Text Mining Final Paper Abstract

Guido Zuidhof, s4160703
October 28th, 2015

Dataset

The dataset consists of 15,000 legal documents retrieved from *www.rechtspraak.nl* (governmental repository of court verdicts), provided in a structured format (XML) by *Legal Intelligence*. These documents contain various meta-data, including the law area(s), which is the ground truth.

Research Question

To which extent can the law area(s) of a plain text verdict be predicted?

Approach

Preprocessing

The first is preprocessing of the verdict files. From these the 'pure text' verdict and the labels have to be extracted.

This text will then be tokenized. I plan on using the *Frog* NLP module software package for tokenizing. Initially this is all the preprocessing that will be performed. If time permits, I will also look into lemmatizing and stop word removal to improve classification performance (which I presume it will, as it decreases the vector space and sparseness).

Classification and validation

One document can entail multiple law areas which makes this a multi-label classification problem. I plan on using inductive methods only.

I plan on constructing a binary classifier for each of the labels which outputs the probability of the document being of a certain law area (such as Naïve Bayes). A probability cut-off value for assigning each of the labels can then be optimized using the test set (note: not the validation set). If this approach performs poorly, I will look into more specialized multi-label classification methods.

Validation will be performed on the held-out validation set, using a metric such as the *Hamming loss*.