Final Project

James Decatur and Claire Dickerson

1 Document Retrieval System

For our final project, we designed and implemented a document retrieval system that matches a given query – either a few words or an entire document – with the documents most similar to it in a data set. The model makes use of a Doc2Vec model – an unsupervised algorithm that learns feature vectors for documents – from the Gensim library.

The model is first trained on a given set of documents – either scikit-learn's 20 Newsgroups data set (the default), the Reuters data set, or a custom data set provided by the user. After training, the user can set the mode to 'run' submitting either a lengthy text query and or a file, as well as a folder with unknown documents within it. The program will output two csv files with the names and scores of the top 10 documents in the unknown data folder that are most relevant to the query or file given. The model also returns a 'vectors metadata' file and a 'vectors' file, which can be exported to e.g. http://projector.tensorflow.org/ for visualization.

2 Evaluation

The model uses cosine distance to evaluate model performance, which is based on the assumption that documents from the same categories will be more similar, and therefore located more closely in the vector space, than documents from different categories. We then loop through all categories, comparing documents within each category to each other, and average their scores. Evaluation returns two csv files, 'all', which contains the average of every matching pair versus the average of every non-matching pair, and 'pairs', which returns the average of every category pair.

Using an online P-value Calculator https://www.gigacalculator.com/calculators/p-value-significance-calculator.php, we obtain a P-Value of 0.475141, T-score of 0.062429, and a significance level of 52.49 percent for the 20 Newsgroups training and evaluation setup. We obtain a P-Value of 0.393001, T-score of 0.271523, and a significance level of 60.70 percent for our Reuters training and evaluation setup. Currently, evaluation only works with the Reuters and 20 Newsgroups data set, so adding an evaluation method for a custom data set would be an avenue for further work in the future.