Queen Mary
University of London

ColBERT v1 and ColBERT v2: The Ranking Model explained

ECS736 - Information Retrieval - Year 2022/23
Group 2
22 February, 2023

Demonstrators

Ayushi
Choudhury
220762546

Arvind
Jadhav
220906863

Edmund
Lepre
170963873

Mohammed Ataaur
Rahaman
220843052

Agenda

Background

The Problem

- Example Documents
- Example Queries
- Queries and answer passages from LoTTE

The Solution

- Representation-Based Similarity
- Query-Document Identification
- All-To-All Interaction

Motivation

- ColBERT v1
- ColBERT v2

ColBERT Pipeline

ColBERT Architecture

ColBERT Results

Conclusion

Background

1. **ColBERT**: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT

   Paper ID : arXiv:2004.12832

   Date : 27 April 2020

   Authors : Omar Khattab, Matei Zaharia: Stanford University

2. **ColBERTv2**: Effective and Efficient Retrieval via Lightweight Late Interaction

   Paper ID : arXiv:2112.01488

   Date : 2 Dec 2021

   Authors :

   Keshav Santhanam, Omar Khattab, Christopher Potts, Matei Zaharia: Stanford University

   Jon Saad-Falcon: Georgia Institute of Technology

The Problem

Problem:

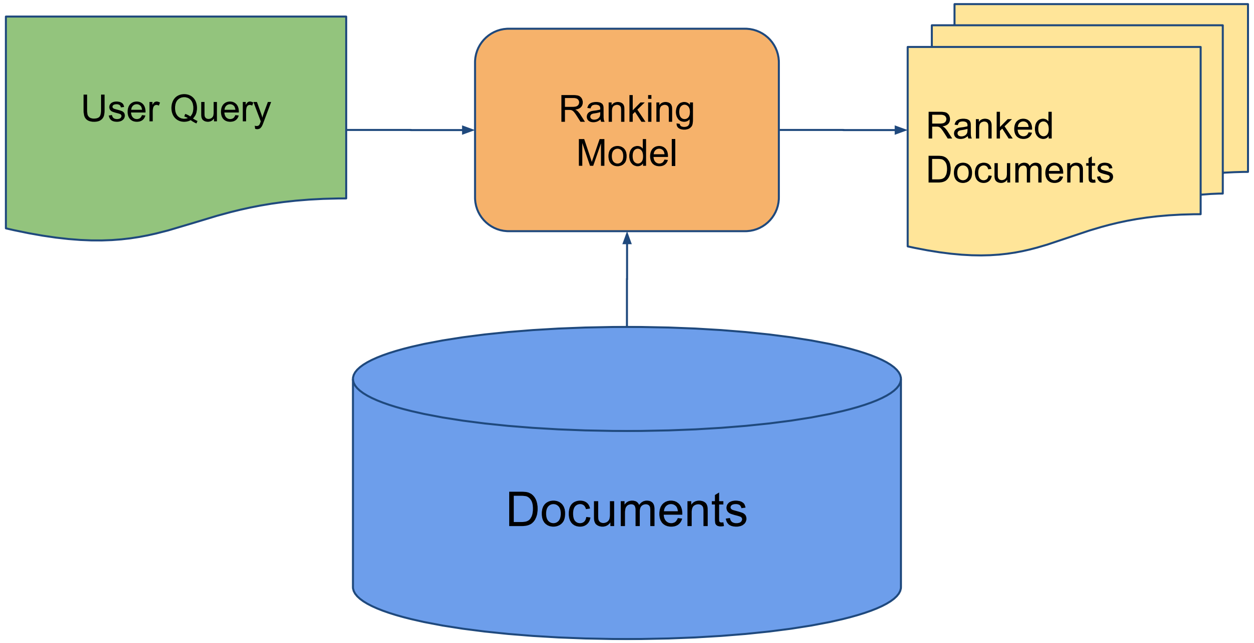
For given documents D, and a given user Query Q, the problem is how to get the top best Documents relevant to the Query.

Solution:

A Ranking model or a retrieval model helps to get the top n Ranked Documents, based on the Query.

