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 $d \in D$. Термины относятся к одному из трех типов, называемых модальностями: слово естественного языка, ссылка на НПА, юридический термин.

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 $d \in D$.

Поставлена задача стохастического матричного разложения.

Необходимость регуляризаторов

Оптимизационная задача в APTM

Для каждой модальности вводится критерий логарифма правдоподобия и с помощью EM-алгоритма максимизируется их взвешенная сумма.

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

Регуляризаторы R_i

Добавляются к сумме как дополнительные критерии с весами τ_i, являющимися гиперпараметрами модели. Регуляризаторы необходимы, поскольку в общем случае задача имеет бесконечно много решений.

Используемые регуляризаторы

Регуляризаторы сглаживания и разреживания

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

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

Регуляризатор декоррелирования

Вводит в модель требование различности тем путем минимизации ковариации между столбцами матрицы Φ.

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

Внешний критерий качества

Идея

Рассчитаем согласованность картины кластеризации векторных предствлений документов с известной нам картиной классификации документов по категориям. В качестве критериев согласованности используем критерии ARI и AMI.

Adjusted Rand Index (ARI)

$$ARI(U,V) = \frac{RI - E\{RI\}}{\max\{RI\} - E\{RI\}}$$

Adjusted Mutual Information (AMI)

$$AMI(U,V) = \frac{MI(U,V) - E\{MI(U,V)\}}{\max\{H(U),H(V)\} - E\{MI(U,V)\}}$$

Внутренние критерии качества

Перплексия языковой модели

Мера неопределенности терминов в тексте.

$$\exp\left(-\frac{1}{n}\sum_{d,w} n_{dw} \ln p(w|d)\right), \quad n = \sum_{d,w} n_{dw}$$

Разреженность распределений терминов в темах

Отвечает гипотезе разреженности.

Равен количеству элементов матрицы Φ, меньших заранее заданного порога $\epsilon = \text{const}$.

Стратегия регуляризации

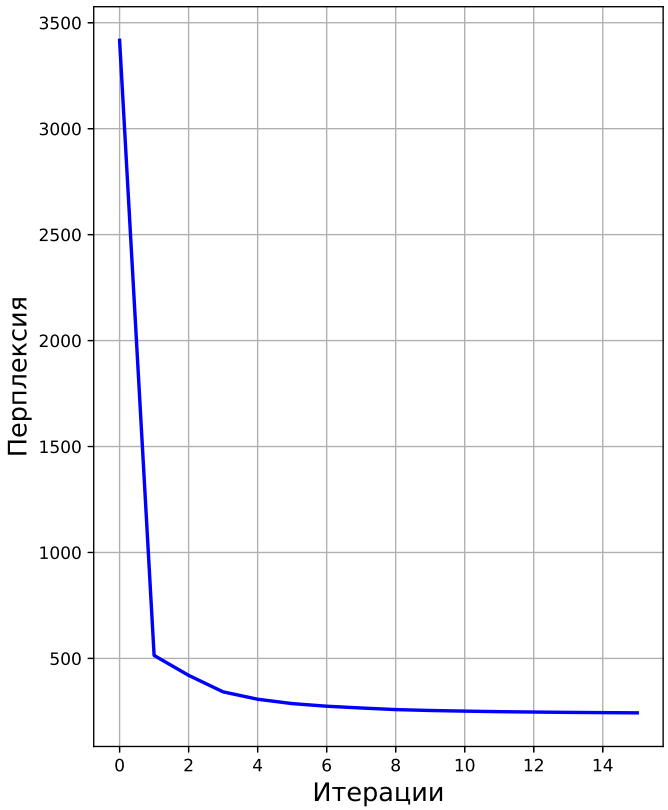
Порядок добавления регуляризаторов и выбор гиперпараметров

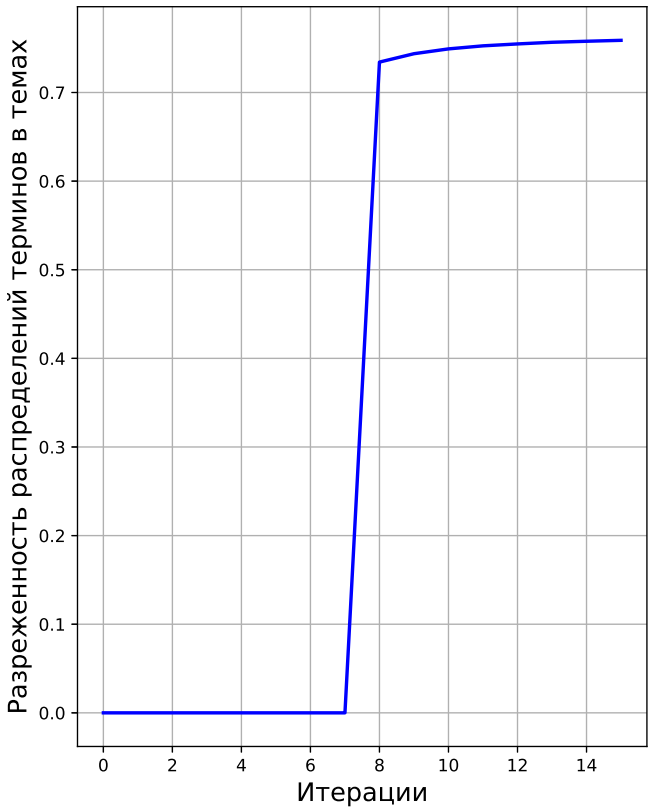
1. Производим 8 итераций EM-алгоритма с регуляризатором декоррелирования для каждого значения коэффициента регуляризации декоррелирования из сетки значений. Выбираем лучший коэффициент регуляризации для регуляризатора декоррелирования.
2. Добавляем регуляризатор разреживания распределений тем в документах. Производим еще 8 итераций EM-алгоритма с обоими регуляризаторами и выбираем лучшее значение коэффициента регуляризации.

Критерии качества

- Перплексия,
- Разреженность распределений терминов в темах.
- Солгласованность с известной картиной классификации.

Зависимость перплексии и разреженности матрицы Φ от количества итераций





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%
BigARTM	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.