Proposal Automatic Generation of a Legal Thesauris

Joris van Vugt, s4279859
November 4, 2016

Research Question

The goal of the project is to develop a model for automatic thesauris generation of legal documents. Thesauri are used by legal professionals to aid the extraction of relevant documents with specialized search engines. Manually building and maintaining a thesauris is a labor-intensive process. Automatic thesauris generation is thus very useful, but often comes at the cost of lower quality. In this project, I will define and validate an approach for thesauris generation using modern machine learning techniques. My research question is: *Can modern machine learning techniques improve the quality of automatically generated thesauri?*

Methods

Recognizing import concepts and their relations is an unsupervised learning task, since the semantic structure has to be inferred from unlabeled text. I will initially experiment with Word2Vec, which has become very popular over the past few years. Additionally, I will use latent semantic analysis with singular value decomposition as a baseline.

Preprocessing is necessary to avoid relating conjugations of words. Hence, I will use lemmatization to prevent this problem and reduce the dimensionality of the data.

Data set

I will use the provided data set, which consists of 15.683 public Dutch legal documents from www.rechtspraak.nl in XML format.