A Parameter Tuning

We performed extensive parameter tuning for the baselines. For LSSI, we tuned over vocabulary sizes of 50K, 100K and 150K. We also tuned over the regularization parameter used for linear-SVR and linear-SVM on a validation set. The values were tuned over a range of [0.1, 1.0, 10.0, 100.0]. For UIUC, we tuned over the vocabulary sizes of 50K, 100K and 150K. We also tuned the kNN classifier for $k \in$ [5, 10, 15, 30]. We also tuned the regularizer of the linearSVM in the range [0.1, 1.0, 10.0, 100.0]. Afterwards, we tuned the threshold of the classifier for ten equidistant values in the range $(0, 1)$. For DML, we performed nested validation after each epoch to save the best performing model parameters. We also tuned the threshold for classification over ten equidistant values in the interval $(0, 1)$.

B Training

The training times for iterating over 1K samples can range in 30-40 minutes on Tesla K60 GPU, although, it can be faster with more advanced GPUs. We can converge with a learn rate of 0.01 in 30-40 epochs. We can converge over a set of 20K documents in $\approx$ 3-4 days.