Additional approaches & complete table of results

During work on the project we also tried to use another set of embeddings obtained through Word2vec called law2vec

law2vec : a set of 200-dimensional embeddings trained on a large amount of legal corpora from various public legal sources in English, including the US code itself.

We also tried a simple "weighted" approach similar to "mixed query" , but we felt the performances of the two approaches were so similar that it wasn't worth it to have both on the official results table.

weighted

We define DE as:

$$DE = TFIDF \times WE$$

The variables $TFIDF$ & WE are defined exactly as in the "mixed query" approach of the paper

The query embedding q is defined as:

$$q = TFIDF_q \times WE$$

This is the exact same definition as $q_{weighted}$ in the paper.

Complete table of results on the whole dataset

APPROACH	Avg. CUSTOM ACCURACY	DOCUMENTS FOUND (%)
Unweighted google word2vec	0.57	0.60
Weighted google word2vec	0.58	0.60
Mixed query google word2vec	0.59	0.62
Weighted law2vec	0.52	0.54
Unweighted law2vec	0.51	0.53
Mixed query law2vec	0.52	0.54
Small bge	0.68	0.71
Small fine-tuned bge	0.72	0.75