Alžbeta Strompová 492945

Alternative Second Term Project: ARQMath Collection: Project Report

System Description

My IR system uses the sentence transformer from https://www.sbert.net/ , which is a pre-trained model that has been fine-tuned for many use cases. It was trained on a large and diverse dataset of over 1 billion training pairs. I did not train this model on our data.

After using the sentence transformer to encode the documents, I use BM25Plus as a re-ranker for the top 10 documents founded by sentence transformer (cosine similarity). BM25Plus is a slightly modified version of *pv211_utils.systems.BM25PlusSystem*, with changes made to the pre-processing function and default parameters.

Training & System Parameters

I did not train model on our data. For sentence transformer, I used all-mpnet-base-v2. In this process model encode documents to vectors.

Development Process & Other Interesting Findings

In addition to reweighting the titles, I've also experimented with changing string representation of the query (For example adding query.tags lower percentage by approximately 3%.) or parameters in BM25Plus. I pick for each(3: k1, b, delta) parameter in BM25Plus three values and by hand made :D grid search I found best combination.

I have also tried replacing BM25Plus with TF-IDF, but I found this to lower the resulting percentage.