Applications:

Search Engines rank pages based on various models and index / rank them.



Example Documents

Science

The Egg Nutrition Center's FAQ page has an entry on this very topic. Basically, the color of the egg does not affect the egg's flavor, nutritional value, etc. It simply depends on the particular breed of chicken that lays the egg -- white eggs from white hens, brown eggs from brown hens. It's also worth noting, as the ENC points out: Generally, brown hens are larger and require more feed and therefore their eggs may be slightly higher priced.

Technology

Go to this link, it has links to many chemical drawing softwares but I will recommend Chemdraw and Chemsketch. Chemdraw is a commercial version but chemsketch is a freeware and it works on both windows and linux operating system. I am currently using Chemsketch and it is very good.

Lifestyle

Close but no cigar. It does dissolve the polymer, but certainly not into its constituent monomers. That would be one heck of a chemical reaction. Make sure you choose the proper gloves for whatever you are working with.

LoTTE (Long-Tail Topic-stratified Evaluation) is introduced for out-of-domain retrievers.

Example Queries

Search queries

Brief, knowledge-based questions with direct answers

Q: what is xerror in rpart?

Q: is sub question one word?

Q: how to open a garage door without making noise?

Q: is docx and dotx the same?

Q: are upvotes and downvotes anonymous?

Q: what is the difference between descriptive essay and narrative essay?

Forum queries

More open-ended questions

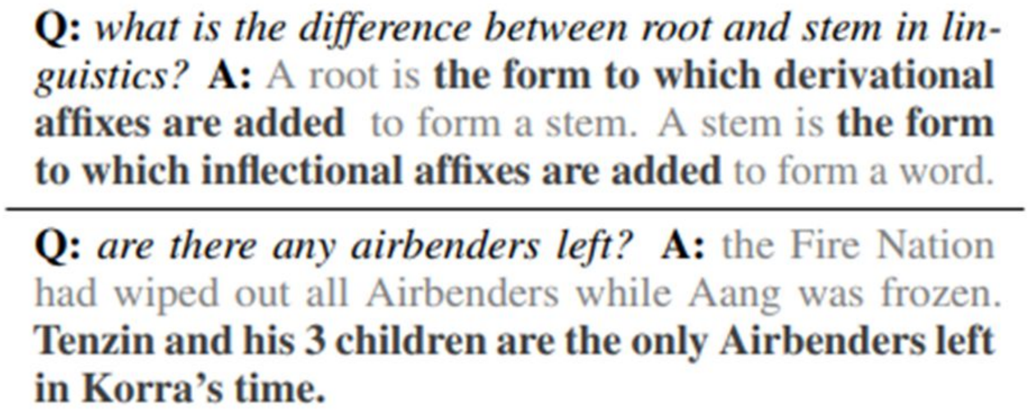
Q: Snoopy can balance on an edge atop his doghouse. Is any reason given for this?

Q: How many Ents were at the Entmoot?

Q: What does a hexagonal sun tell us about the camera lens/sensor?

Q: Should I simply ignore it if authors assume that Im male in their response to my review of their article?

