1 Introduction of Assumption Miner

Assumption Miner is composed of three modules: **Data Collection**, **Dataset Management**, and **Data Analysis**. **Data Collection** aims to (1) show the data models of Repository, Release, Tag, PR, Commit, and Issue, (2) search and show data collection information of repositories and collect issues, PRs, and commits based on the data models, (3) monitor data collection processes, and (4) show data collection history. **Dataset Management** aims to (1) manage labels, (2) label the collected data, and (3) organize the labeled data into datasets. **Data Analysis** aims to (1) identify assumptions, (2) search data from the databases, and (3) show knowledge graphs of the data.

2 An example of how to use Assumption Miner

1. Users need to register an account and set a personal access token of GitHub when using Assumption Miner. The token is used to access the GitHub Application Programming Interface (API), since Assumption Miner needs to communicate with the GitHub API to get data (e.g., issues, PRs, and commits). We also provide a default token for Assumption Miner users. However, since GitHub has limitations in place to protect against excessive or abusive calls to GitHub servers (e.g., the rate limit is 5,000 points per hour and individual calls cannot request more than 500,000 total nodes), using the default token may lead to errors in data collection because of these limitations. After registration of the Assumption Miner account, users can login Assumption Miner with the account. Below is the process of using Assumption Miner to identify and extract assumptions from the TensorFlow project on GitHub.

2. Create the TensorFlow repository. Users need to click on the Repository Management module, then click on the “Add” button, enter the owner as “tensorflow” and the name as “tensorflow”, and click on the “Save” button to create the TensorFlow repository on Assumption Miner. For each release of a repository, Assumption Miner provides users a link to download the source code in the Repository Management module (this is an optional step). Then users can use tools such as Visual Studio Code and PyCharm to further browse the code and search assumptions in the code.

3. Collect issues, PRs, and commits on TensorFlow. After the TensorFlow repository is created on Assumption Miner, users can further use the Data Collection module to collect issues, PRs, and commits of the TensorFlow repository. Users can start multiple tasks simultaneously, but this could cause errors because of the limitation by GitHub.

4. Create lables. Users need to click on the Label Management module, then click on the "Add New" button, create three labels, i.e., NA with a value 0, PA with a value 1, and SCA with a value 2, which represents non-assumptions, potential assumptions, and self-claimed assumptions, respectively.

5. Label the data. Users need to click on the Data Labeling module, then input the search conditions, and click on the "Init" button. After showing the results, users can select the sentences that need to be labeled, and then right-click on the mouse, prvide the selected sentences labels.

6. Download the dataset. Users need to click on the Datasets module, find the dataset that need to be downloaded, and then click on the "download" button to download the dataset.