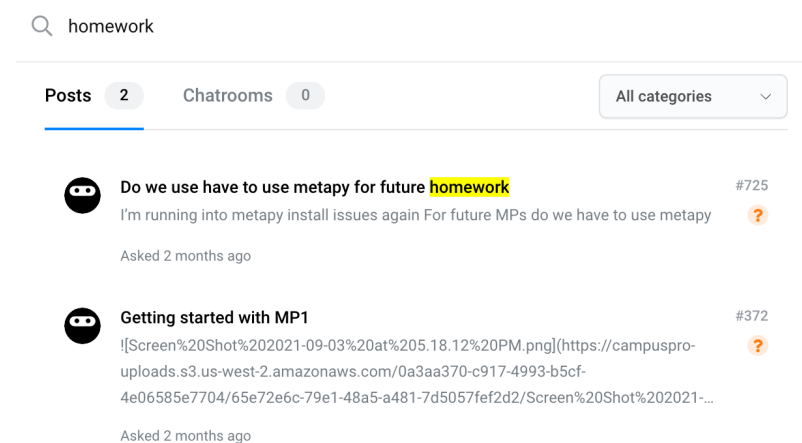


Search for Neural search framework

Intro

This is a capture screen from campuswire. When I search with the word 'homework' then the search result gives the document that has word homework. actually it could have contained 'assignment', 'hw'.



If there's no word that overlaps completely, you can't get it as a search result. And i think this is limitation of

indexed based IR. So i would like to find out the trend of model based neural information retrieval

Simply I compare two approaches between Index based IR and Model based Neural IR.

Index based IR(BM25)	Model based Neural IR(SentenceEmbedding)
<ul style="list-style-type: none">• Search requirements are limited and time is limited• If it is unlikely to change in the past limited domain• When searching for text• More faster than getting the most accurate results if it matters• If we want to understand why we get results when computing resources are limited	<ul style="list-style-type: none">• If search requirements are likely to extend to various languages and data types(ex. Image, text, video..)• Even if you can't fully understand why you got the result, if a plausible result is important• If you don't have much knowledge about search areas• If you want something that works in the black-box

Body


And with some research, i found Jina framework which is an open-source neural search framework that offers the building blocks for designing and implementing neural network-based search applications Jina provides large-scale indexing and queries for various types of data, including video, image, text, music, source code, and PDF, with the Neural Search Framework. Among the examples, I found the use of Cross-Modal Search System, a model that allows me search for images given a caption description

This is the searching result from jina framework

<https://link.ainize.ai/3fCyHdI>

Input was “horses are running through the seaside”

And output really gave me some reasonable images



Jina - Cross Modal Search System

It allows the user to search for images given a caption description.


Github : [Jina - Cross Modal Search System](#)
Open API : [On Ainize](#)

Example

Input


a dog running in the meadow

Output



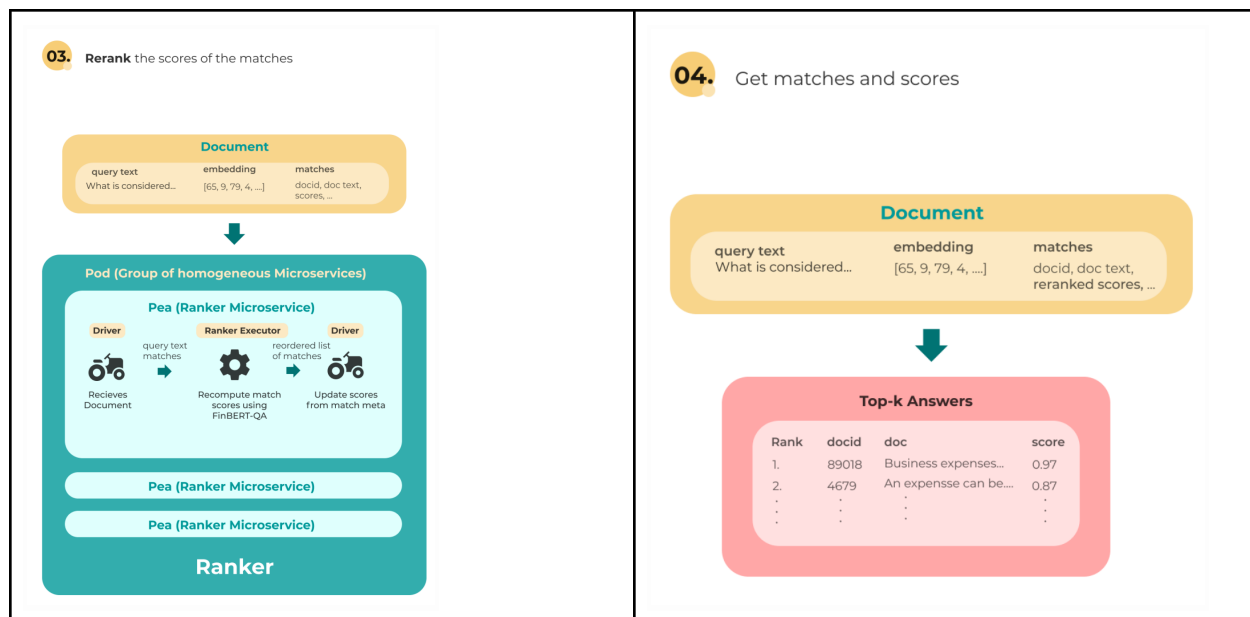
Try it!

