Тематический поиск схожих дел в коллекции актов арбитражных судов

Герасименко Николай Александрович

Московский авиационный институт

Курс: Численные методы обучения по прецедентам (практика, В. В. Стрижов)/Группа 8О-103М, весна 2019

Цель исследования

Цель работы

Решение задачи информационного поиска по коллекции актов арбитражных судов.

Проблема

Специалисты в области юриспруденции вынуждены тратить большое количество времени на поиск релевантной судебной практики.

Метод решения

Построение тематической модели коллекции с помощью открытой библиотеки BigARTM, реализующей вероятностное тематическое моделирование на основе аддитивной регуляризации.

Литература

Теория APTM:

- Konstantin Vorontsov and Anna Potapenko. Additive regularization of topic models. Machine Learning, 2015.

Библиотека BigARTM:

- 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.

Опорное исследование:

- 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.

Постановка задачи

Дано

Коллекция текстовых документов D

- n_{dw} - частоты терминов w в документах коллекции $d \in D$.

Найти

Параметры тематической модели $p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$

- ϕ_{wt} - вероятности терминов w в темах $t \in T$.
- θ_{td} - вероятности тем t в каждом документе $d \in D$.

Максимизируя логарифм правдоподобия с регуляризаторами:

$$\sum_{d,w} n_{dw} ln \sum_{t \in T} \phi_{wt}\theta_{td} + \sum_{i} \tau_i R_i(\Phi, \Theta) \to \max$$

Внешний критерий качества

Используемая информация

Для оценки качества построения векторных представлений документов коллекции используем информацию о принадлежности каждого документа определенной категории дел таких как, например, «Особенности банкротства отдельных категорий должников» или «Внешнее управление».

Идея

Рассчитав согласованность картины кластеризации векторных предствлений документов с картиной принадлежности документов их категориям, получаем оценку адекватности векторных предствлений.

Критерии согласованности

- Adjusted Rand Index (ARI).
- Adjusted Mutual Information (AMI).

Необходимость регуляризаторов

Оптимизационная задача в APTM

В соответствии с теорией аддитивной регуляризации тематических моделей для каждой модальности вводится критерий логарифма правдоподобия и с помощью EM-алгоритма максимизируется их взвешенная сумма.

Регуляризаторы R_i

Добавляются к сумме как дополнительные критерии с весами τ_i, необходимые поскольку в общем случае задача имеет бесконечно много решений.

$$\sum_{d,w} n_{dw} ln \sum_{t \in T} \phi_{wt} \theta_{td} + \sum_i \tau_i R_i(\Phi, \Theta) \to \max$$

Используемые регуляризаторы
Регуляризаторы сглаживания и разреживания

Общий вид

Данные два регуляризатора имеют одинаковый вид и отличаются только знаками коэффициентов α и β, для регуляризатора разреживания они отрицательны.

$$R(\Phi, \Theta) = \beta \sum_{m \in M} \sum_{t \in T} \sum_{w \in W^m} \beta_w ln \phi_{wt} + \alpha \sum_{d \in D} \sum_{t \in T} \alpha_t ln \theta_{td} \to \max.$$

Назначение

Регулятор сглаживания вводит в модель требование схожести распределений ϕ_{wt} с распределением β_w и θ_{td} с распределением α_t. Регуляризатор разреживания, в свою очередь, способствует появлению нулевых элементов в распределенях ϕ_{wt} и θ_{td}, что позволяет находить более компактные представления документов.

Используемые регуляризаторы
Регуляризатор декоррелирования

Общий вид

$$R(\Phi) = -\tau \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \phi_{wt}\phi_{ws} \to max.$$

Назначение

Вводит в модель требование различности тем путем минимизации ковариации между столбцами матрицы Ф. Также побочным эффектом работы регуляризатора декоррелирования является разреживание матрицы Ф, поэтому в случае его применения, можно не применять регуляризатор разреживания для нее.

