UNIVERSITY
OF MANNHEIM
School of Business Informatics
and Mathematics

Data Mining

Exercise 7: Text Mining

7.1. Which documents are similar?

1. The file documents.zip is provided in ILIAS and contains three corpora. Load and vectorize the 4-documents corpus using load_files function. How many different attributes has the generated example set?
2. Examine the generated word list. What are the most common words? Look for the three most common words that might be helpful for text mining tasks?
3. Remove stopwords and apply the porter stemmer. By how many attributes do the operators reduce the size of your example set?
4. Compute the cosine similarity between the documents with the cosine_similarity function. Which documents are most similar? Can you confirm the judgment of the algorithm by reading the documents?
5. Experiment with different similarity metrics as well as with different vector creation methods. Which combination produces the best similarity scores?

7.2. Learn a Classifier for the 300-Documents Corpus

1. The 300-documents corpus contains postings from three different news groups. Vectorize the 300-documents corpus and learn a classifier for classifying the postings. Evaluate the classifying using 10-fold cross-validation. Which accuracy does your classifier reach? Increase the performance of your classifier by pruning the document vectors.

2. Try to do the same classification as in 7.2.1 using word2vec embeddings. You can aggregate word embeddings to get a document representation by applying mean pooling (elementwise average of word vectors).

3. Now do the same using BERT embeddings from the huggingface library. Experiment with mean pooling as well as using the [CLS] token representation as document representations.

7.3. Learn a Classifier for the Job Postings

1. The Job Postings corpus contains 500 descriptions of open positions belonging to 30 different job categories. The corpus is provided as an Excel file in ILIAS. Vectorize the corpus and learn a Naïve Bayes classifier for classifying the job adds. Evaluate the classifying using 10-fold cross -validation. Analyze the classifier performance and the word list. What do you discover?
2. Experiment with different vector creation and pruning methods as well as different types of classifiers in order to increase the performance. What is the highest accuracy you can reach? Which problem concerning precision and recall does remain?