

# Text Mining of Supreme Administrative Court Jurisdictions

Ingo Feinerer and Kurt Hornik

## Introduction

Investigation of jurisdictions is a fundamental part in jurisprudence since convictions give insight to law interpretations.

Text mining has become a major factor in business intelligence.

Automated textual analysis of law corpora (Conrad et al., 2005) and international court jurisdictions (Schweighofer, 1999).

## Overview

Dataset by tax law experts for Oesterreichische Nationalbank project.

Perform clustering and classification tasks validated against results by law experts.

Automatic derivation of document properties (like senate size).

## Oesterreichische Nationalbank (OeNB) project

Tax law experts (Nagel and Mamut, 2006) supported by a OeNB grant investigate legal norm changes.

Achatz et al. (1987) analysed Austrian supreme administrative court jurisdictions in 1980s. Nagel and Mamut (2006) compare their results and trends to jurisdictions from 2000–2004.

Unveil quality of executive and juristic authorities (e.g., complex law leads to more convictions per subject) or discriminations.

## Dataset

994 text documents containing jurisdictions in German language.

Obtained through the legal information system (Rechtsinformationssystem, <http://ris.bka.gv.at/>) of the Republic of Austria coordinated by the Austrian Federal Chancellery.

HTML documents without explicit metadata.

Four splits of about 250 documents.

## Data Preparation

Removal of malformed HTML tags and extra white space.

Custom parsing function in extension to text mining infrastructure provided by tm in R.

Stemming via Snowball package.

## Clustering Jurisdiction Documents

Creation of a term-document matrix with two weightings:

**Term Frequency** counting the frequency of each term in the documents:  $tf_{t,d}$ ,

**Term Frequency Inverse Document Frequency** normalizing the term frequencies under consideration of the number of all documents:  $tf_{t,d} \cdot \log_2 \frac{n_d}{df_t}$ .

Large matrices with up to  $250 \times 18500$  entries.

## *k*-means Clustering

Domain specific background suggests  $k = 3$  since the grouping creates clusters identifying two concepts and a third capturing the remaining effects.

Clusters can be interpreted modelling *income tax*, *value-added tax* and *none*. Annotations by tax experts are used for validation.

Dataset Split	<i>tf</i> Rand	<i>tf-idf</i> Rand
Split 1	0.49	0.48
Split 2	0.51	0.51
Split 3	0.55	0.54
Split 4	0.53	0.52
Average	0.52	0.51

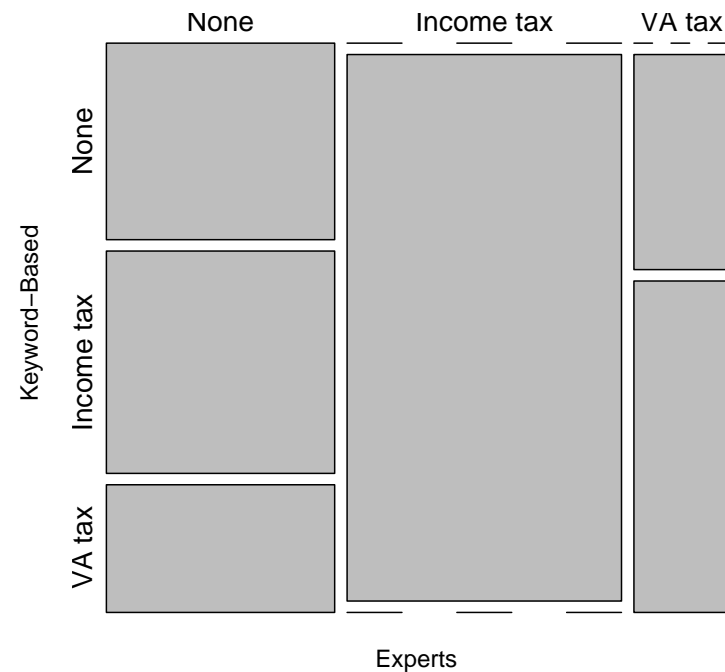
## Keyword Based Clustering

Simulates the behaviour of tax law students preprocessing the documents.

