13.SmartSearch into CodeBase

Organizational and thematic responsibility

KF: Software-Engineering and Architecture
Topics: SW-Development-Tools, Applied-AI
Contact: Bernhard Wallisch (walliscb@technikum-wien.at)

Vision

The goal of this project is the development of a tool for searching a project's codebase using prompts, resulting a list of referenced sources which are most likely appropriate to the searched text.

Description

To investigate the function of software projects and find the right places where specific behavior is implemented in a code-base can be a very time-consuming and complicated task, e.g. the question "where are new users created…" is hard to answer when using "find-in-files" or through debugging. In this project AI-technologies are applied to enable programmers asking a question, and the application will return a list of results of source-files and positions where the context applies best.

Cloud Solutions, like GitHub CoPilot would serve this, but for the companies it is often not acceptable to upload their complete code-base into the cloud and therefore need a local solution!

The implementation may be done with the help of the Haystack framework. In short it means, that in a first step the source-files needs to be indexed (Dense-Passage-Retrieval) and stored in a datastore (e.g. elastic-search, which is already supported by the Haystack framework).
In a second step - the search function, actually - the data-base will be searched (using the Extractive Question Answering Model) and the result list containing the file- and line-references is returned.
See the following article about application of the Haystack framework: https://ix.de/zy3m

For the user - the software developer - the project needs to provide two major functions:

1. A command (at least a command-line command) which will index the complete codebase and store it in a datastore (e.g. elastic-search) for later use. This command should be executable during build or along a CI/CD-pipeline (e.g. Java/Maven)
2. A command (at least a command-line command) which will take a prompt from the user as input, will execute the search, and return the result list with links to the corresponding positions in the source-files

The exact features are to be discussed with the supervisor.

Requirements and Conditions

- The supported programming language and tooling of the users can be freely defined.
- The type of the data-store may also be freely chosen (alternatives to elasticsearch)
- Pay attention, that the users will be required to have a GPU installed to enable for performant indexing and searching!
- **Overall estimation of required skills:**
  Mostly doable with the skills you earned in your FHTW studies.
  No sophisticated AI-skills/research is necessary! Turn a "feasibility sample" into a product.

Value for our faculty

Support students & lecturers in doing software development/coding projects.