Стратегия регуляризации

Общий подход к подбору параметров

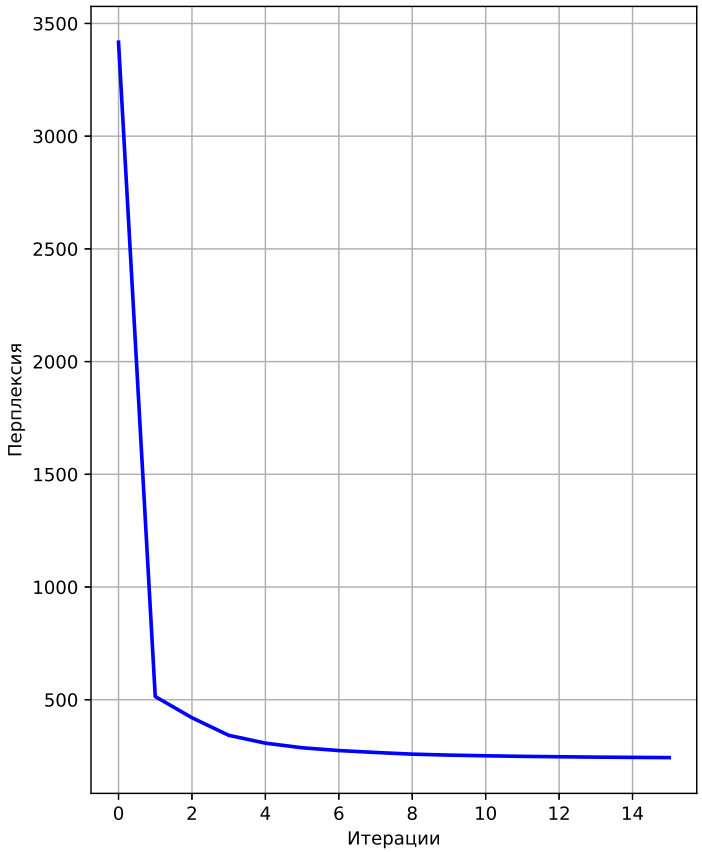
Выбор параметров модели производился путем перебора значений по сетке с использованием набора критериев качества: перплексия, разреженность распределений токенов в темах. Также при оценке учитывался внешний критерий качества.

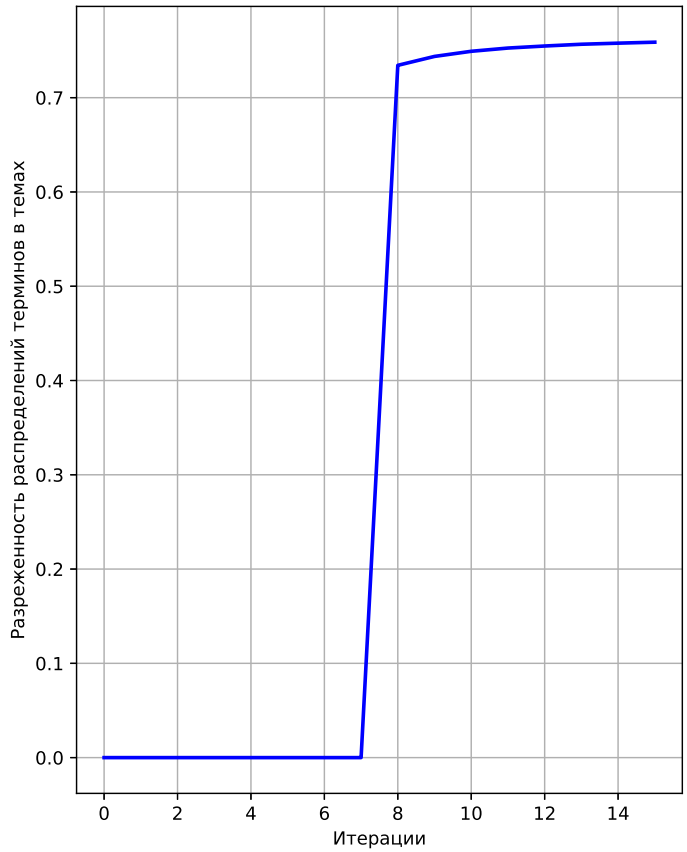
Порядок добавления регуляризаторов и выбор параметров

Регуляризатор декоррелирования матрицы Φ добавлялся в модель изначально, а спустя 8 итераций EM-алгоритма добавлялся регуляризатор разреживания для матрицы Θ. Для каждого значения параметра регуляризации производилось по 8 итераций EM-алгоритма, после чего выбиралось то значение, при котором модель улучшилась по одному или нескольким критериям качества, существенно не ухудшившись ни по одному из них.

Результаты вычислительного эксперимента
Внутренние критерии качества.

Зависимости перплексии и разреженности распределений терминов в темах





Результаты вычислительного эксперимента
Внешние критерии качества.

Сравнение с базовыми алгоритмами

Для оценки качества модели были использованы в качестве базовых алгоритмов

- TF-IDF по словам
- TF-IDF по выделенным из текста ссылкам на нормативно-правовые акты (НПА), например, «пункт 3 статьи 6 УК РФ».

Модель	ARI	AMI
TF-IDF по словам	10%	12.5%
TF-IDF по ссылкам на НПА	17%	22%
APTM	37.5%	42%

Заключение

- При помощи библиотеки BigARTM построена тематическая модель, строющая для документов коллекции сжатые векторные представления и, таким образом, позволяющая реализовать поиск, используя косинусную меру близости.
- Был реализован метод внешней оценки качества тематической модели, заключающийся в вычислении согласованности между картиной кластеризации тематических векторов документов коллекции и их принадлежности различным категориям дел.
- Результаты показывают, что тематическая модель способна корректно строить векторные представления юридических документов, что, в перспективе, дает возможность построения системы разведочного поиска, с использованием тематического моделирования в качестве ключевой технологии.