19.08.2024 **Prepared by Yakup Zengin**

Project Report

AI-Powered Q&A System for Web Traffic Logs

1. Introduction

This report covers the development and performance evaluation of an artificial intelligence (AI)-powered question-answer (Q&A) system that can generate answers to user questions from web traffic logs. The project designed a system that can generate meaningful and accurate answers by processing log data. This report will cover the development process of the system, the challenges encountered, and the performance of the system in detail.

2. Project Definition and Goals

This project aims to develop a robust system that processes web traffic logs and answers user queries effectively. The primary goals include:

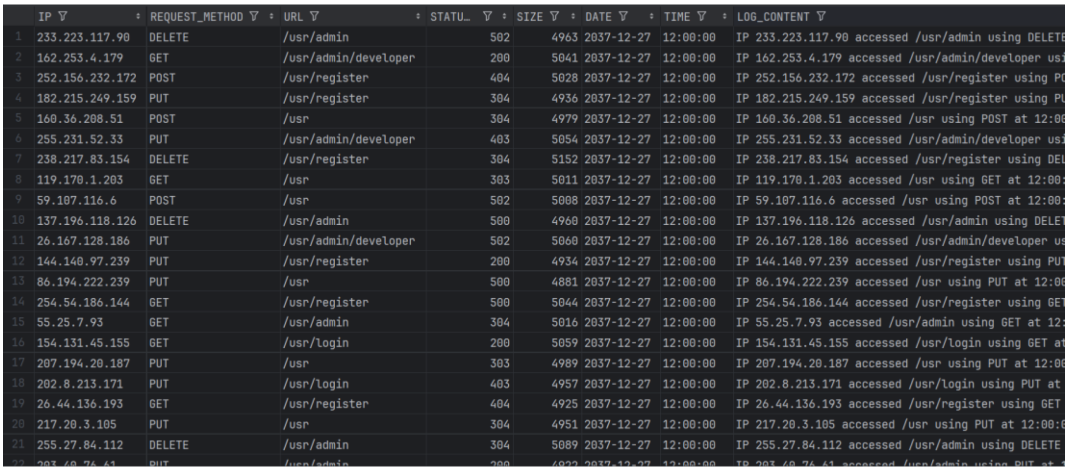
- **Data Processing:** Efficiently clean and prepare log data for analysis.
- **Vectorization:** Convert cleaned data into a format suitable for querying and indexing.
- **Indexing:** Create a FAISS index to facilitate fast and accurate search queries.
- **Query Handling:** Implement a query processing module that can generate relevant responses based on the indexed data.

3. System Architecture and Components

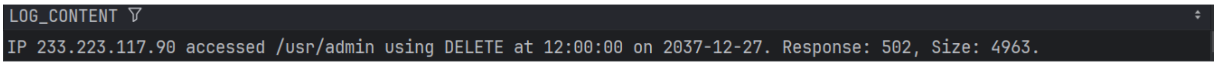
System is designed with the following components:

3.1. Data Preparation

- **Data Reader:** Reads raw web log data from a specified file.
- **Data Cleaner:** Cleans and preprocesses the raw data to remove noise and irrelevant information.
- **Data Saver:** Saves the cleaned data to a CSV file for further processing.



