Тематический поиск в коллекции юридических документов

Герасименко Николай Александрович

Московский авиационный институт (национальный исследовательский университет)
Институт №8 «Информационные технологии и прикладная математика»
Кафедра 810Б «Информационные технологии в моделировании и управлении»
Магистерская программа «Машинное обучение и управление большими данными»

Научный руководитель: д.ф.-м.н. Абгарян К. К.,
Научный консультант: профессор РАН, д.ф.-м.н. Воронцов К. В.

Москва
2020 г

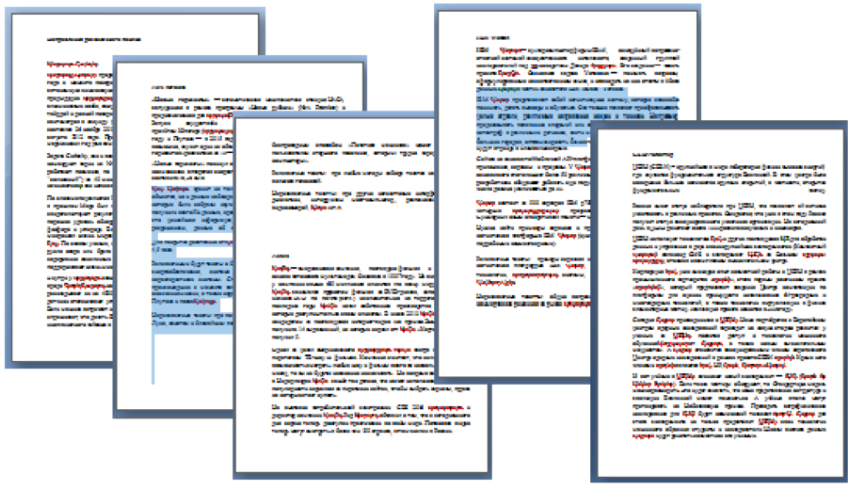
Цель и задачи исследования

Проблема: традиционные системы поиска по короткому запросу и ключевым словам (known-item search) не позволяют специалисту в области юриспруденции описать описать все характеристики дела, для которого идет поиск релевантной практики.

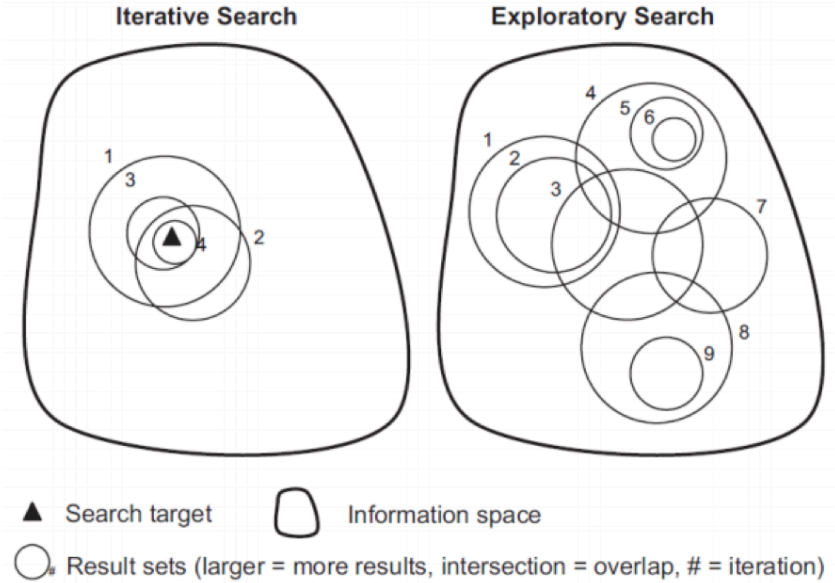
Цель: построить систему разведочного поиска, допускающую использование длинных специализированных запросов, например, целых документов. В качестве ключевой технологии при построении системы предлагается использовать тематическое моделирование с аддитивной регуляризацией (АРТМ).

Задачи

1. Определить совместно со специалистами в предметной области дополнительные признаки, которые могут улучшить качество поиска.
2. Организовать экспертную разметку поисковой выдачи репрезентативного набора запросов для последующей оценки качества поиска.
3. Обучить тематическую модель коллекции юридических документов.
4. Оценить качество поиска с помощью тематической модели и сравнить его с результатами других подходов.



