Report

Dev Goyal

11th August 2023

1 Problem

The goal was to work on a next gen search engine which parses the web for information and gives a concise summary of the query. My work involved on starting with the domain of different professors. Given a professors name, extract various attributes related to the professor by leveraging artificial intelligence models and NLP techniques to analyse uncleaned free form text on the web and make sense of it.

2 Approach

2.1 Data Collection

the first step was to extract relevant documents from the web pertaining to the professor in question. In lieu of focusing the time and effort on building a web crawler, I used a Google search api using SERP and then a used a prebuilt web scraper library called Trafilatura to extract the relevant text from the web pages. The text was then cleaned and divided into different sections by extracting the headers and the text between them. This part of the pipeline remained fairly constant throughout the project.

2.2 Attribute Extraction

The meat of the project was to extract the relevant attributes from the text. The basic approach was to use a Large Language Model(LLMs) to prompt the text as context and then ask questions to extract the attributes. As one might have guessed this approach had limitations of it's own namely the length of the context and the speed of the evaluation of the model. To overcome these limitations, it was decided to first extract the relevant sections with respect to each question and then prompt the model given the most relevant context. The following techniques were looked into to extract the relevant sections:

2.2.1 Sentence Similarity

This involved using different models to embed the paragraphs and then using cosine similarity to find the most similar paragraph to the question. some models used were:

- Sentence BERT
- BERT
- instruction based encoders

out of these sentence BERT performed the best. The problem with this approach was that it was very slow.

2.2.2 Retrieval Augmented Generation

This approach involved using a retriever to find the most relevant paragraphs and then using a generator to generate the answer. The retriever used was DPR and the generator used was T5. This approach was faster than the previous one but the retrieval part was sometimes inaccurate even though the Generation was reliable based on the context.

2.2.3 Hybrid Approach

So it was decided to use a hybrid approach by embedding the paragraphs using sentence BERT and use KNN to find the most similar paragraphs and then use the generator to generate the answer. This approach turned out to be ideal.

3 Results

With sentence BERT capturing the semantic relationship between amongst the paragraphs and being more immune to the noise in the prompt, the regular KNN approach worked very accurately in finding the most relevant context to provide to the generator by choosing the top 5 most relevant passages. The generator was able to generate the answer to straightforward questions involving the educational background and research interests with a very high degree of accuracy but had issues with more complex questions like extracting the address of the professor and the recent courses taught. This was mainly due to the different ways the information was represented in the context which made it hard for the model to understand what information from the context the user really wants. Other limitations included the unavailable information on the web and the restricted access to the web pages which made it hard to extract the information. Also, the web scraping library sometimes failed to extract all the text from the web page in order to avoid the noise.

4 Assessment

The progress made in the project was a small step towards a larger goal of building a more open domain search engine. With the model being able to extract readily available information relatively quickly and accuately I would say that considerable progress was made and it would not be tough to scale the project and improve it.

5 Reflection

The project was a great learning experience for me. I got to learn about the difference between the theoretical approach and practical implementations of a technique. While many of the techniques I used were theoretically sound, they were not very practical. Reading a new research paper every week and implementing it was a great way to learn about the latest research in the field. Seeing the progress that I made in the project helped me realise the depth of the field and the amount of work that goes into finding a solution to a research problem.

6 Future Work

The project has a lot of scope to improve and almost every aspect of the project can be worked on to boost performance.

1. different web crawlers and web scraping techniques can be integrated to imporve the data collection process.
2. Playing around with different models to embed the text can help in capturing more relevant information among the text.
3. Using different models to generate the answer can help in generating more accurate answers.
4. prompt engineering can be used to improve the quality of the prompt.
5. iterative search can be used to find information that wasn't found in the first attempt.