	IP	REQUEST_METHOD	URL	STATU…	SIZE	DATE	TIME	LOG_CONTENT
1	233.223.117.90	DELETE	/usr/admin	502	4963	2037-12-27	12:00:00	IP 233.223.117.90 accessed /usr/admin using DELETE
2	162.253.4.179	GET	/usr/admin/developer	200	5041	2037-12-27	12:00:00	IP 162.253.4.179 accessed /usr/admin/developer usi
3	252.156.232.172	POST	/usr/register	404	5028	2037-12-27	12:00:00	IP 252.156.232.172 accessed /usr/register using PO
4	182.215.249.159	PUT	/usr/register	304	4936	2037-12-27	12:00:00	IP 182.215.249.159 accessed /usr/register using PU
5	160.36.208.51	POST	/usr	304	4979	2037-12-27	12:00:00	IP 160.36.208.51 accessed /usr using POST at 12:00
6	255.231.52.33	PUT	/usr/admin/developer	403	5054	2037-12-27	12:00:00	IP 255.231.52.33 accessed /usr/admin/developer usi
7	238.217.83.154	DELETE	/usr/register	304	5152	2037-12-27	12:00:00	IP 238.217.83.154 accessed /usr/register using DEL
8	119.170.1.203	GET	/usr	303	5011	2037-12-27	12:00:00	IP 119.170.1.203 accessed /usr using GET at 12:00:
9	59.107.116.6	POST	/usr	502	5008	2037-12-27	12:00:00	IP 59.107.116.6 accessed /usr using POST at 12:00:
10	137.196.118.126	DELETE	/usr/admin	500	4960	2037-12-27	12:00:00	IP 137.196.118.126 accessed /usr/admin using DELET
11	26.167.128.186	PUT	/usr/admin/developer	502	5060	2037-12-27	12:00:00	IP 26.167.128.186 accessed /usr/admin/developer us
12	144.140.97.239	PUT	/usr/register	200	4934	2037-12-27	12:00:00	IP 144.140.97.239 accessed /usr/register using PUT
13	86.194.222.239	PUT	/usr	500	4881	2037-12-27	12:00:00	IP 86.194.222.239 accessed /usr using PUT at 12:00
14	254.54.186.144	GET	/usr/register	500	5044	2037-12-27	12:00:00	IP 254.54.186.144 accessed /usr/register using GET
15	55.25.7.93	GET	/usr/admin	304	5016	2037-12-27	12:00:00	IP 55.25.7.93 accessed /usr/admin using GET at 12:
16	154.131.45.155	GET	/usr/login	200	5059	2037-12-27	12:00:00	IP 154.131.45.155 accessed /usr/login using GET at
17	207.194.20.187	PUT	/usr	303	4989	2037-12-27	12:00:00	IP 207.194.20.187 accessed /usr using PUT at 12:00
18	202.8.213.171	PUT	/usr/login	403	4957	2037-12-27	12:00:00	IP 202.8.213.171 accessed /usr/login using PUT at
19	26.44.136.193	GET	/usr/register	404	4925	2037-12-27	12:00:00	IP 26.44.136.193 accessed /usr/register using GET
20	217.20.3.105	PUT	/usr	304	4951	2037-12-27	12:00:00	IP 217.20.3.105 accessed /usr using PUT at 12:00:0
21	255.27.84.112	DELETE	/usr/admin	304	5089	2037-12-27	12:00:00	IP 255.27.84.112 accessed /usr/admin using DELETE
22	203.40.76.61	PUT	/usr/admin	200	4922	2037-12-27	12:00:00	IP 203.40.76.61 accessed /usr/admin using PUT at 1



```
LOG_CONTENT
IP 233.223.117.90 accessed /usr/admin using DELETE at 12:00:00 on 2037-12-27. Response: 502, Size: 4963.
```

3.2. Vectorization and FAISS Index

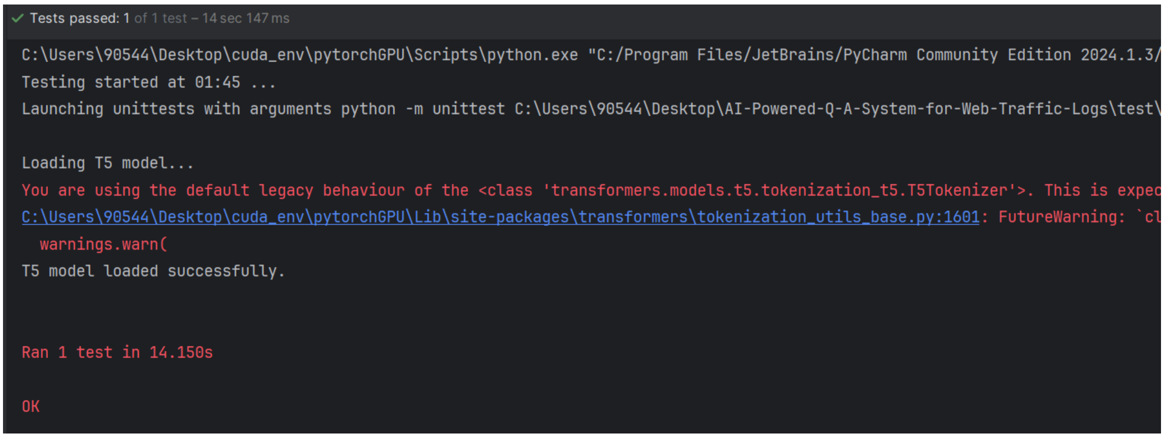
- **Vectorizer:** Converts cleaned data into vectors for efficient querying.
- **FAISS Index Builder:** Builds a FAISS index to enable fast similarity searches on the vectorized data.