Queries and answer passages from LoTTE

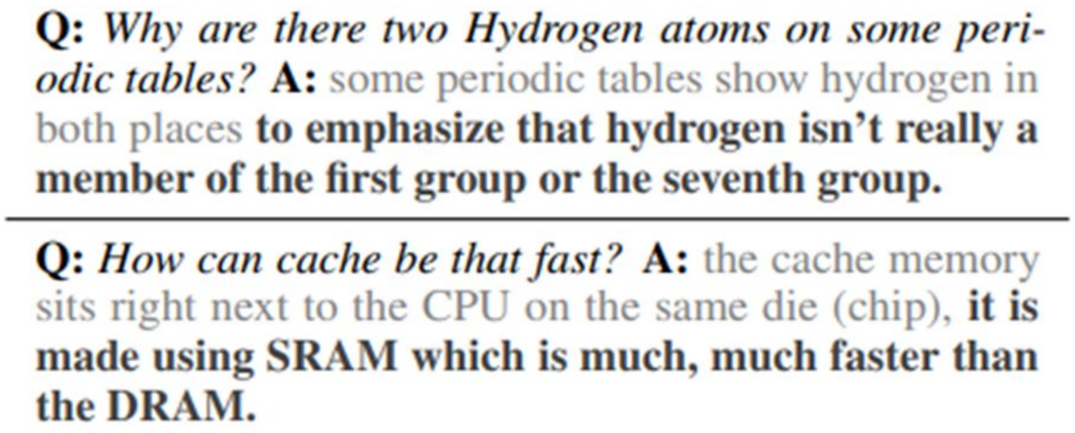


Q: *what is the difference between root and stem in linguistics?* **A:** A root is **the form to which derivational affixes are added** to form a stem. A stem is **the form to which inflectional affixes are added** to form a word.

Q: *are there any airbenders left?* **A:** the Fire Nation had wiped out all Airbenders while Aang was frozen. **Tenzin and his 3 children are the only Airbenders left in Korra's time.**

"search" queries

"forum" queries



Q: *Why are there two Hydrogen atoms on some periodic tables?* **A:** some periodic tables show hydrogen in both places **to emphasize that hydrogen isn't really a member of the first group or the seventh group.**

Q: *How can cache be that fast?* **A:** the cache memory sits right next to the CPU on the same die (chip), **it is made using SRAM which is much, much faster than the DRAM.**

The Solution

Previous Solutions for ranking models :

a. Representation-Based Similarity
b. Query-Document Interaction
c. All-To-All Interaction

Novel Solution:

a. ColBERT v1
b. ColBERT v2

Representation-Based Similarity

BiEncoder : Uses and encodes collections as a single-vector similarity paradigm

Individually creates embeddings for queries and documents, then generates a similarity score between two vectors.

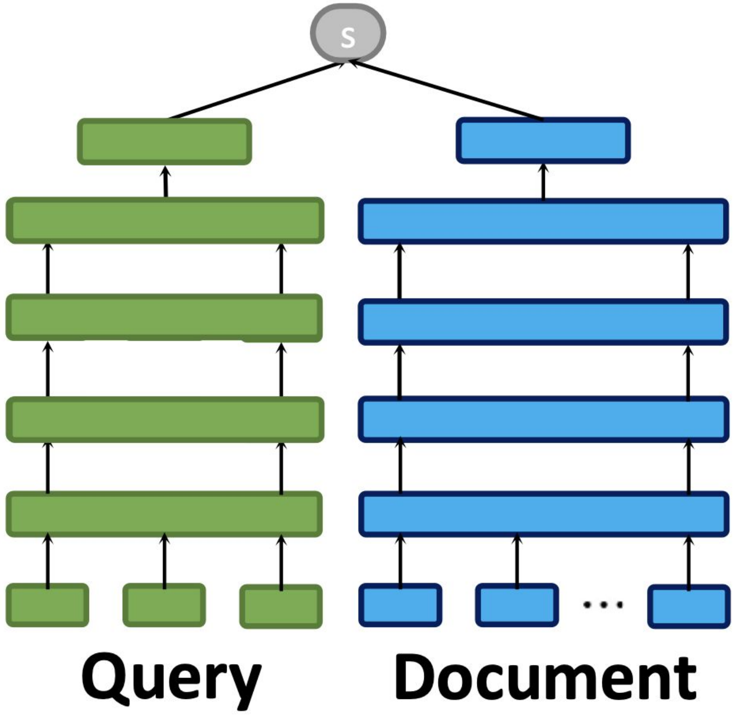
Examples: DSSM, SNRM

Advantages:

- Documents are pre-indexed

Disadvantages:

- Word level and Phrase level relationships within and across the queries and documents are not taken into consideration.So this model does not perform fine-grained matching.



Query-Document Identification

Instead of creating individual embeddings, it models a word and phrase level relationship across query and documents and match them using deep neural network.

