10/30/2024

Terms – Condition Reader

Vidya, Ranjan, Shivam

SITARE UNIVERSITY

This project implements a RAG (Retrieval Augmented Generation) based question-answering system. Below, we highlight the methods used to develop this system, including efforts to improve accuracy and mitigate hallucination risks.

How Our RAG Pipeline Works

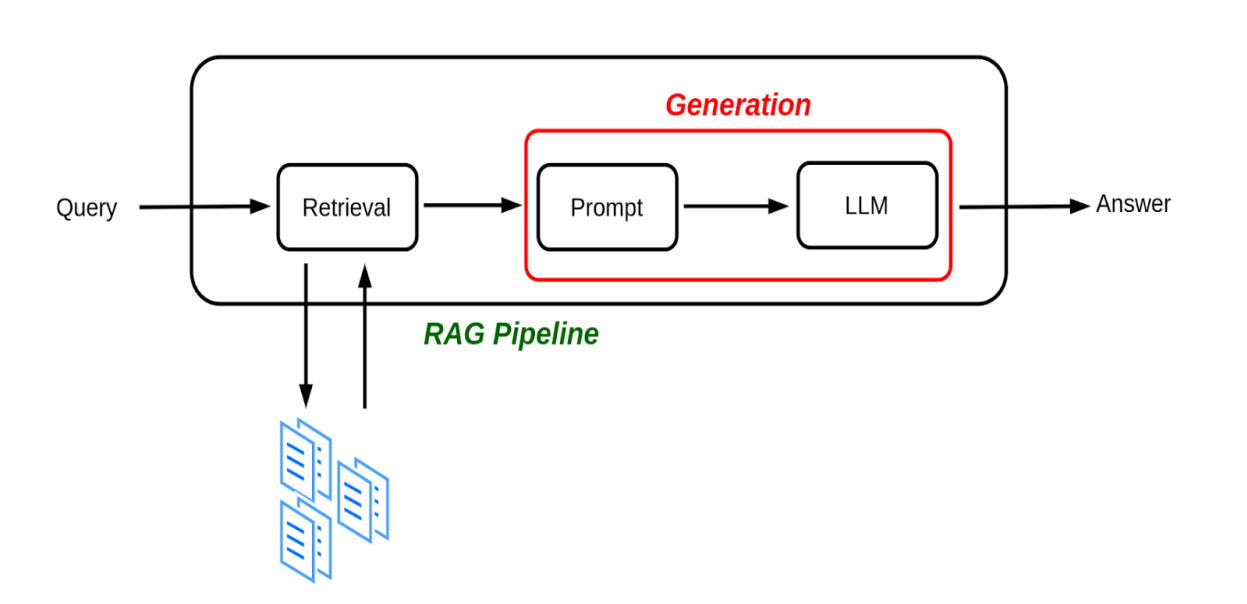
1. User Input 📝: The process begins with a user prompt or question. The prompt establishes context, while the question specifies the desired information.

2. Document Retrieval:

- Corpus Creation 📚 : Our document corpus consists of 100 Terms & Conditions documents from various companies. These documents are divided into approximately 3,820 chunks, each comprising seven sentences: three new sentences per chunk and a window of two sentences from the previous chunk.
- Chunking Strategy ✂️: We separate sentences assuming ". " as the delimiter between sentences. Each chunk provides enough contextual information for our retrieval model to accurately select relevant content.
- Embedding Generation 🔍: Using the `bert-large-uncased` model, embeddings are created for each chunk and stored in a PostgreSQL table, alongside TF-IDF scores, chunk texts, and document-level metadata. These embeddings encapsulate the context of each chunk, allowing the retrieval model to prioritize relevance accurately.
- Query Processing 💭 : When a user submits a query, we convert it into an embedding using the same transformer model. We then compute dot products with all chunk embeddings to determine relevance, ranking the top 10 matches from the 3,820 possibilities.

3. Creating Enhanced Context 🧩: The system combines the retrieved information with the user's question to form a rich context for the LLM. This augmented context provides our model, `llama3-70b-8192`, with comprehensive background information.

4. Response Generation 🧠: The `llama3-70b-8192` model uses the enriched context to generate an accurate answer. The inclusion of retrieved context enables the model to produce complete, precise responses.