Запросы для разведочного поиска



Запрос

1-2 страничный документ с описанием поисковой задачи. В частности, один из документов коллекции, в которой идет поиск.

Результат поиска

Набор релевантных документов, изучение которых поможет юристу получить представление о том, какая юридическая практика складывается для дел, похожих на то, над которым он работает.

Алгоритм тематического поиска

Дано:

- коллекция документов D.
- множество запросов Q.

Алгоритм:

1. Обучить тематическую модель коллекции D.
2. Получить с помощью обученной тематической модели векторные представления документов коллекции D.
3. Получить с помощью обученной тематической модели векторные представления документов-запросов Q.
4. Пользуясь косинусной мерой близости, найти k ближайших документов коллекции D для каждого запроса из Q.

Критерии качества:

- Precision@k - доля релевантных документов среди первых k найденных
- Recall@k - доля k первых найденных релевантных документов среди всех релевантных.

Vorontsov, K., Ianina, A. Multimodal topic modeling for exploratory search in collective blog. In Intelligent Data Processing: Theory and Applications: Book of abstracts of the 11th International Conference (pp.186-187).

Модальности терминов

Мультимодальные тематические модели позволяют учитывать не только модальность текста, но и дополнительные данные как токены других модальностей.

Дано

- W^m – словарь токенов модальности $m \in M$
  $W = W^1 \cup ... \cup W^M$ – объединенный словарь,
- D – коллекция текстовых документов $d = \{w_1, ..., w_{n_d}\}$,
- n_d – длина документа d, n_{dw} – частота термина w в документе d.

Предположения:

- каждый термин $w \in W$ в $d \in D$ имеет тему $t \in T$;
- $D \times W \times T$ – дискретное вероятностное пространство;
- Коллекция – это i.i.d. выборка $(d_i, w_i, t_i)_{i=1}^n \sim p(d, w, t)$;
- d_i, w_i – наблюдаемые, темы t_i – скрытые;
- гипотеза условной независимости: $p(w|d, t) = p(w|t)$.

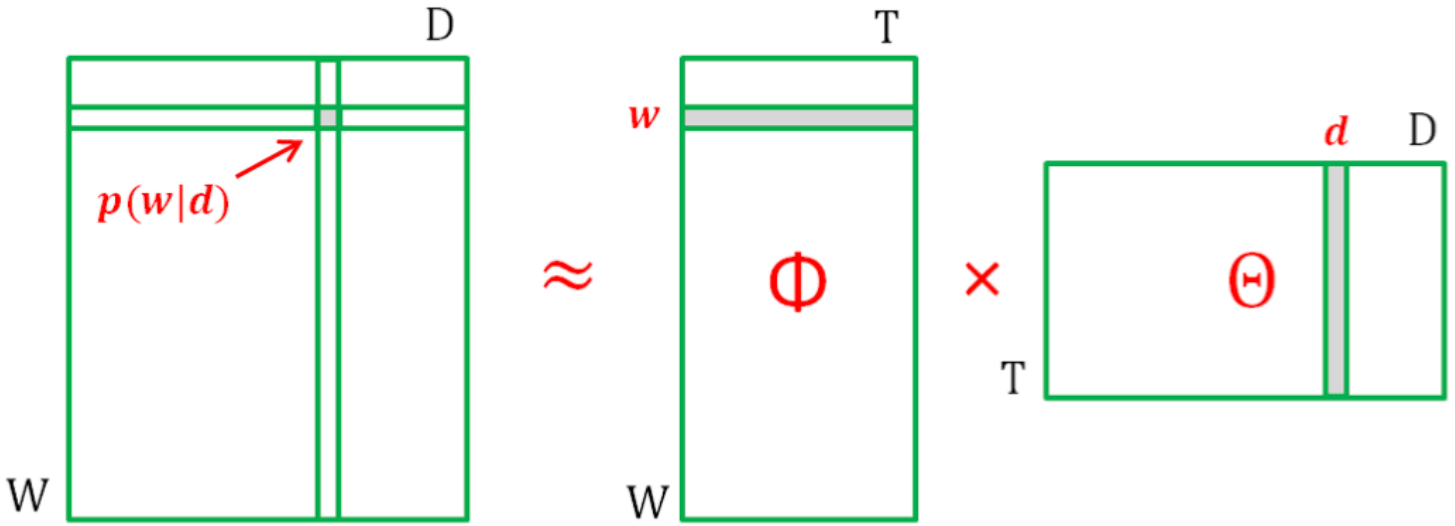
Найти

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

$\phi_{wt} = p(w|t)$ – вероятности терминов w в каждой теме t.

