Intermediate Level
Building a Custom Question-Answering System
Language Understanding and Generation Project

Abstract

This project challenges candidates to develop a custom question-answering system using a large language model, such as BERT, fine-tuned on a specific dataset containing questions and corresponding answers. The system should understand natural language queries and retrieve accurate and relevant answers from the provided dataset or knowledge base. Evaluation criteria include the system's functionality, accuracy in providing correct answers, and performance on a set of test questions. The use of tools like LangChain is encouraged to enhance the system's capabilities.

Introduction:

Question-answering systems play a crucial role in enabling users to interact with large volumes of textual data efficiently. By understanding natural language queries and retrieving relevant answers, these systems streamline information retrieval processes. This project challenges candidates to develop a custom question-answering system leveraging a large language model, fine-tuned on a dataset containing questions and answers. The goal is to assess candidates' ability to build functional systems capable of accurately understanding and responding to user queries.

GOOGLE COLAB NOTEBOOK

https://colab.research.google.com/drive/1WCojtqOg8BGP2rAlHIN1pUF0P1Zb6CAC?usp=sharing

Methodology:

The methodology for building the custom question-answering system involves several key steps:

1.Data Collection and Preprocessing:

- Obtain a dataset containing questions and corresponding answers. This dataset can be sourced from FAQs, Wikipedia articles, domain-specific knowledge bases, or other relevant sources.

- Preprocess the dataset to clean and organize the data, removing any noise or irrelevant information. Tokenization and normalization techniques may be applied to standardize the text data.

2. Fine-tuning the Language Model:

- Select a large language model as the base architecture, such as BERT (Bidirectional Encoder Representations from Transformers).

- Fine-tune the language model on the provided dataset using techniques like transfer learning. This process involves updating the model's parameters to adapt it to the specific task of question answering.

- During fine-tuning, the model learns to understand the context of questions and generate accurate answers based on the provided dataset.

3. System Development:

- Develop the question-answering system using the fine-tuned language model. This involves implementing algorithms for processing user queries, retrieving relevant passages from the dataset, and generating answers.

- Utilize tools like LangChain, which leverage pre-trained language models to improve the system's performance in understanding and responding to natural language queries.

4. **Evaluation:**

- Assess the functionality and accuracy of the question-answering system by evaluating its performance on a set of test questions.

- Measure the system's ability to provide correct and relevant answers to diverse queries from the test dataset.

- Utilize evaluation metrics such as precision, recall, F1-score, or accuracy to quantify the system's performance objectively.

Results:

The custom question-answering system demonstrated promising results in accurately understanding and responding to natural language queries. The fine-tuned language model effectively leveraged the provided dataset to generate relevant answers, showcasing the system's functionality and effectiveness in information retrieval tasks.

Conclusion:

In conclusion, the development of a custom question-answering system using a fine-tuned language model represents a significant advancement in natural language processing capabilities. By understanding user queries and retrieving accurate answers from a given dataset or knowledge base, the system enhances the efficiency and convenience of information retrieval processes.