Implementation of RAG 🛠️:

Aspects	Details
Data Sources 📊	Structured (tabular) text corpora for efficient document retrieval.
Retrieval Techniques 🔄	Dot products between stored Embeddings and query Embedding for ranking document relevance.
Model Selection 🤖	llama3-70b-8192 for generating responses with fine-tuning.
Performance Metrics ⏱️	Response time and relevance of retrieved documents.
Tools and Frameworks 🔧	**Hugging Face Transformers**: For access to pre-trained models like bert-large-uncased and llama3-70b-8192. **SQLAlchemy**: For efficient management of PostgreSQL databases containing chunk embeddings and document information.

Relevance Measurement 🎯 : To assess the model's response relevance, we created a set of 15 test questions. Based on this test set, we observed that most answers were accurate. However, a question remains: are these answers correct due to the strength of the generation model, or is it our retrieval system that provides such effective context, enabling the model to generate accurate responses?

1. **Generation Model Check** 📜 : the generation model, "llama3-70b-8192," has not been explicitly trained to read terms and conditions documents and answer questions on this specific domain. Therefore, it's unlikely that "llama3-70b-8192" alone could handle such specialized questions effectively without additional contextual support.
2. **Retrieved Document Verification** ✅ : we also verify the retrieved documents. Both the generated response and the retrieved documents are shown to the user, allowing them to cross-verify the accuracy and relevance of the response. This approach provides the user with the context needed to evaluate the correctness and reliability of the answer when there are any doubts.

Drawbacks of this approach ⚠️ :

High Response Time ⏳ : The main problem is, it's taking quite a high time.

Since it has to generate the query Embeddings first, then calculate dot product with every (3820) chunks in the database so it's quite natural to take that much time.

Improvements 🚀 : We'll use BM25 algorithm for initial level of matching and then we can use Embedding based dot product method to retrieve top 10 documents. Here we can use parallelization also as follows.

Initially the moment we got the query, we passed it to the Embedding generator model on one thread and on another thread we will use the BM25 algorithm for finding top 100 or 200 documents.

Let's see an example scenario 🔍 ,

Question: What types of personal information does 17TRACK collect and how is it used across their services?

Retrieved documents with using BM25 (using combination of dot product & BM25):

1. ['the column on the right provides a short explanation of the privacy policy and is not legally binding this privacy policy covers 500pxs treatment of personal information that 500px gathers when you are on the 500px website and when you use 500px services this policy does not apply to the practices of third parties that 500px does not own or control, or to individuals that 500px does not employ or manage information collected

by 500px we only collect personal information that is relevant to the purpose of our website this information allows us to provide you with a customized and efficient experience we collect the following types of information from our 500px users basically, we collect information to make 500px useful for you and to provide a personalized experience information you provide to us 1 we receive and store any information you enter on our website or provide to us in any other way',

2. 'this privacy policy has been designed and developed to help you understand the following the type of personal information including sensitive personal data or information that 1mg collects from the users the purpose of collection, means and modes of usage of such personal information by 1mg how and to whom 1mg will disclose such information how 1mg will protect the personal information including sensitive personal data or information that is collected from the users and how users may access and or modify their personal information this privacy policy shall apply to the use of the website by all users sellers accordingly, a condition of each users use of and access to the website and to the other services provided by 1mg to users is their acceptance of this privacy policy',

3. 'information collected by 500px we only collect personal information that is relevant to the purpose of our website this information allows us to provide you with a customized and efficient experience we collect the following types of information from our 500px users basically, we collect information to make 500px useful for you and to provide a personalized experience information you provide to us 1 we receive and store any information you enter on our website or provide to us in any other way you can choose not to provide us with certain information, but then you may not be able to take advantage of many of our features 2 registration in order for you to use 500px services you must complete a registration form as part of this registration form, we require certain personal information',

4. 'among these cookies are, for example, those used for the setting of language and currency preferences or for the management of first party statistics employed directly by the owner of the site other types of cookies or third parties that install cookies some of the services listed below collect statistics in an anonymized and aggregated form and may not require the consent of the user or may be managed directly by the owner depending on how they are described without the help of third parties if any third party operated services are listed among the tools below, these may be used to track users browsing habits in addition to the information specified herein and without the owners knowledge please refer to the privacy policy of the listed services for detailed information analytics the services contained in this section enable the owner to monitor and analyze web traffic and can be used to keep track of user behavior google analytics for firebase google llc google analytics for firebase or firebase analytics is an analytics

service provided by google llc in order to understand googles use of data, consult googles partner policy',

5. …

Retrieved documents without using BM25 (only using dot product over Embeddings):

