**HACK TO HIRE – 2024**

****

**CASE STUDY FOR DATA SCIENTIST(4010)**

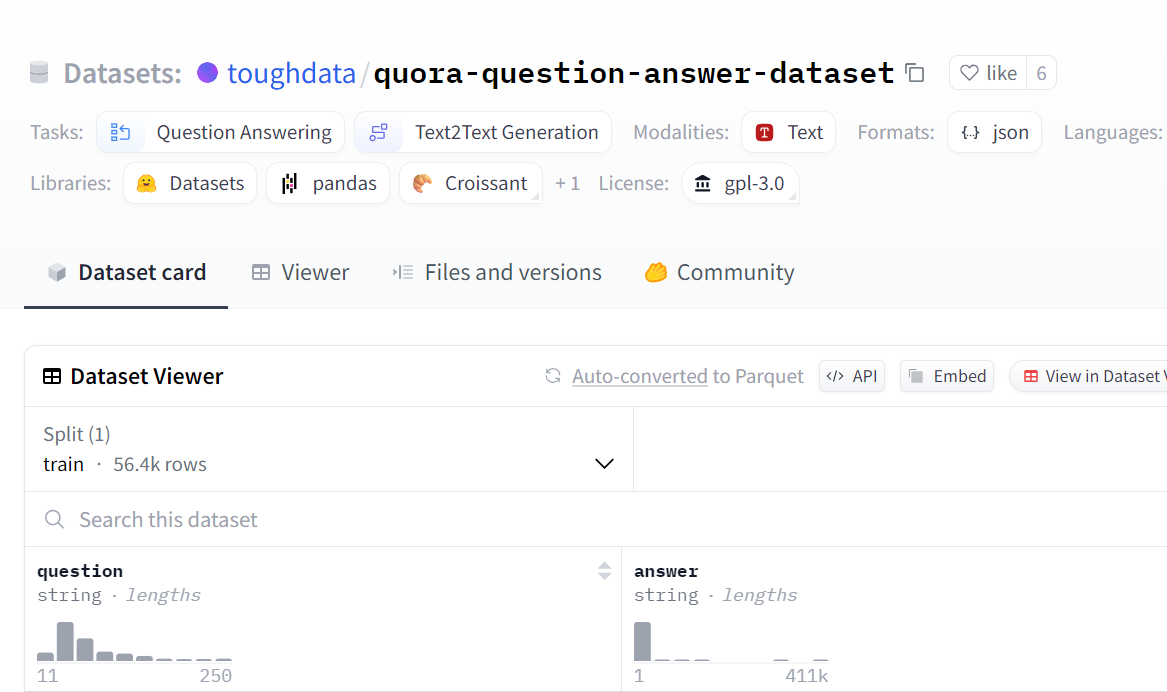
By: Cheshta Nagpal

**Problem Statement:** Develop a state-of-the-art question-answering model leveraging the QuoraQuestion Answer Dataset. The objective is to create an AI system capable of understanding and generating accurate responses to a variety of user queries, mimicking a human-like interaction.

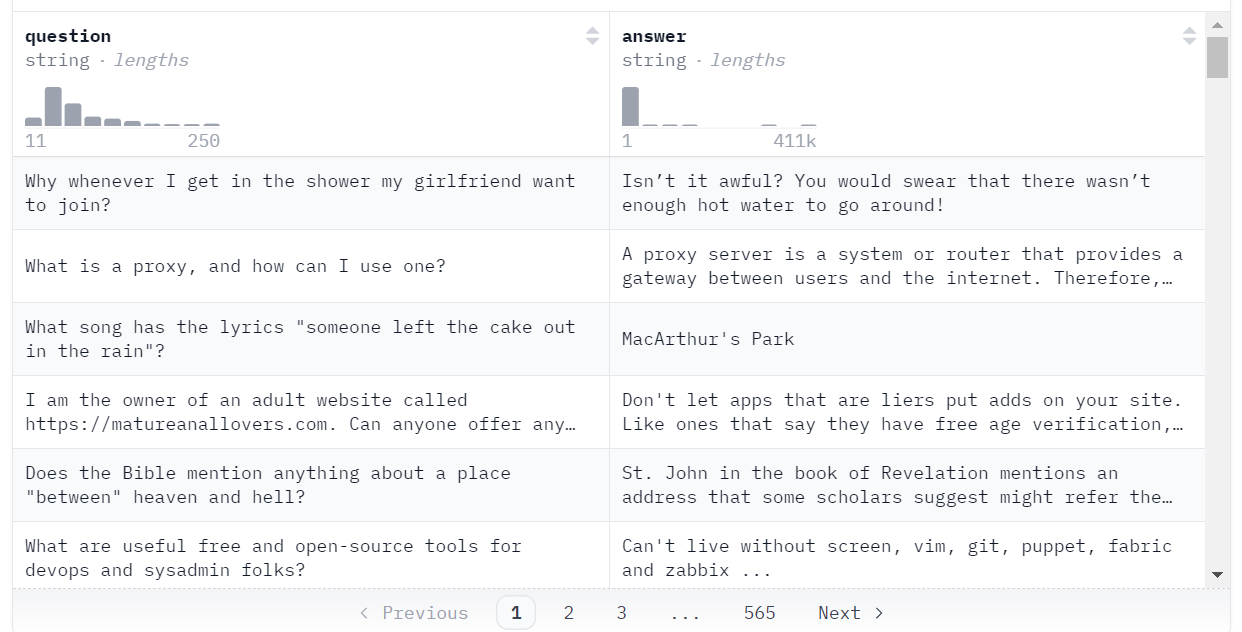
**DATASET:** https://huggingface.co/datasets/toughdata/quora-question-answer-dataset

1. **Data Exploration, Cleaning, and Preprocessing:**

As we can infer from the dataset it is a collection of 56402 questions and answers without context. Before performing any processing over this data, it was useful to tokenize it.