3.3. Query Processing and Response Generation

- **Query Processor:** Processes user queries and generates answers by leveraging the FAISS index and vectorized data

3.4. Testing

- **test_file_reader.py:** Tests the Data Reader component to ensure it correctly reads log data from files and handles different log formats properly.
- **test_data_cleaner.py:** Evaluates the Data Cleaner component to check if it effectively removes noise and irrelevant information while preserving necessary data.
- **test_data_saver.py:** Assesses the Data Saver component to ensure that cleaned data is saved correctly to CSV files and the file format is appropriate for further processing.
- **test_vectorizer.py:** Tests the Vectorizer component to confirm it accurately converts cleaned data into vectors and performs vectorization efficiently.
- **test_faiss_index.py:** Validates the FAISS Index Builder component to ensure the FAISS index is constructed correctly and performs similarity searches as expected.
- **test_query_processor.py:** Evaluates the Query Processor component to check if it generates relevant and accurate responses based on the FAISS index and vectorized data.
- **test_model_loader.py:** Tests the Model Loader component to ensure the model is loaded correctly and is ready for use in query processing.



```
✓ Tests passed: 1 of 1 test – 14 sec 147 ms
C:\Users\90544\Desktop\cuda_env\pytorchGPU\Scripts\python.exe "C:/Program Files/JetBrains/PyCharm Community Edition 2024.1.3/
Testing started at 01:45 ...
Launching unittests with arguments python -m unittest C:\Users\90544\Desktop\AI-Powered-Q-A-System-for-Web-Traffic-Logs\test\

Loading T5 model...
You are using the default legacy behaviour of the <class 'transformers.models.t5.tokenization_t5.T5Tokenizer'>. This is expec
C:\Users\90544\Desktop\cuda_env\pytorchGPU\Lib\site-packages\transformers\tokenization_utils_base.py:1601: FutureWarning: `cl
  warnings.warn(
T5 model loaded successfully.


Ran 1 test in 14.150s

OK
```

4. Challenges and Solutions

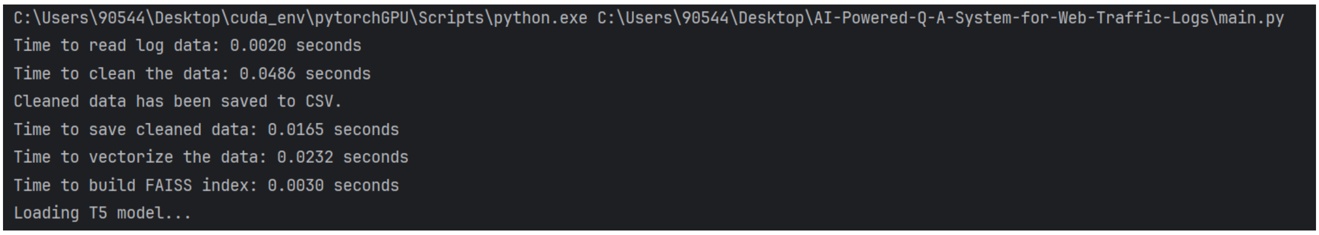
- **Challenge:** Handling large volumes of log data efficiently.
    - **Solution:** Implemented optimized data processing and indexing techniques to manage large datasets effectively.
- **Challenge:** Ensuring accurate query processing and response generation.
    - **Solution:** Refined the query processing algorithm and incorporated feedback loops for continuous improvement.