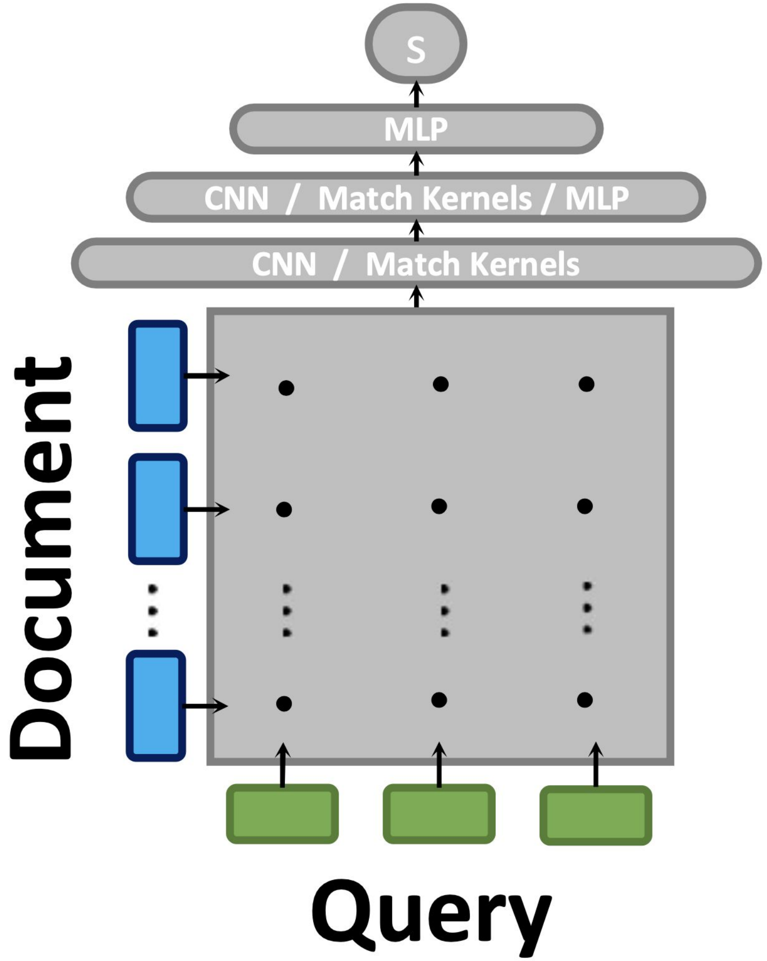
Examples: DRMM, KNRM, Conv-KNRM

Advantages:

- Similarity matrix is generated for interaction between Q and D.

Disadvantages:

- Documents are not pre-indexed increasing computational cost.



All-To-All Interaction

CrossEncoder: It’s a more powerful interaction-based paradigm which models the interaction between words within as well as across the query and documents at the same time.

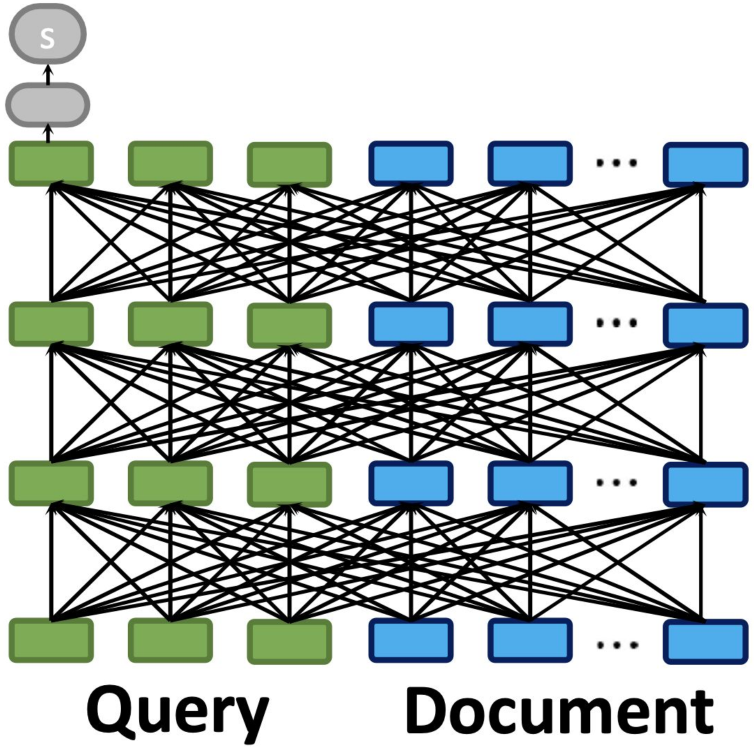
Examples: BERT

Advantages:

- Powerful Interaction-based paradigm: Cross matrix is generated for interaction within and across Q and D.