1. '1314 of regulation eu 2016679 general data protection regulation this privacy policy relates solely to this application, if not stated otherwise within this document latest update september 18, 2018',
2. 'you are advised to review this privacy policy periodically for any changes changes to this privacy policy are effective when they are posted on this page contact us if you have any questions about this privacy policy, please contact us log out',
3. 'you are advised to review this privacy policy periodically for any changes changes to this privacy policy are effective when they are posted on this page contact us if you have any questions about this privacy policy, please contact us by visiting this page on our website',
4. 'privacy at 33across 33across, inc and its corporate group affiliates collectively, 33across provides online publishers and marketers our clients with a suite of products and technologies products, or technology that provide insights into how content is consumed and shared on their web sites, monetize this content, and drive incremental traffic we consider the protection of user privacy to be of paramount importance and provide this privacy policy to inform consumers about how our technology collects and uses consumer data this privacy policy also outlines how you, as an enduser, may optout of data collection by 33across this policy also covers privacy practices on our corporate website website, privacy practices for our technology overview of how we use and protect information 33across provides products that enable clients to gain insights into how their
5. …

Observation & findings:

1. We can see a clear difference between the retrieved documents of both approaches.
2. Hence the generated answer would be significantly different.

3. **Time analysis:** In both approaches, we couldn't track any timing differences. So they were around 18 to 20 seconds.
4. **Relevancy analysis:** None of these 8 chunks are from the 17TRACK's privacy policy documents. The relevancy is very bad here. However if we try a query which has no company name specified, just general query, it performs pretty well.

Another improvement:

We'll pre-process the query first.

1. Use LLM models for correcting the syntactical error in the query
2. Then ask LLM for providing if company name's in the query
3. Get the company name and search among that company's documents only.
4. This will lower down our search space and hence lower the time taken.

Let's see an example. We'll take the same question taken previously.

Retrieved chunks after using these 2 improvements:

1. 'this agreement is made in chinese in case of any dispute as to the interpretation between the translated version and chinese version, the chinese version shall prevail about us contact us help terms privacy about us contact us help terms privacy',
2. 'network error, please try again later refresh view all welcome back login log in with no account register login sign up for free products bulk tracking 17track tracking api shopify app mobile app apps developers carriers integration tools tracking widget carrier list logistics center links help help center contact us login register termslast updated 8th aug 2021 this licensing agreement hereinafter referred to as this agreement is made between you and demon network tech co., ltd hereinafter called 17track please carefully read this agreement before using the software or 17track site hereinafter called licensing software by downloading, installing or using licensing software, you agree that you fully understand and accept all terms and conditions of this agreement if you do not accept the terms and conditions of this agreement, please do not use licensing software and remember to destroy all copies of licensing software 17track reserves the right to amend this agreement from time to time, and in case of that, we will post relevant notice on 17track official website without any separate notice',
3. 'you shall be liable for any losses and actual damages suffered by 17track or other users, as a result of your breach of this agreement governing law and dispute resolution the

validity, interpretation, modification, execution and dispute settlement shall be governed and construed by the laws and regulations of the peoples republic of China in the event that there are no relevant regulations, reference shall be made to the relevant provisions of local laws andor general international practices any disputes arising from this agreement shall be settled by both parties through friendly negotiations in the event that such disputes cannot be settled through negotiations, such disputes shall be submitted to the court of competent jurisdiction where 17track is located if any of the provisions of this agreement is held to be invalid by any court of competent jurisdiction, then such invalidity shall not affect any of the remaining provisions of this agreement and such remaining provisions shall be implemented by you and 17track',

4. 'based on this agreement, 17track grants you a license other than sell you licensing software you shall only use licensing software subject to this agreement, 17track reserves all rights that are not expressly granted usage specification you shall not use licensing software in the following ways any behavior in violation of laws and regulations, public order and good morals, or harmful for public interests rent, lend, copy, amend, link, reproduce, compile, issue, publish, and set up mirror sites for licensing software without authorization based on licensing software, developing any licensing software related derivative products, works, services, plugins, compatibility, interconnections, etc log in or use licensing software through any third party compatible software system, which is not developed, authorized or approved by 17track, or use plugins, which are not developed, authorized or approved by 17track on licensing software delete any modify, delete or intentionally avoid technique security measures in the application, which are set up by 17track for intellectual property protection purposes',
5. …

Observation:

1. If the query contains any company name, which is quite natural for this scenario, this is working much better.
2. **Time analysis:** It just took 5-6 seconds to execute the query with the company name.
3. **Relevancy analysis:** We only received the chunks from 17TRACK’s privacy policy documents.