Introduction

Dataset: Reuters-21578 Text Categorization Collection [link]

Goal: Classify the TOPIC for given news article (135 topics)

Method: SVM

Approach

Create a vocabulary from the news articles. Lemmatize/stem the words. Eliminate the stop words. Eliminate very rare words. Use reduced vocabulary as features. Use SVM to train the model on different kernels. Evaluate on the test set. Try to find more way to modify features (for example use document length or word occurrences).