10 results in 0.92 seconds



And I thought text retrieval could also apply similarly. After some searching, I found a Question Answering search engine from the github repository, and followed the tutorial from this [repository](#). And it was about BERT-based Financial Question Answering System

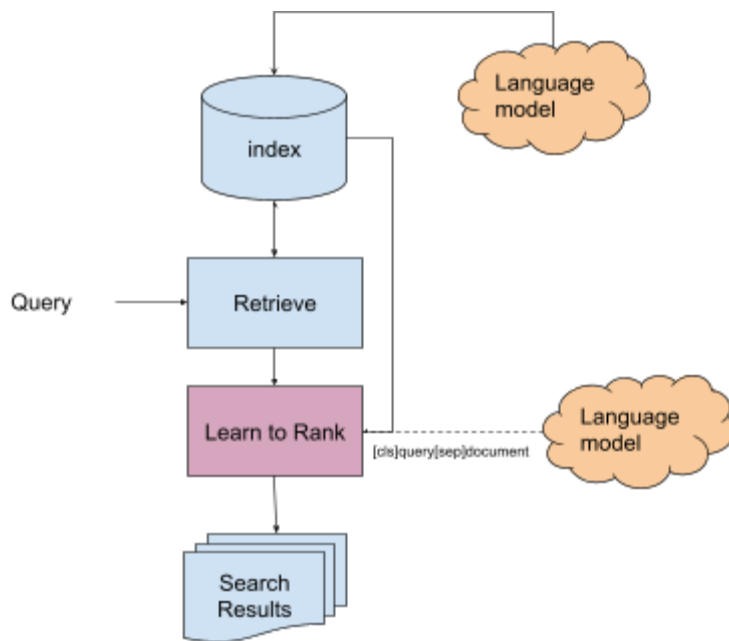
<p><u>Encoder in the query flow</u> (encoder can be done with pre trained language model, they used FINBERT trained by a large financial corpus</p>	<p><u>The Indexer will search for the answers with the most similar embeddings(ex.cosine distance)</u></p>
<div><p>01. Encode the query to an embedding</p><p>Q: "What is considered a business expensive?"</p><div><div>Document</div><div>query text What is considered a business expensive?</div></div><div><div>Pod (Group of homogeneous Microservices)</div><div><div>Pea (Encoder Microservice)</div><div><div>Driver</div><div>text</div><div>Encoder Executor</div><div>embedding</div><div>Driver</div></div><div><div>Receives Document</div><div>Get embedding</div><div>Add embedding to Document</div></div></div><div><div>Pea (Encoder Microservice)</div><div>Pea (Encoder Microservice)</div></div><div>Encoder</div></div></div>	<div><p>02. Find the most similar matches from the Index</p><div><div>Document</div><div>query text What is considered..</div><div>embedding [65, 9, 79, 4, ...]</div></div><div><div>Pod (Group of homogeneous Microservices)</div><div><div>Pea (Indexer Microservice)</div><div><div>Driver</div><div>embedding</div><div>Indexer Executor</div><div>matches</div><div>Driver</div></div><div><div>Receives Document</div><div>Get most similar embeddings (matches) and corresponding meta data (scores, text, ...)</div><div>Add matches to Document</div></div></div><div><div>Pea (Indexer Microservice)</div><div>Pea (Indexer Microservice)</div></div><div>Indexer</div></div></div>
<p>The Ranker recomputes the scores of the matches using FinBERT-QA(Fine-tuned bert-qa on the FiQA dataset)</p>	<p>get matches and scores stored in the Document</p>



Normally the search space of document searching is very large. In order to avoid this problem, They divide it into two steps. The key to this method was to quickly extract the primary candidate results with pre pre-trained language model embedding layer. And rank them with domain specific models..

Conclusion

From this research, I could imagine how to make a text retrieval system with a Neural search engine for a specific domain.



For Ranking model, it would be also good idea to mixture of bm25 ranking model and fine-tuning bert model with nsp(next sentence prediction power)

cite relevant references

JINA github: <https://github.com/jina-ai/jina>

Question Answering example: <https://github.com/yuanbit/jina-financial-qa-search-template>

<https://towardsdatascience.com/how-to-build-a-production-ready-financial-question-answering-system-with-jina-and-bert-48335103043f>

FinBERT : <https://arxiv.org/pdf/1908.10063.pdf>