DIVERSIFIED ANSWER GENERATION QA SYSTEM

Milestone Meeting: 10.01.2025

Agenda

1. Data Collection:
    1. Ground Truth QA pairs
    2. Knowledge Base
2. RAG Pipeline
3. Evaluation Methods for Diversified Answer Generation

1. Data Collection: Ground Truth QA Pairs

Completed Tasks:

- Created an OpenSearch index which stores the URLs of all blog posts by AskEP.
- Crawled and stored in a separate OpenSearch index the following information from each blog post: *html code, URL, publication date, question, answer* (as a dictionary where the section titles are the keys and the section texts are the values), *links, tags* and *tag URLs*.
- Both indices are periodically updated. At the moment there are 168 ground truth QA pairs.

Problems:

- 23 documents could not be indexed, they throw the error:
  `RequestError(400, 'mapper_parsing_exception', 'failed to parse')`
- Inexact questions (title or first paragraph can be the question).
- (Same answer in multiple languages.)

Necessary improvements:

- Index the remaining documents.

1. Data Collection: Knowledge Base

1. Web Scraping

Completed Tasks:

- Extracted, analysed and filtered the external URLs (eliminated duplicates, links throwing various errors, YouTube, mailto links and insecure links).
- Separated PDF and Non-PDF URLs for further processing.
- For regular web pages I used Beautiful Soup to extract the text. For the PDFs, I used the MarkItDown tool to convert each PDF to Markdown and then extract the text (is particularly useful for PDFs with tables). There were 1651 Non-PDF URLs, 134 PDF URLs and 359 URLs which threw errors. Currently I crawl the text from all URLs without taking possible internal hierarchies into account.

Possible Improvements:

- Crawl also the documents which certain pages link (e.g. https://epthinktank.eu/?s=TTIP).
- Take internal hierarchy into account.
- Use Apify Website Content Crawler to crawl the contents as another method.

1. Data Collection: Knowledge Base

2. Chunking, Embedding and Storing

Completed Tasks:

- I use the RecursiveCharacterTextSplitter from Langchain with various chunk sizes and overlaps to chunk the content of each crawled page.
- The chunks of each page are embedded using mxbai-embed-large embedding model provided through Ollama. It is a 335M parameter model with embedding dimension 1024 (SOTA in March 2024 on MTEB).
- Each OpenSearch index for the knowledge base consists of documents containing the following fields: URL, chunk id, chunk text and embedding. Currently, I created one index using chunk size 400 and chunk overlap 50. Indexing ~60% of the contents of the Non-PDF URLs took 12 hours. This is due to web pages like: link.

Possible Improvements:

- Experiment with other text splitters and better embedding models.
- Optimize the current code.
- Create further indices with different configurations.

2. RAG Pipeline

Completed Tasks:

- To select documents which are relevant to a query and diverse, I use the MMRRetriever from Langchain, which ranks documents using the Maximal Marginal Relevanve score and then retrieve the top-k documents for a given query.
- To generate a diversified answer, I use Llama 3.3 through Ollama with the following prompt template:

> **Question**: **{question}**
> The following documents provide relevant information: **{context}**
>
> Please answer the question, but make sure to provide a diversified response that covers different perspectives and details from the provided documents. Your answer should include multiple viewpoints and insights from the context, not just a single perspective. If necessary, highlight different interpretations, opinions, or additional context that is relevant to the question.
> Answer the question comprehensively, using the information from the documents provided.

Possible Improvements:

- Experiment with hybrid retrieval strategies, increase the nr. of retrieved documents.

3. Evaluation Methods for Diversified Answer Generation

Ideas for Evaluation:

- Semantic Similarity with the ground truth answer.
- Check coverage (How many key points from the ground truth answer does the generated answer address?). Maybe use an LLM to judge.
- Classic evaluation metrics for the retrieval component and then the RAG pipeline as a whole.
- What to do regarding diversity in the answer generation?