UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Data and Web Science Group
B6 – B1.15
68159 Mannheim

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. Cluster the 30-Documents Corpus

1. The 30-documents corpus contains postings from three news groups. Vectorize the 30-documents corpus, remove stopwords and stem the corpus.
2. Use K-Means to cluster the corpus. Print the ground truth and the prediction. Compute the adjusted Rand index.
3. Examine the distribution of frequent words over the three different classes in the word list. Does the distribution give you an idea how you could improve the clustering using any of the prune methods (max_df, min_df)?

7.3. Learn a Classifier for the 300-Documents Corpus

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.

7.4. 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?

ChatGPT Bonus Exercises

Reminder: Do not take the answers of ChatGPT at face value! Always cross-check with lecture slides, literature and/or the teaching staff!

C.1. Discuss about Text Mining and preprocessing steps for textual data

- Discuss with ChatGPT the common preprocessing steps used in text mining. Ask it to explain to you a preprocessing step listed that you did not understand. Example: What is POS tagging and how can you use it?
- Pre-trained language models such as BERT outperform traditional, bag-of-words based NLP methods on many text classification tasks. Discuss with ChatGPT the potential reasons for the improved performance.

C.2. Learn about Feature Generation and Feature Selection

- Ask ChatGPT how you can generate features from textual data. Further ask it to provide you an example code of generating features from text using TF-IDF. For the same documents, ask it to show you how you can generate features using word embeddings.
- In applications, not all features that are generated are helpful and/or may lead to the "curse of dimensionality". Ask ChatGPT what the curse of dimensionality is (in the context of text mining) and some common feature selection methods that are used to overcome it. Further ask for an example textual dataset and how to perform feature selection on it. Choose a classification algorithm and learn a model using the data from before feature selection and after and compare the performance of the two models learned. Did feature selection help in this case regarding training time and model performance?

C.3. Self-Assessment

- Ask ChatGPT to give you some multiple-choice questions for graduate students related to text mining. Request the correct answers and compare with your own answers.
- Ask ChatGPT to generate you some short documents and to give you a paper and pen exercise to calculate binary term occurrence vectors of the documents. Find the most similar documents using Jaccard similarity. Ask ChatGPT to give you the answers when you have finished the exercise.