Problem Statement-

Attached with document is Json file containing 37000 news articles from different sources reporting about the Israel Hamas war. It contains articles between October 2023 to March 2024. Not all articles talk about the Israel-Hamas war, the Json file also contain some spurious articles talking about different topics. Also, since the data has been scrapped from the web, the articles may contain noise such as punctuation, special characters, and irrelevant symbols.

Q1. Create a rudimentary Question Answering system that can answer questions related to the Israel Hamas war.

Solution-

I build the question-answering system using a pre-trained BERT model to answer questions related to the Israel-Hamas war based on a dataset containing 37000 news articles or related information.

Approach:-

1. The first step is to choose a model for question-answering.
   I have chosen the suitable Bert based pre trained model (bert-large-uncased-whole-word-masking-finetuned-squad). The reason for choosing the model is it handling large size of data and can uncover semantic relationship between the text data and one more reason is that it take both upper case and lower case letters same.
2. Load the data into the google colab which is in json format containing 37420 articles and have 5 columns. I worked only on articleBody because it contain the complete information and can help to answer the question in effective manner.
3. The question and the context are tokenized using the BERT tokenizer. Tokenization is basically the conversion of text into tokens (words or subwords) and converting them into numerical representations suitable for input to the model.
4. This function, answer_question, takes a question and a context as input. It first tokenizes the question and the context using the BERT tokenizer, converting them into a format suitable for the model. Then, it passes these tokenized inputs to the pre-trained BERT model. After obtaining the model's outputs, it identifies the start and end positions of the answer span within the context based on the highest logits (scores). Finally, it converts these token IDs back into a string representation to represents the predicted answer to the question.
5. Then combines all context strings into one, then uses a function to find and print the answer to the question within the combined context.
6. My Question is – “ What caused the Israel-Hamas war?”