Contribution ID: **42**   Type: **Short Talk**

AI assisted user support

Thursday, 3 October 2024 12:20 (10 minutes)

Users may have difficulties to find the needed information in the documentation for products, when many pages of documentation are available on multiple web pages or in email forums. We have developed and tested an AI based tool, which can help users to find answers to their questions. The Docu-bot uses Retrieval Augmentation Generation solution to generate answers to various questions. It uses github or open gitlab repositories with documentation as a source of information. Zip files with documentation in a plain text or markdown format can also be used for input. Sentence transformer model and Large Language Model generate answers.

Different LLM models can be used. For performance reasons, in most tests we use the model Mistral-7B-Instruct-v0.2, which fits into the memory of the Nvidia T4 GPU. We have also tested a larger model Mixtral-8x7B-Instruct-v0.1, which requires more GPU memory, available for example on Nvidia A100, A40 or H100 GPU cards. Another possibility is to use the API of OpenAI models like gpt-3.5-turbo, but the user has to provide his/her own API access key to cover expenses.

Topic

Topic not listed

Primary authors: CHUDOBA, Jiri (CESNET);   Mr CHUDOBA, Michal (Faculty of Mathematics and Physics, Charles University)

Co-author: Dr HEJTMÁNEK, Lukáš (Institute of Computer Science, Masaryk University)

Presenter: CHUDOBA, Jiri (CESNET)

Session Classification: Processing Research Data with Artificial Intelligence and Machine Learning