## Code search using Natural Language Queries

Dmitrii Shabalin d.shabalin@innopolis.university

Aidar Khuzin ay.khuzin@innopolis.university

github: <https://github.com/shabalin13/code-search>

During this iteration we measured the performance of CodeBert on test split CodeXGLUE dataset. We also observed the details of CodeBert: what is correct input, how to get it, how to pass batched data to it.

Firstly we obtained a tokenized version of test split and made embeddings for code and docstrings. The next step was to use knn classifier to retrieve a code item depending on its natural language description. We used faiss for finding nearest neighbors as it is much faster than scikit-learn implementation.

To measure performance we randomly shuffled the data and divided it into bins of size 1000. So for each docstring there will be 1 correct code and 999 incorrect ones. Using faiss we obtained 1000 nearest neighbors sorted by distance for each docstring embedding. Using this data we computed MRR(Mean Reciprocal Rank) for the test split.

Surprisingly for us, MRR was very low, about 0.007. Our assumption is that CodeBert should be fine-tuned on our dataset to get an acceptable result.

The next iteration we will fine-tune CodeBert and try GraphCodeBert. We are also looking for other approaches.