Automatic subject classification of Swedish DDC: Impact of tuning and training data set

Koraljka Golub

The presentation builds on the NKOS 2018 presentation of automatically produced Dewey Decimal Classification (DDC) classes for Swedish union catalogue (LIBRIS). Based on a data set of 143,838 records, Support Vector Machine with linear kernel outperforms Multinomial Naïve Bayes algorithm. Impact of features shows that using keywords or combining titles and keywords gives better results than using only titles as input. Stemming only marginally improves the results. Removed stop-words reduced accuracy in most cases, while removing less frequent words increased it marginally. Word embeddings combined with different types of neural networks (Simple linear network, Standard neural network, 1D convolutional neural network, Recurrent neural network) produced worse results than Naïve Bayes / Support Vector Machine, but reach close results. The greatest impact is produced by the number of training examples: 81.37% accuracy on the training set is achieved when at least 1,000 records per class are available, and 66.13% when few records on which to train are available.