**EXAMPLES OF THE QUESTION ANSWERS IN DATASET**



1. **Model Selection and Evaluation:**
   1. **BERT MODEL**

[BERT](https://github.com/google-research/bert), or Bidirectional Encoder Representations from Transformers, is a method of pre-training language representations which obtains state-of-the-art results on a wide array of Natural Language Processing tasks.

Since no BERT model works without context we cannot use BERT model here as our dataset is closed book ,i.e, questions answers without context.

* 1. T5 MODEL

FLAN-T5 is an open-source, sequence-to-sequence, large language model that can be also used commercially. The model was published by Google researchers in late 2022, and has been fine-tuned on multiple tasks.

**FINDINGS:**

Rogue score is increasing with every epoch in training which shows that fine tuning our model to the dataset is yielding positive results.

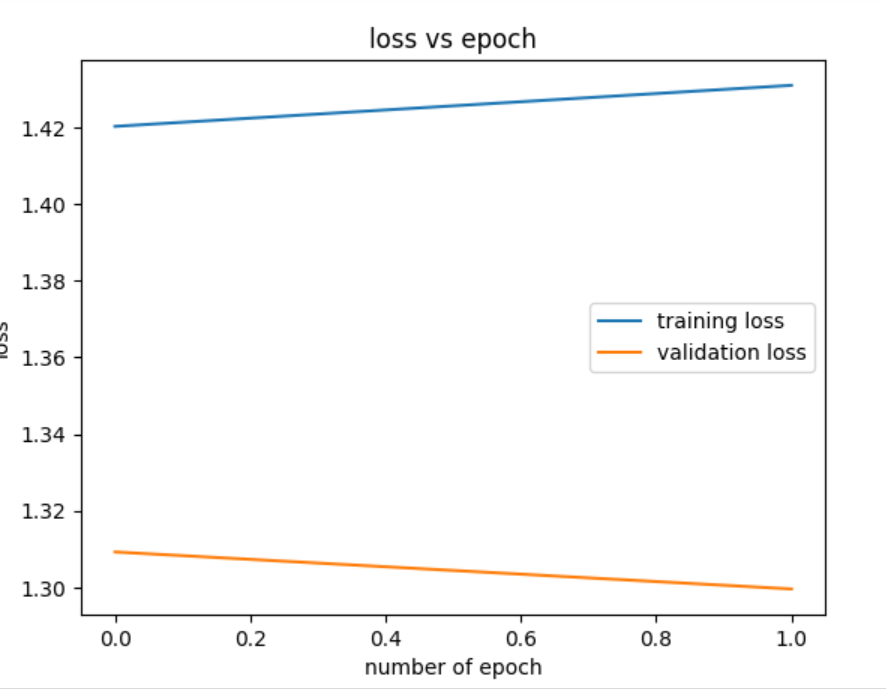
**RESULT:**

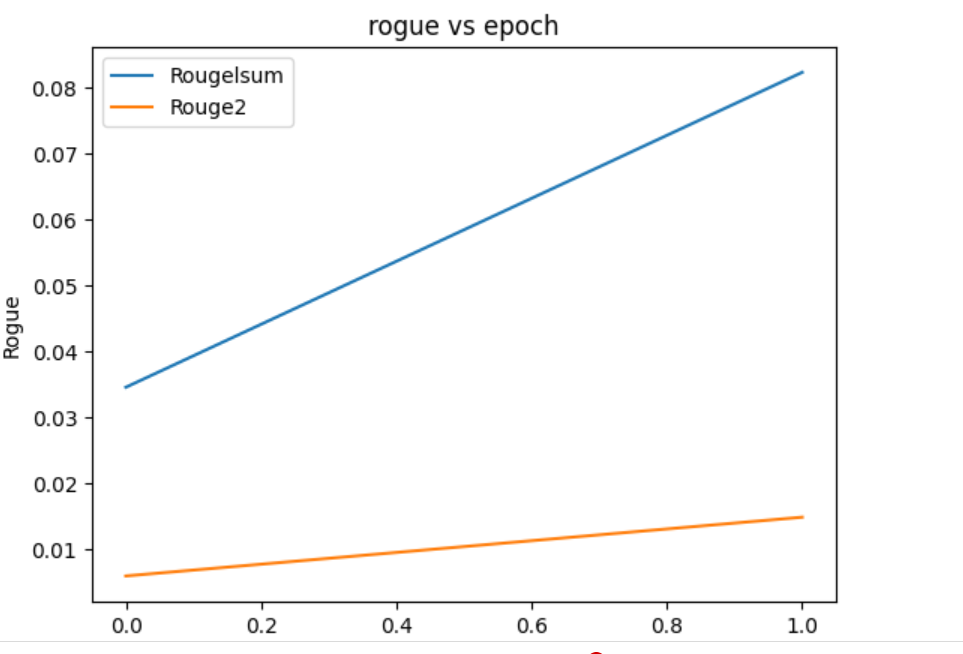
T5 model works for closed book question-answering which means that T5 model is effectively trained on a universal context

* 1. GPT MODEL

we will try to fine tune the pretrained gpt2 model of Hugging Face library.

1. **Visualization:**





**CONCLUSION:** Transfer learning, where a model is first pre-trained on a data-rich task before being fine-tuned on a downstream task, has emerged as a powerful technique in natural language processing (NLP). T5, GPT uses an encoder decoder model whereas BERT uses an encoder only model. Our findings show that encoder decoder model is able to answer questions without any context while only encoder model requires an external context. One may also say that encoder decoder model stores more information in its parameters than data itself.