
            "Deloris sneaked off to a bar.",
            "For drinking in a bar"
        ],
        "kind": "movie"
    },
    {
        "id": "2453d062843edc379bdae3be69859e18bf1abd9d",
        "question": "What does Augustus reveal to Hazel after staying in Amsterdam together?",
        "answers": [
            "That his cancer is terminal",
            "his cancer has relapsed"
        ],
        "kind": "movie"
    },
```


- File JSON contente 83 domande univoche e rispettive risposte
- Contiene le query e answers di riferimento (ground truth)
- Gli ids corrispondono agli indici dei testi a cui le domande fanno riferimento

Librerie e ulteriori modelli

- **LANGCHAIN**: libreria che semplifica la creazione di applicazioni basate su modelli per processare il linguaggio naturale supportando anche l'**integrazione** con piattaforme come OpenAI e Hugging Face e funzionalità per lo storing in Vector Store

- **RAGATOUILLE**: libreria che facilità l'**integrazione** di ColBERT all'interno della **pipeline di IR**

- **MistralAI/Mixtral-8x7B-Instruct-v0.1:** modello di linguaggio **addestrato per generare risposte** informative e coerenti alle domande degli utenti. Mixtral 8x7B è un modello di lingua di tipo **Sparse Mixture of Experts (SMoE)**, una metodologia di modellazione in cui più "esperti" (modelli individuali) sono combinati insieme e un "gating network" decide quale esperto o quali esperti dovrebbero essere utilizzati per fare una previsione per un dato input. La particolarità è che, invece di coinvolgere tutti gli esperti per ogni previsione, utilizza solo un sottoinsieme; da qui il termine "**sparse**". Questo modello, estensione del Mistral 7B, è caratterizzato da strati composti da 8 blocchi feedforward (esperti). Utilizza una rete di instradamento per una lavorazione dinamica, consentendo a ciascun token di accedere a 47B parametri con **solo 13B parametri attivi** durante l'inferenza

Retrieval step

- Tramite Ragatouille si invoca il modello pre-addestrato ColBERT (v2) da 110M parametri
- Le informazioni passate al modello sono divise in tre arrays:
  - array dei documenti: stringhe composte dal testo e dal summary
  - array degli ids: stringa alfa numerica che viene utilizzata per l'identificazione dei testi
  - array dei metadati: contiene solo l'id numerico generato per il documento

es.

```
# Query example
query = "Why did Reverend Mother chastise Deloris?"
RAG = RAGPretrainedModel.from_index(index_path)
results = RAG.search(query)
```

```
[{'content': 'Mary Clarence objects to following the strictures and simple life of the convent, but comes to befriend several of the nu
  'score': 18.953125,
  'rank': 1,
  'document_id': 'f7bb9eb9306b79cad4b6466f2ac3dcbd0e5fa63a',
  'passage_id': 34473,
  'document_metadata': {'source_id: ': 46}},
 {'content': 'Deloris convinces Monsignor O\'Hara that the nuns should be going out to clean up the neighborhood. This they do, and the
  'score': 18.546875,
  'rank': 2,
  'document_id': 'f7bb9eb9306b79cad4b6466f2ac3dcbd0e5fa63a',
  'passage_id': 34474,
  'document_metadata': {'source_id: ': 46}},
 {'content': "After Deloris walks in on Vince having his chauffeur Ernie executed for betrayal, Vince orders his two henchmen Joey and
  'score': 18.328125,
  'rank': 3,
  'document_id': 'f7bb9eb9306b79cad4b6466f2ac3dcbd0e5fa63a',
  'passage_id': 34472,
  'document_metadata': {'source_id: ': 46}},
```

Gaia Assunta Bertolino Mario Saccomanno Gianluca Ferrari

Generation step ed evaluation

- Viene standardizzato l'input (le queries) da sottoporre al modello come prompt

- Per ogni domanda, il modello genera una risposta e viene calcolata una metrica rispetto a una lista di risposte di riferimento (**golden answers**) utilizzando una funzione per il calcolo del BEM score

- Lo **score Bert Matching (BEM)** usa un modello BERT per svolgere un task di equivalenza tra risposte: il task è risolto addestrando un classificatore che determina se due risposte sono equivalenti attraverso il calcolo di un punteggio di equivalenza. Tra i tre tipi di confronto, si è ricorso all'uso delle sole due risposte (**ground truth e candidata**)

Generation step ed evaluation

BEM score without ColBERT indexing: **0.45**

```
df = pd.DataFrame(res)
print(df['BEM score'].mean())
df.head()
```

```
0.5543980582524272
```

	Query	Candidate answer	First gt answer	BEM score
0	Why did Reverend Mother chastise Deloris?	Reverend Mother chastised Deloris for not wea...	Deloris sneaked off to a bar.	0.752
1	What does Augustus reveal to Hazel after stayi...	Augustus reveals to Hazel that he is actually...	That his cancer is terminal	0.467
2	What happen to the hideout spot?	The hideout spot is the room where Chris and ...	It was blown up	0.421
3	How much did Christian pay for the video?	The price Christian paid for the video is not...	One million	0.400
4	Where is the underground shelter where James C...	The underground shelter where James Cole is b...	Under the ruins of Philadelphia.	0.849

BEM score with ColBERT indexing: **0.58**

```
# Computing BEM medium score

bem_scores = res['BEM score']

mean_bem_score = sum(bem_scores) / len(bem_scores)

print(f"BEM score generale: {mean_bem_score}")
```

```
BEM score generale: 0.5814854368932042
```

Gaia Assunta Bertolino Mario Saccomanno Gianluca Ferrari

Conclusioni

CONTRO:

- L'indice richiede un **maggiore spazio di memoria**
- La creazione dell'indice richiede un **maggior tempo di esecuzione**
  - **4 minuti** circa per l'embedding di base vs **9 minuti** circa tramite ColBERT
- Nessuna integrazione con il **Vector Store**

PRO:

- ColBERT migliora di **13 punti percentuali** la metrica di base
- Crea una rappresentazione che **conserva dettagliatamente le informazioni** (180 token per documento) nonostante non si sia ricorso al **pre-processing**
- Permette un **embedding automatizzato** ed intuitivo tramite le librerie a disposizione