

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 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.

Теория АРТМ:

- 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:

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

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 представлений документов с картиной принадлежности документов их категориям, получаем оценку адекватности векторных представлений.

Критерии согласованности

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

Необходимость регуляризаторов

Оптимизационная задача в АРТМ

В соответствии с теорией аддитивной регуляризации тематических моделей для каждой модальности вводится критерий логарифма правдоподобия и с помощью ЕМ-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.

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

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.

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

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 распределениях ϕ_{wt} и θ_{td} , что позволяет находить более компактные представления документов.

Используемые регуляризаторы

Регуляризатор декоррелирования

Общий вид

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

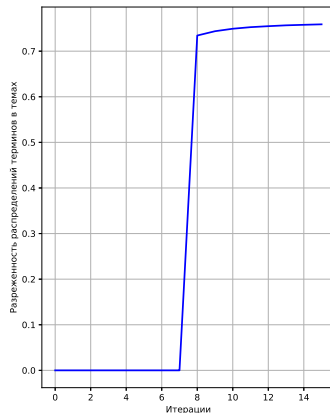
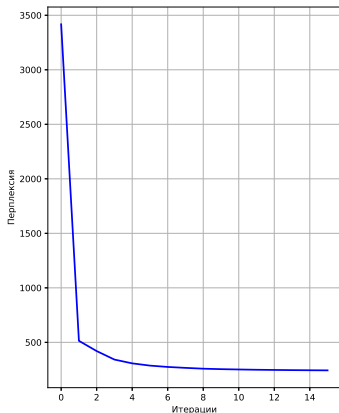
Назначение

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

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%