Disadvantages:

- Documents are not pre-indexed increasing computational cost.

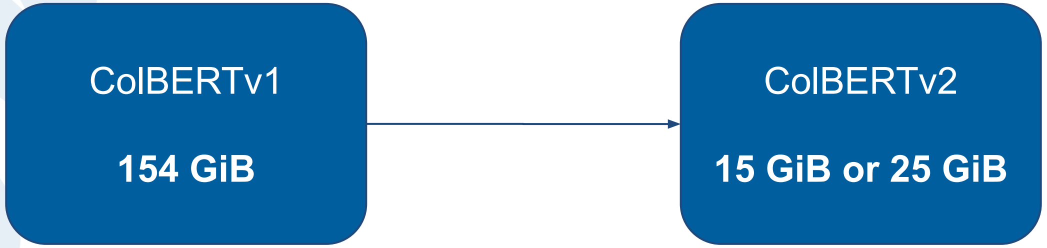


Motivation

Deep Learning language models have high computation cost as they feed each query document pair through a massive neural network to get the document ranks.

To tackle this problem **colBERTv1** was introduced which is a **Multi Vector Retrieval Model**.

ColBERTv2 Improves the quality of the retrieval models while reducing their space footprint by **10x**.


ColBERTv1

154 GiB

ColBERTv2

15 GiB or 25 GiB

ColBERT v1

Novelty : Late Interaction, of Query and Documents.

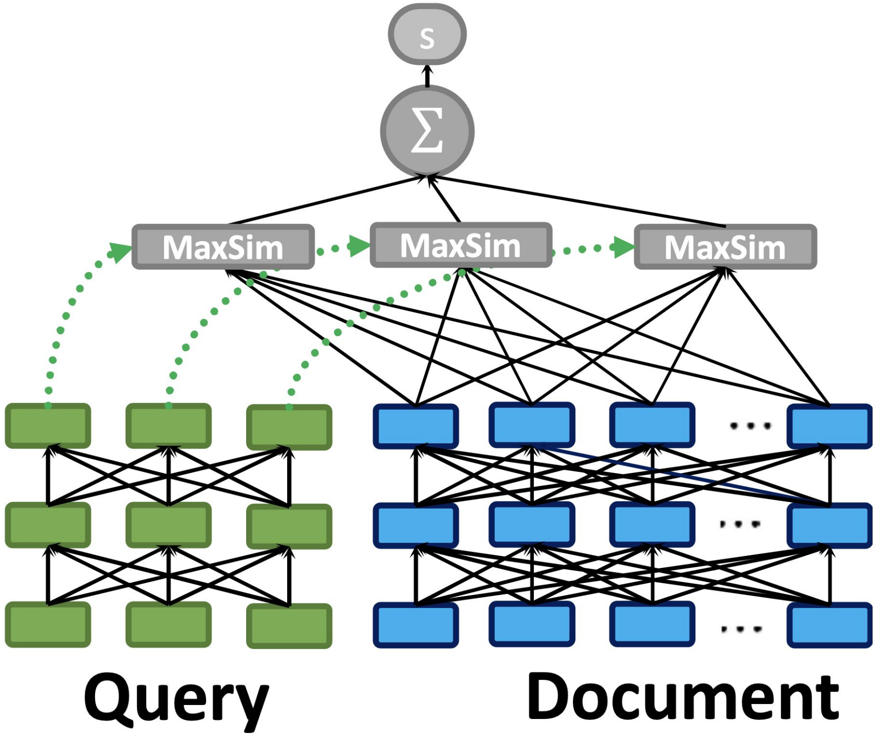
Uses Late interactions, indexing document representations offline, its interaction via MaxSim (Maximum Cosine Similarity) operator and crucial design choice in the BERT based encoder are all essential to colBERT's effectiveness.