5. Performance Evaluation

5.1. Time Measurements

The following time measurements illustrate the duration of different stages in the system:

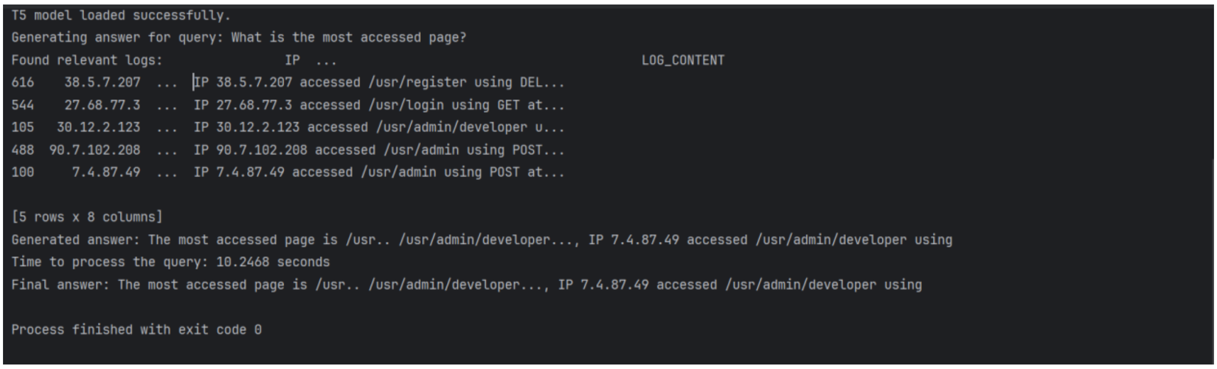
- **Log Data Reading Time:** 0.0020 seconds
- **Data Cleaning Time:** 0.0486 seconds
- **Cleaned Data Saving Time:** 0.0165 seconds
- **Vector Creation Time:** 0.0232 seconds
- **FAISS Index Creation Time:** 0.0030 seconds
- **Query Processing Time:** 10.2468 seconds



```
C:\Users\90544\Desktop\cuda_env\pytorchGPU\Scripts\python.exe C:\Users\90544\Desktop\AI-Powered-Q-A-System-for-Web-Traffic-Logs\main.py
Time to read log data: 0.0020 seconds
Time to clean the data: 0.0486 seconds
Cleaned data has been saved to CSV.
Time to save cleaned data: 0.0165 seconds
Time to vectorize the data: 0.0232 seconds
Time to build FAISS index: 0.0030 seconds
Loading T5 model...
```

5.2. Accuracy and Quality

The system produces accurate and meaningful responses to test queries. However to enhance performance and accuracy, the following recommendations are suggested.



```
T5 model loaded successfully.
Generating answer for query: What is the most accessed page?
Found relevant logs:                IP  ...                                      LOG_CONTENT
616    38.5.7.207  ...  IP 38.5.7.207 accessed /usr/register using DEL...
544    27.68.77.3  ...  IP 27.68.77.3 accessed /usr/login using GET at...
105   30.12.2.123  ...  IP 30.12.2.123 accessed /usr/admin/developer u...
488  90.7.102.208  ...  IP 90.7.102.208 accessed /usr/admin using POST...
100     7.4.87.49  ...  IP 7.4.87.49 accessed /usr/admin using POST at...

[5 rows x 8 columns]
Generated answer: The most accessed page is /usr.. /usr/admin/developer..., IP 7.4.87.49 accessed /usr/admin/developer using
Time to process the query: 10.2468 seconds
Final answer: The most accessed page is /usr.. /usr/admin/developer..., IP 7.4.87.49 accessed /usr/admin/developer using

Process finished with exit code 0
```

6. Improvement Recommendations

- **Training with More Data:** Increasing the amount and variety of training data can improve the model's accuracy.
- **Model Parameter Tuning:** Optimizing model parameters may enhance both performance and accuracy.
- **Advanced Algorithms:** Utilizing more advanced algorithms and techniques can improve the overall performance of the Sistem.
- **Data Quality Enhancements:** Improving the quality of raw data can lead to better results in the cleaning process.

7. Conclusion

The developed AI-powered question-answering system effectively operates on web traffic data. The system’s accuracy and performance have been evaluated through testing and time measurements. The proposed improvements will guide future enhancements to the system’s performance.

8. Appendices

Github Repository : https://github.com/yakupzengin/AI-Powered-Q-A-System-for-Web-Traffic-Logs