D20 – Method for creating and describing clusters (EN)

Project funded by the European Union – Next GenerationEU through the Recovery and Resilience Plan of the Slovak Republic under project No. 09I05-03-V02-00049

Financované Európskou úniou NextGenerationEU
PLÁN OBNOVY
ÚRAD PODPREDSEDU VLÁDY SLOVENSKEJ REPUBLIKY PRE PLÁN OBNOVY A ZNALOSTNÚ EKONOMIKU
VAIA VÝSKUMNÁ A INOVAČNÁ AUTORITA
VÝSKUMNÁ AGENTÚRA

INTRODUCTION

The project "**Automation of Legal Text Analysis Based on Machine Learning"** (hereinafter referred to as "ALTAML") represents a key initiative aimed at integrating innovative approaches in the field of data processing and subsequent analysis, specifically within the legal domain, which also includes the law of information and communication technologies. The objective of this project is to develop and validate effective methods for the automated analysis of legal texts using machine learning techniques. This includes, in particular, the development of tools that facilitate the processing and analysis of large volumes of data in the form of various legal documents, as well as the extraction of relevant information (attributes) from such documents, including the identification of key terms, references to other legal acts, and other attributes.

The ALTAML project thus aims to contribute to a more efficient access to legal information and to the acceleration of legal processes, thereby ensuring a higher degree of legal certainty and improving the accessibility of legal texts, the results of their analysis, and relevant legal information for both professionals and the general public.

Within Work Package KPB4-3, Deliverable Method for creating and describing clusters was developed. This document presents notes on a graph-based computational framework for categorizing court decisions, specifically targeting cases labeled as "Others" in the Slovak Republic judicial collection. The methodology utilizes k-partite graphs to model the intricate relationships between court decisions, legal references, and keywords. By applying the Louvain method for community detection and modularity optimization, the system identifies thematic clusters that emerge organically from the data. This approach enhances the organization and accessibility of legal information, transforming unstructured judicial data into a connected and searchable ecosystem.

1. Graph-Based Decisions Categorizer

The Decisions Categorizer is a Python-based project designed to process court decisions, construct multi-partite graph representations, and identify latent legal domains through community detection. The following sections describe the main functionalities, graph construction methodology, and experimental results.

Financované Európskou úniou NextGenerationEU | PLÁN OBNOVY | ÚRAD PODPREDSEDU VLÁDY SLOVENSKEJ REPUBLIKY PRE PLÁN OBNOVY A ZNALOSTNÚ EKONOMIKU | VAIA | VÝSKUMNÁ AGENTÚRA

2. Features

The key functionality of the project is summarized in several core processes:

- **Data Preparation:** Extract legal references using regular expressions and similarity metrics such as Levenshtein distance, and derive keywords via unsupervised methods like TF-IDF and TextRank.
- **Three-Partite Graph Construction:** Capture relations between court decisions (CD), legal references (LR), and keywords (KW) into a unified graph G.
- **Weight Normalization:** Prioritize rare legal references and keywords by downscaling common terms using global usage statistics.
- **Graph Projection**: Transform the three-partite structure into a weighted courtdecision-to-court-decision graph where edges represent thematic similarity.
- **Community Detection:** Partition the graph into distinct legal clusters using the Louvain algorithm to maximize modularity.

3. Requirements

Before deployment, the system requires a structured dataset of judicial decisions and specific computational modules for graph analysis. The project relies on Python and the NetworkX library for graph operations.

Required components:

- Python 3.x and NetworkX (for graph construction and community detection)
- Legal Reference Extractor (based on regular expressions and Levenshtein distance)
- Keyword Extractor (TF-IDF or TextRank implementations)

Financované Európskou úniou NextGenerationEU | PLÁN OBNOVY | ÚRAD PODPREDSEDU VLÁDY SLOVENSKEJ REPUBLIKY PRE PLÁN OBNOVY A ZNALOSTNÚ EKONOMIKU | VAJA | VÝSKUMNÁ AGENTÚRA

4. Graph Construction and Weights

The graph G, consisting of court decisions, legal references, keywords, and edges, captures the multi-dimensional nature of legal texts. Weights are calculated to ensure that significant but rare citations carry more weight than common ones.

Weight calculation for court decision to legal reference edges:

The weight between a court decision and a legal reference is calculated as:
Weight equals the local citation count of the reference within a decision divided by the global citation count of that reference across the entire dataset.
Where:

- The local weight is the number of times a legal reference appears in a specific decision.
- The global weight is the total number of times the same legal reference appears across all decisions.

All weights are normalized per decision so that the sum of weights for a decision equals one.

5. Community Detection (Louvain Method)

The system identifies clusters where connections within communities are denser than connections between communities.

Phase 1: Nodes are moved between communities to maximize modularity gain.

Phase 2: Communities are collapsed into super-nodes, and the process is repeated recursively.

Modularity: Measures the quality of the partition and typically ranges from minus one to one.

6. Experimental Results

An analysis of one thousand court decisions in the Criminal Law domain yielded fourteen distinct communities with a modularity score of 0.53.

Financované Európskou úniou NextGenerationEU | PLÁN OBNOVY | ÚRAD PODPREDSEDU VLÁDY SLOVENSKEJ REPUBLIKY PRE PLÁN OBNOVY A ZNALOSTNÚ EKONOMIKU | VAJA | VÝSKUMNÁ AGENTÚRA

Selected Cluster Examples:

Community C1 (55 cases)
Key law references: Section 415 (Conditional release), Section 37 (Retaliation)
Key keywords: student, minor, school, family

Community C2 (34 cases)
Key law references: Section 289 (Addictive substances), Section 61 (Prohibition)
Key keywords: health, test, vehicles, competence

Community C3 (21 cases)
Key law references: Section 352 (Forgery), Section 394 (Renewal of proceedings)
Key keywords: forgery, justification, complaints

Community C4 (71 cases)
Key law references: Section 117 (Prison sentence), Section 469 (Expunging conviction)
Key keywords: sanctions, security, intervention

7. Source Files

The project's source code and dataset are maintained in the official repository. It can be accessed at the following address:
https://github.com/simonHorvat/GCDC

Financované Európskou úniou NextGenerationEU
PLÁN OBNOVY
ÚRAD PODPREDSEDU VLÁDY SLOVENSKEJ REPUBLIKY PRE PLÁN OBNOVY A ZNALOSTNÚ EKONOMIKU
VAIA VÝSKUMNÁ A INOVAČNÁ AUTORITA
VÝSKUMNÁ AGENTÚRA