Advantages:

- These target attention-based re-ranking
- Documents are pre-indexed
- Q terms are individually matched with D terms.

Disadvantages:

- Very high space complexity (to store D embeddings, pre-indexed Documents)
- Computational time is high due to Q terms and D Terms MaxSim finding.



ColBERT v2

Novelty : Training with Hard negatives, and Document embedding clustering.

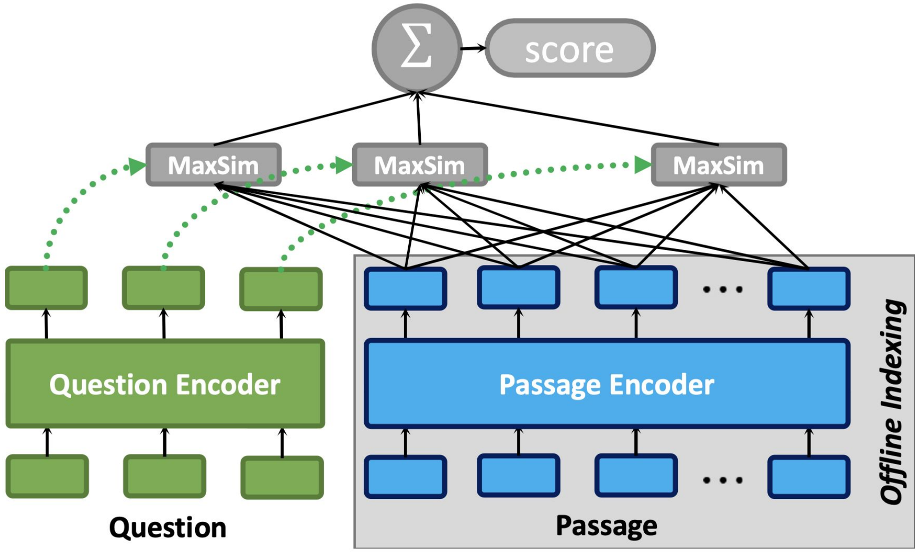
Addresses ColBERTv1's scalable MaxSim end-to-end retrieval

Advantages:

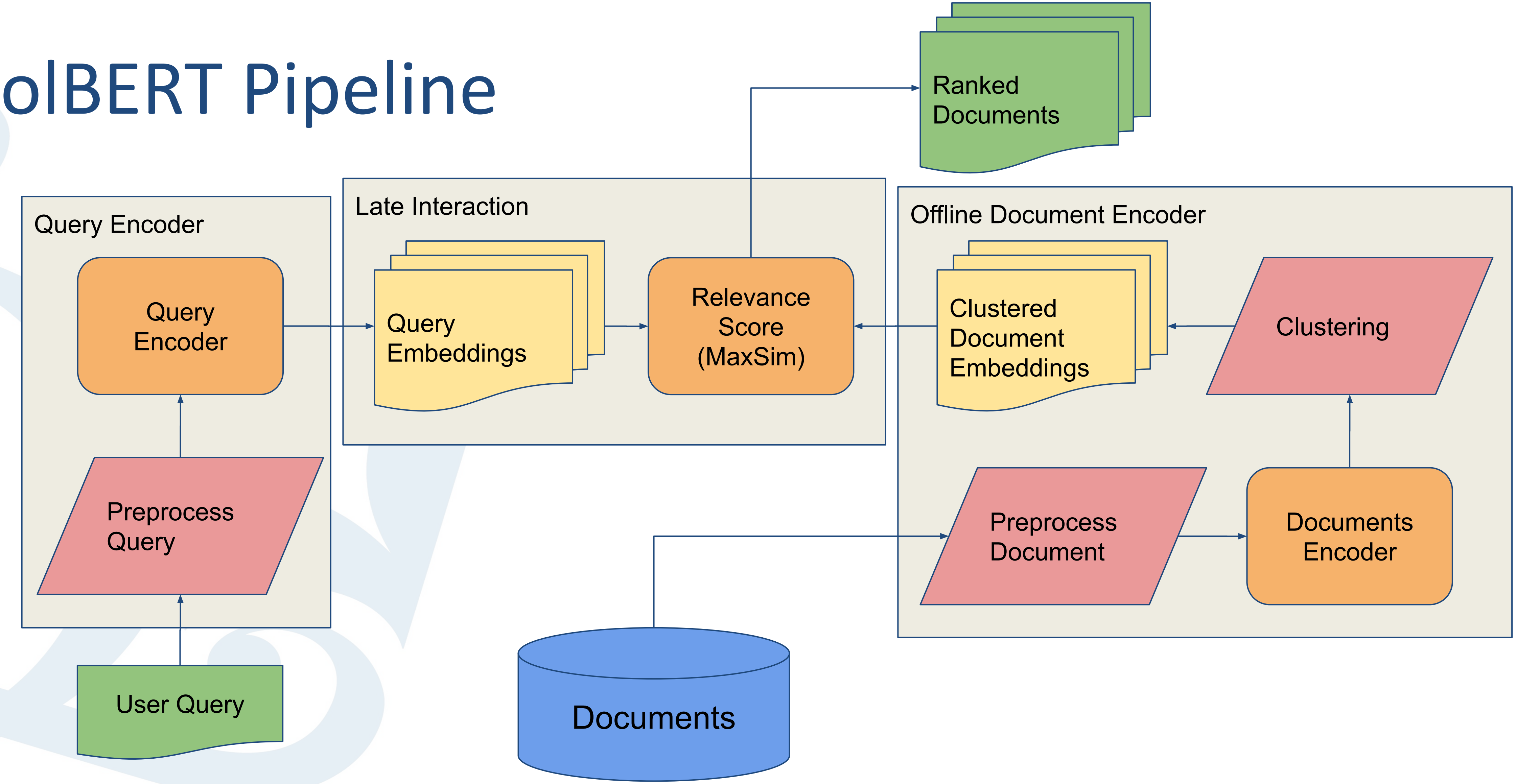
- **Hard Negative Mining**: Train for mining quote and unquote negatives for a model to learn representation for retrieval
- **In-batch Negatives:** Hard (image) negatives and distillation like standard cross-entropy
- **Generating training**: Some combination of in-batch negatives and distillation, as known as, with hard negatives so this alone already improves a lot in performance
- **Lower search cost efficiency:** Storing the document embeddings using clustering and centroids
- **Quality Improvement:** Multi-vector retrieval models while reducing their space footprint by 10x.

Disadvantages:

- **Storage Space:** When loading these documents into memory (cross-encoder)
- **Low-level Systems Optimizations:** Under extreme resource constraints, simpler model designs like **SPLADEv2** or **RocketQAv2** could be used themselves for simple optimization.



