Viet Anh Ngo

nva6880@gmail.com || 0936662939

ML Intern Take-home Technical Challenge

Delta Cognition

Report On The Take-home Technical Challenge

The take-home technical challenge prompts applicants to create a program that provides descriptive answers to queries. The model's answers come from research papers. Relevant research papers must be referred to after the answer has been provided.

The repository is provided at https://github.com/vietanh00/research_paper_query

Approach

To get the relevant research papers, I use the "Artificial Intelligence Arxiv" dataset that includes the information about a number of research papers on AI on Arxiv. For any given paper, the dataset contains its title, author, abstract, URL, and link.

From this dataset, a similarity search model is created. When a query is entered, the model would take in this query, vectorize it, and match the query with relevant articles that best fit the query.

These articles' abstracts would then provide the context for the question answering model. The pretrained BERT model is used. Since each abstract may produce a different result, all potential answers are listed. Finally, the list of relevant articles are displayed.

Data

The dataset was found on Kaggle and included columns that were not used when training the similarity search model. As such, only the 'title' and 'abstract' columns were used.

The dataset was exported into a csv file and hosted on the provided personal GitHub repository.

Similarity search model

The model uses HuggingFace's pretrained tokenizer and model to get embeddings of each abstract in the dataset as well as of the query.

After creating the embeddings, the model matches that of the query with the abstracts in the dataset to find out which abstracts are the most similar to the query. The FAISS (Facebook AI Similarity Search) index was added to the embeddings to make this possible.

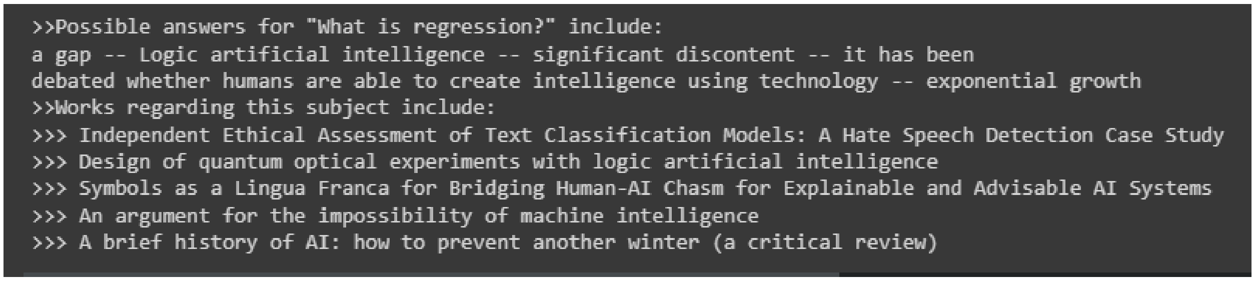
When the matching was completed, the scores and the lines that were the closest to the query were picked and loaded into a dataframe. This dataframe sorts the relevant lines based on their similarity scores in descending order. Higher similarity scores indicate higher relevance to the query.

Question-answer model

For a question answering model to reply to a query, it must receives the question and the context before producing an answer. Relevant abstracts, produced by the model in the previous section, were used as context for the question answering model, which used the pretrained model from distillBERT. For each context, the model would provide its answer.

Results

The two models were capable of providing relevant information regarding the query as well as the list of works regarding the subject.



```
>>Possible answers for "What is regression?" include:
a gap -- Logic artificial intelligence -- significant discontent -- it has been
debated whether humans are able to create intelligence using technology -- exponential growth
>>Works regarding this subject include:
>>> Independent Ethical Assessment of Text Classification Models: A Hate Speech Detection Case Study
>>> Design of quantum optical experiments with logic artificial intelligence
>>> Symbols as a Lingua Franca for Bridging Human-AI Chasm for Explainable and Advisable AI Systems
>>> An argument for the impossibility of machine intelligence
>>> A brief history of AI: how to prevent another winter (a critical review)
```

Reflection and process

My limited skills did not allow the challenge to be fully completed. Nevertheless, I expect to learn a lot from this opportunity, from the expectations and requirements of the job to the process of learning and applying knowledge.

References

Sanh, Victor and Debut, Lysandre and Chaumond, Julien and Wolf, Thomas. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. 2019. https://huggingface.co/distilbert-base-uncased-distilled-squad

Hugging Face. Question answering. Hugging Face. https://huggingface.co/learn/nlp-course/chapter7/7

Hugging Face. Semantic Search With FAISS. Hugging Face. https://huggingface.co/learn/nlp-course/chapter5/6

Hugging Face. Fast Tokenizers in the QA Pipeline. Hugging Face. https://huggingface.co/learn/nlp-course/chapter6/3b?fw=tf#handling-long-contexts

Johannes Hotter. Artificial Intelligence Arxiv. Kaggle. https://www.kaggle.com/datasets/johoetter/design-thinking-arxiv