Preprocessors skim over the text looking for discriminative terms (e.g., “income”, “income tax”, . . . , for the cluster identifying the income tax grouping).

Rand index of 0.66.

Agreement plot of the contingency table between the keyword based clustering results and the expert rating:



## Classification with String Kernels

Promising results of string kernels in text classification (Lodhi et al., 2002) and text clustering (Karatzoglou and Feinerer, 2007).

We use a full string kernel

$$k(x, y) = \sum_{s \in \Sigma^*} \lambda_s \cdot \nu_s(x) \cdot \nu_s(y)$$

for “C-svc” classification according to federal due regulations with support vector machines in  $\mathbb{R}$ .

Rand index of 0.49, very long running time.

## Classification with Term-Document Matrices

	<i>tf</i>	<i>tf-idf</i>
Rand	0.59	0.61

Despite large term-document matrices far better performance.

Reasonable results useful for a preclassification to be investigated by specialised law experts.

## Deriving the Senate Size

Formulations are quite standardised:

*Der Verwaltungsgerichtshof hat durch den Vorsitzenden Senatspräsident Dr. Weiss und die Hofräte Mag. Heinzl, Dr. Zorn, Dr. Robl und Dr. Bässer als Richter, im Beisein der Schriftführerin Dr. Doralt, über die Beschwerde des . . . , zu Recht erkannt.*

Monotonous, time consuming, and expensive task.

Investigate punctuation marks and copula phrases to derive the senate size.

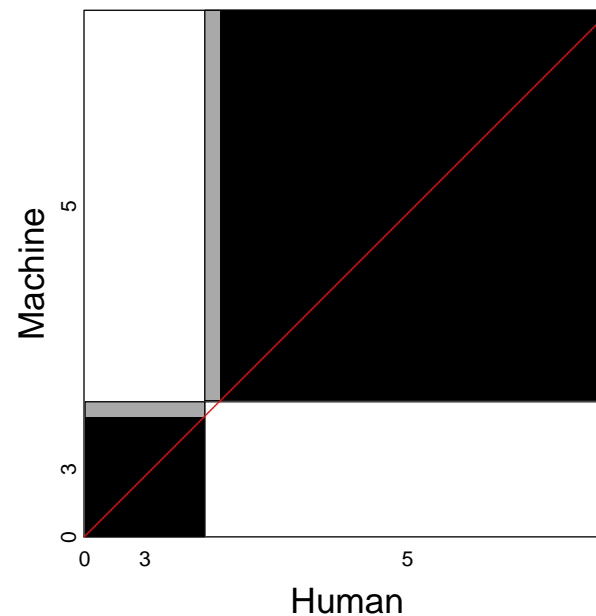
Number of jurisdictions ordered by senate size obtained by fully automated text mining heuristics. The percentage is compared to the percentage identified by humans:

Senate size	0	3	5	9
Documents	0	255	739	0
Percentage	0.000	25.654	74.346	0.000
Human Percentage	2.116	27.306	70.551	0.027

## Text Mining of Supreme Administrative Court Jurisdictions

---

Mosaic plot of the contingency table between the senate size reported by text mining heuristics and the senate size reported by humans (Rand index of 0.94):



## Conclusion

Presented clusterings and classifications for supreme court jurisdictions.

Keyword based clustering works well for text corpora with well defined formulations (i.e., common expressions or keywords).

Useful for determining reasonable working sets to be investigated by law experts.

Classical term-document matrix approaches work better than string kernels in our context.

Deriving of specialised properties works very well.

## Outlook

Improve algorithms for deriving the senate size, either,

**rule-based** with a set of (learning, i.e., adapting) rules, or

**heuristics-based** with algorithms given probabilities and confidence intervals for their answers,

such that metadata with very high accuracy can automatically be produced.

Achatz et al. (1987) originally investigated dozens of tax questions which would be interesting to be considered in further tests.

## Coordinates

Ingo Feinerer and Kurt Hornik  
Department für Statistik und Mathematik  
Wirtschaftsuniversität Wien  
Augasse 2–6, A-1090 Wien

Tel: +43/1/313-36x4756  
Fax: +43/1/313-36x774  
Email: {h0125130|Kurt.Hornik}@wu-wien.ac.at