$\theta_{td} = p(t|d)$ – вероятности тем t в каждом документе d.

Поставлена задача стохастического матричного разложения:



Максимизация log правдоподобия с регуляризаторами $R_i(\Phi, \Theta)$

$$\sum_{m \in M} \eta_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + \sum_i \tau_i R_i(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

EM-алгоритм

Метод простой итерации для системы уравнений:

$$\text{E-шаг:} \quad p_{tdw} \equiv p(t|d,w) = \underset{t \in T}{\mathsf{norm}}\big(\phi_{wt}\theta_{td}\big)$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w \in W}{\mathsf{norm}}\big(\sum_d n_{dw} p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\big) \\ \theta_{td} = \underset{t \in T}{\mathsf{norm}}\big(\sum_w n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\big) \end{cases}$$

где $\underset{t \in T}{\mathsf{norm}}(x_t) = \frac{\max\{x_t, 0\}}{\sum_{s \in T} \max\{x_s, 0\}}$ – операция нормирования вектора

Vorontsov K.V., Potapenko A.A. Additive Regularization of Topic Models
Machine Learning Journal, S.I. “Data Analysis and Intelligent Optimization with Applications”

Регуляризатор разреживания распределений тем в документах

используется для выделения относительно небольшой доли предметных тем в каждом из документов. Максимизация расстояний распределений θ_{td} от заданных распределений α_t:

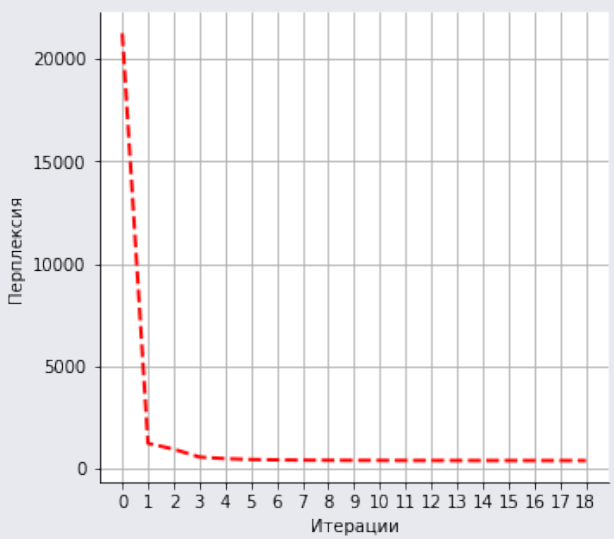
$$\sum_{d \in D} KL(\alpha_t || \theta_{td}) \to \max_{\Theta} \Rightarrow R(\Theta) = -\alpha \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td} \to \max$$

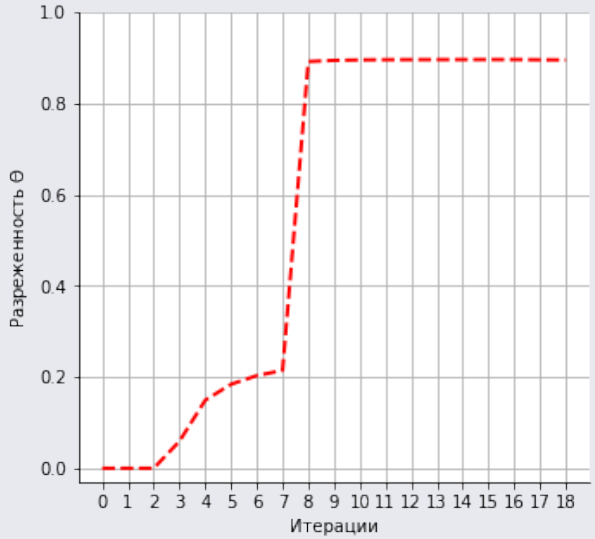
Регуляризатор декоррелирования распределений терминов в темах

