Тематический поиск схожих дел в коллекции актов арбитражных судов*

Герасименко Н. А.[1]***, Артёмова Е. Л.***[2]***, Воронцов К. В.***[3]
nikgerasimenko@gmail.com
[1]МАИ; [2]НИУ ВШЭ; [3]МФТИ

В работе рассматривается задача информационного поиска по коллекции актов арбитражных судов. В качестве запроса поисковой системе может выступать любой документ коллекции. В ответ на поисковый запрос генерируется список документов коллекции, ранжированный по убыванию релевантности. Для решения поставленной задачи построена тематическая модель коллекции актов арбитражных судов с помощью открытой библиотеки BigARTM. При построении модели учтена специфика предметной области: добавлены модальность ссылок на нормативно-правовые акты, а также модальность юридических терминов, выделенных полуавтоматически, с использованием алгоритма TopMine.

Ключевые слова: *LegalTech, BigARTM, тематическое моделирование, аддитивная регуляризация мультимодальных иерархических тематических моделей.*

1 Введение

Целью данной работы является построение тематической модели коллекции актов арбитражных судов для осуществления разведочного поиска по ней.

Специалисты в области юриспруденции регулярно сталкиваются в своей работе с необходимостью поиска документов в базах судебной практики. Зачастую они ищут дела, схожие с теми, над которыми работают, не всегда при этом зная, по какому запросу или в какой именно тематике искать. Такого рода поиск называется разведочным [1], для его осуществления разработано множество методов. В данной работе сделана попытка применения одного из методов разведочного поиска, основанного на построении тематической модели, к задаче поиска по коллекции юридических документов.

Для построения тематической модели используется библиотека BigARTM [2]. Одним из ее преимуществ является возможность использования механизма модальностей, с помощью которого сделана попытка учесть специфику предметной области, выделив ссылки на нормативно-правовые акты и юридические термины. Задача выделения именованных сущностей решается с помощью регулярных выражений и алгоритма TopMine для выделения коллокаций.

Исследование основано на подходе, описанном в работе [1], где он использован при построении тематической модели для разведочного поиска по новостям технологий на порталах Habrahabr.ru и TechCrunch.com. Построенная авторами модель показала хорошие результаты на размеченных с помощью ассесоров данных.

*Задачу поставил: Воронцов К. В. Консультант: Артёмова Е. Л.

Литература

[1] Anastasia Ianina, Lev Golitsyn, and Konstantin Vorontsov. Multi-objective topic modeling for exploratory search in tech news. In *Communications in Computer and Information Science*, pages 181–193. Springer International Publishing, nov 2017.

[2] Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, and Marina Dudarenko. Bigartm: open source library for regularized multimodal topic modeling of large collections. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 370–381. Springer, 2015.