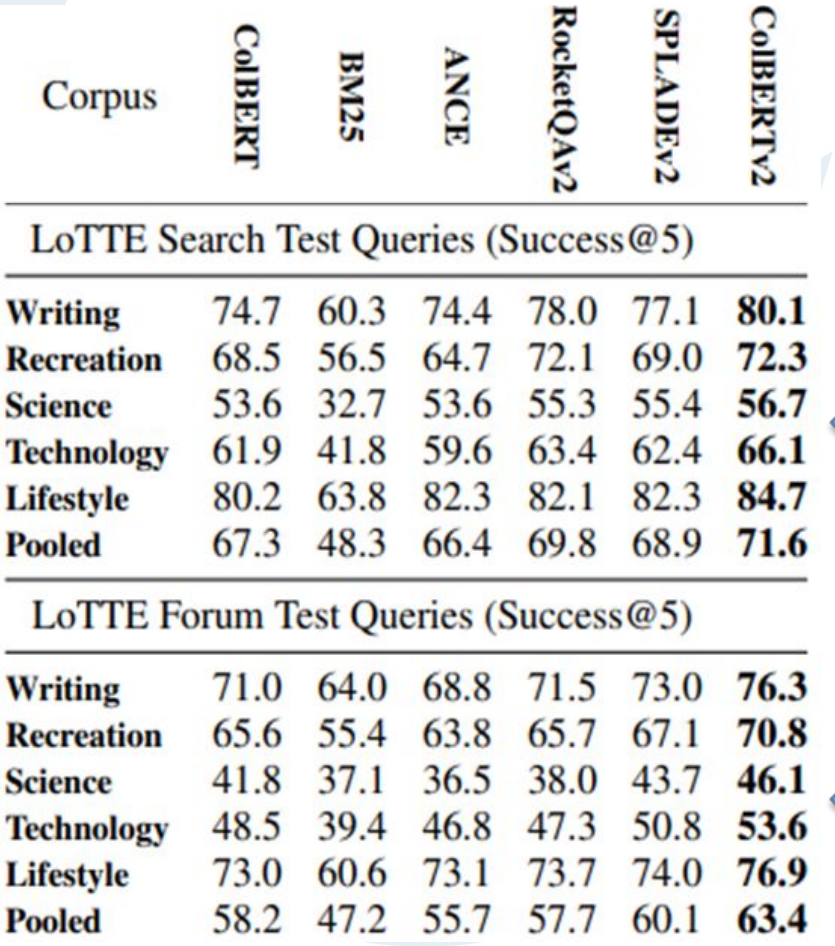
ColBERT Pipeline



ColBERT Results

ColBERTv2 outperforms RocketQAv2, SPLADEv2 (Distillation based Models), and BM25, ANCE (non-Distillation based models on LoTTE Dataset, as seen in the results table here.

Success@5, as known as Recall@5, is the metric that measures the accuracy of the correlation between the percentage of questions and extracting potential short answer strings from overlapping with one or more of the top-5 passages.



Corpus	ColBERT	BM25	ANCE	RocketQAv2	SPLADEv2	ColBERTv2
LoTTE Search Test Queries (Success@5)						
Writing	74.7	60.3	74.4	78.0	77.1	**80.1**
Recreation	68.5	56.5	64.7	72.1	69.0	**72.3**
Science	53.6	32.7	53.6	55.3	55.4	**56.7**
Technology	61.9	41.8	59.6	63.4	62.4	**66.1**
Lifestyle	80.2	63.8	82.3	82.1	82.3	**84.7**
Pooled	67.3	48.3	66.4	69.8	68.9	**71.6**
LoTTE Forum Test Queries (Success@5)						
Writing	71.0	64.0	68.8	71.5	73.0	**76.3**
Recreation	65.6	55.4	63.8	65.7	67.1	**70.8**
Science	41.8	37.1	36.5	38.0	43.7	**46.1**
Technology	48.5	39.4	46.8	47.3	50.8	**53.6**
Lifestyle	73.0	60.6	73.1	73.7	74.0	**76.9**
Pooled	58.2	47.2	55.7	57.7	60.1	**63.4**

Automatic Google rankings through GooAQ

Organic StackExchange question–answer pairs

Conclusions

Our views :

- **Late interaction** helps to capture and learn term embeddings which makes it efficient matching during inference.
- **Efficient** storage of Document Embedding reducing computational cost during inference.
- Faster centroid approach (optimization with fewer centroids).
- When pruning, don't compare against terms from every centroid
- Query to centroid closeness relation to ignore irrelevant document terms, helping reconstruct the terms with centroid plus and vector faster.

obrigado
ank U
Merci
ahalo
köszi
спасибо
Grazie
Thank you
mauruuru
Takk
Gracias
Dziękuję
Děkuju
danke
Kiitos

References

A Comparison of Text Retrieval Models

- https://academic.oup.com/comjnl/article/35/3/279/525681?login=true

Neural ranking models for document retrieval

- https://arxiv.org/abs/2102.11903#:~:text=Ranking%20models%20are%20the%20main,learning%20models%20in%20information%20retrieval.

ColBERTv1: arXiv:2004.12832v2 [cs.IR] 27 April 2020

- https://arxiv.org/abs/2004.12832

ColBERTv2: arXiv:2112.01488v2 [cs.IR] 2 Dec 2021

- https://arxiv.org/abs/2112.01488