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

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. Задача, которую решают специалисты называется задачей информационного поиска, которая может быть решена методами тематического разведочного поиска.

Тематический разведочный поиск - это парадигма поиска, в которой в качестве запроса может выступать целый документ или даже коллекция документов.

Юридические тексты обладают особой спецификой и, не смотря на то, что могут иногда казаться обычными текстами на естественном языке, не могут в полной мере рассматриваться таким образом. Понимание истинного смысла юридического текста зачастую требует консультации с экспертом.

В ходе совместной работы с экспертами в области юриспруденции был выявлен ряд важных модальностей, без учета которых невозможно построение адекватной модели. Библиотека BigARTM [3], с помощью которой строится тематическая модель, обладает возможностями для учета этих модальностей. В данной работе учтена модальность ссылок на нормативно-правовые акты и модальность юридических терминов.

Данное исследование основано на подходе, описанном в работе [2], где он использован при построении тематической модели для разведочного поиска по статьям на порталах Habrahabr.ru и TechCrunch.com. Построенная авторами модель показала хорошие результаты на размеченных с помощью ассесоров данных.

2 Постановка задачи

Дана коллекция, состоящая из <n> юридических документов, актов арбитражных судов. Для каждого документа известна его тематика, которая не тождественна теме в тематической модели, а носит более общий характер: тематика состоит из набора тем.

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

Задача состоит в построении тематической модели данной коллекции с помощью библиотеки BigARTM. Проанализировав структуру документов, требуется выделить из них, при помощи контекстно-свободных грамматик и регулярных выражений, ссылки на нормативно-правовые акты (НПА). Также требуется выделить юридические темины с помощью алгоритма TopMine [1]. Ссылки на НПА и юридические термины будут учтены в модели в качестве модальностей.

Предварительная оценка модели будет проводиться по критериям перплексии, разреженности распределений тем в документах, а также разреженности распределений токенов в темах для модальностей ссылок на НПА и юридических терминов [4].

Окончательная оценка будет строиться следующим образом. Полученные в результате моделирования тематические вектора для каждого документа кластеризуются при помощи алгоритма k-means. Затем, по критерию Rand Index делается оценка, в какой степени картина кластеризации соответствует принадлежности документов их тематикам: оказались ли в одном кластере документы из одной тематики, и оказались ли документы из разных тематик в разных кластерах.

Таким образом, формальная постановка задачи может быть сформулирована в следующем виде. Пусть D - коллекция документов, $|D| = n$. Каждый документ $d \in D$ принадлежит одной из тематик $T_i \in T$. С другой стороны, каждый документ d принадлежит одному из кластеров $Y_j \in Y$, полученных в результате применения алгоритма k-means к множеству тематических векторов документов коллекции. Требуется построить тематическую модель таким образом, что

$$\frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}} \to min, \tag{1}$$

где:
$a = |S^a|,\ where\ S^a = \{(o_i, o_j) \mid o_i, o_j \in Y_k, o_i, o_j \in T_l\}$,
$b = |S^b|,\ where\ S^b = \{(o_i, o_j) \mid o_i \in Y_{k_1}, o_j \in Y_{k_2}, o_i \in T_{l_1}, o_j \in T_{l_2}\}$,
$c = |S^c|,\ where\ S^c = \{(o_i, o_j) \mid o_i, o_j \in Y_k, o_i \in T_{l_1}, o_j \in T_{l_2}\}$,
$d = |S^d|,\ where\ S^d = \{(o_i, o_j) \mid o_i \in Y_{k_1}, o_j \in Y_{k_2}, o_i, o_j \in T_l\}$,
$1 \leqslant i, j \leqslant n, i \neq j, 1 \leqslant k, k_1, k_2 \leqslant r, k_1 \neq k_2, 1 \leqslant l, l_1, l_2 \leqslant s, l_1 \neq l_2$.

Литература

[1] Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare Voss, and Jiawei Han. Scalable topical phrase mining from text corpora. *Proceedings of the VLDB Endowment*, 8(3):305–316, 2014.

[2] 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.

[3] 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.

[4] Konstantin Vorontsov and Anna Potapenko. Additive regularization of topic models. *Machine Learning*, 101(1-3):303–323, 2015.