MedQuAD (Medical Question Answering Dataset)

The MedQuad dataset provides a comprehensive source of medical questions and answers for natural language processing. With over 43,000 patient inquiries from real-life situations categorized into 31 distinct types of questions, the dataset offers an invaluable opportunity to research correlations between treatments, chronic diseases, medical protocols and more. Answers provided in this database come not only from doctors but also other healthcare professionals such as nurses and pharmacists, providing a more complete array of responses to help researchers unlock deeper insights within the realm of healthcare. This incredible trove of knowledge is just waiting to be mined - so grab your data mining equipment and get exploring!

How to use the dataset

In order to make the most out of this dataset, start by having a look at the column names and understanding what information they offer: qtype (the type of medical question), Question (the question in itself), and Answer (the expert response). The qtype column will help you categorize the dataset according to your desired question topics. Once you have filtered down your criteria as much as possible using qtype, it is time to analyze the data. Start by asking yourself questions such as “What treatments do most patients search for?” or “Are there any correlations between chronic conditions and protocols?” Then use simple queries such as SELECT Answer FROM MedQuad WHERE qtype='Treatment' AND Question LIKE '%pain%' to get closer to answering those questions.

Once you have obtained new insights about healthcare based on the answers provided in this dynmaic data set - now it’s time for action! Use all that newfound understanding about patient needs in order develop educational materials and implement any suggested changes necessary. If more criteria are needed for querying this data set see if MedQuad offers additional columns; sometimes extra columns may be added periodically that could further enhance analysis capabilities.

Link: https://www.kaggle.com/datasets/thedevastator/comprehensive-medical-q-a-dataset/data

TASK) Questioning Answering using Transformer based model

Implement following transformer based variants for the Question Answering task.

1. BERT
2. MobileBERT
3. RoBERTa

Link: https://simpletransformers.ai/docs/qa-specifics/

From the link given above you can get information about the model you need to fine-tune. Moreover you can find guideline on how input is tailored to pass to Transformer based models. Use 75% for training and 25% for testing.

For each of these models, try different hyper parameters and report the best results with parameter values. Like changing number of Encoder Layers etc.
Dropout rate, 0.3 or 0.7

Set n_best_size = 5 and for few questions show models top 5 predicted answers along with actual.

Use “wandb” to record training visualization.

Calculate BLUE Score and Rouge for both the models and report the results in table.

Also report parameter values which were used to get the results.