LLM Zoomcamp

Bilal Naseem[*]

June 15, 2024

Contents

1 Week 1 **2**

1.1 Introduction to LLM and RAG 2

1.2 Preparing the Environment 2

1.3 Retrieval 2

1.4 Generation with OpenAI 3

1.4.1 OpenAI API Alternatives 3

1.5 Cleaned RAG flow 3

1.6 Searching with ElasticSearch 3

Abstract

The following notes were made based on the course **LLM Zoomcamp**. This course does not cover the theory behind LLMs and treats them as black boxes.

[*] https://www.linkedin.com/in/bilalnaseem96/

1 Week 1

1.1 Introduction to LLM and RAG

[1]A language model is a model which predicts the next word based on the words which you have typed so far. A **Large Language model** also does the same thing, but has a lot more parameters (billions). The input to the LLMS (text/image/video etc.) is called prompt.

RAG stands for Retrieval Augmented Generation. Retrieval means search, so a RAG system uses search to augment the generation (make it better) of the text.

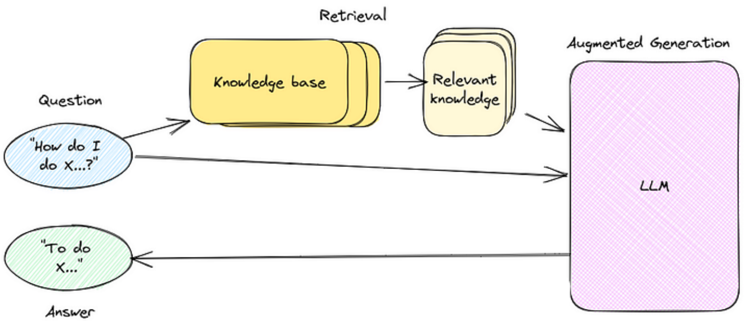

Figure 1: RAG system overview

1.2 Preparing the Environment

```
python3 -m venv myenv
source myenv/bin/activate
```

[2] Commands `which python3` and `python3 -V` can be used to check source and version of python.

```
pip install tqdm openai elasticsearch scikit-learn pandas
pip freeze > requirements.txt
```

OpenAI api key can be obtained from here.

```
load_dotenv()
openai_key = os.getenv('OPENAI_KEY')
###
client.chat.completions.create(
    model='gpt-3.5-turbo',
    messages=[
       {
        "role": "user",
        "content": "What is up?"
       }
    ]
)
```

- messages is what we write to the client

1.3 Retrieval

xxx

[1] Lecture 1

[2] Lecture 2

1.4 Generation with OpenAI

1.4.1 OpenAI API Alternatives

1.5 Cleaned RAG flow

1.6 Searching with ElasticSearch