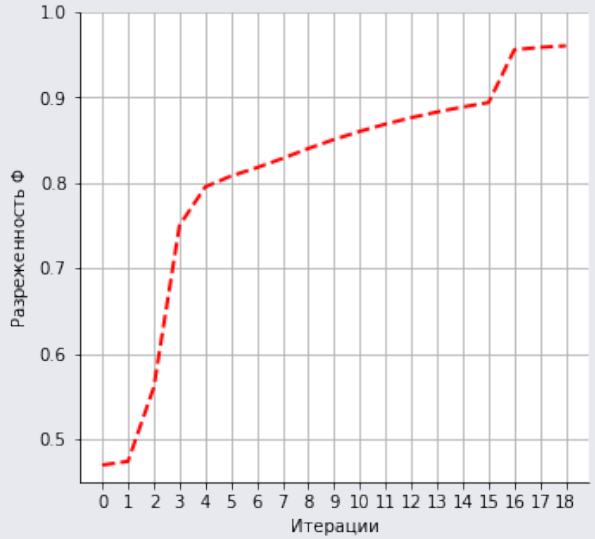
используется для повышения различности лексических ядер предметных тем. Минимизация ковариаций между вектор-столбцами ϕ_t:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \phi_{wt} \phi_{ws} \to \max$$

Зависимость внутренних критериев качества тематической модели от количества итераций EM-алгоритма







Перплексия языковой модели

Мера неопределенности терминов в тексте.

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

Матрица Φ обученной тематической модели

Когда тематическая модель обучилась, столбцы матрицы Φ представляют собой темы коллекции D – дискретные распределения на множестве терминов.

Ниболее вероятные термины нескольких тем

Topic 4	**Topic 11**	**Topic 31**
пожар	участок	убыток
мчс	земельный	нарушить
ппб	собственность	произвести
противопожарный	кадастровый	возмещение
автоматический	администрация	ущерб
недра	предоставление	вред
сигнализация	площадь	причинить
пожаротушение	аренда	потерпевший
полезный	расположить	выплата
эвакуация	здание	транспортный
оповещение	пользование	гражданский

Качество тематического поиска по критериям Precision@k и Recall@k

Рассмотрены модели с разными сочетаниями модальностей (Слова, Ссылки на НПА, Юридические Термины) и разным числом тем.

	Количество тем					**Модальности**			
Метрика	**25**	**50**	**100**	**200**	**300**	**C**	**CH**	**CT**	**TH**
Precision@5	0.7	0.71	0.79	0.75	0.75	0.73	0.78	0.74	0.78
Precision@10	0.79	0.8	0.87	0.84	0.83	0.84	0.86	0.84	0.85
Precision@15	0.85	0.87	0.93	0.91	0.91	0.9	0.89	0.9	0.91
Precision@20	0.9	0.91	0.94	0.93	0.91	0.9	0.91	0.91	0.92
Recall@5	0.09	0.09	0.11	0.1	0.09	0.09	0.1	0.09	0.1
Recall@10	0.13	0.14	0.16	0.16	0.16	0.13	0.13	0.12	0.14
Recall@15	0.16	0.16	0.18	0.16	0.17	0.15	0.16	0.15	0.17
Recall@20	0.2	0.2	0.23	0.21	0.22	0.2	0.21	0.2	0.2

Качество поиска с помощью классической и нейросетевых моделей

Рассмотрены классическая модель Tf-Idf и нейросетевая модель Doc2Vec с разными разными размерностями векторных представлений для документов коллекции D.

	APTM	**PLSA**	**TF-IDF**	**Doc2Vec**	
Метрика	**100**	**100**	**100**	**100**	**200**
Precision@5	0.79	0.71	0.74	0.75	0.75
Precision@10	0.87	0.8	0.83	0.84	0.83
Precision@15	0.93	0.87	0.89	0.91	0.91
Precision@20	0.94	0.91	0.9	0.93	0.91
Recall@5	0.11	0.09	0.08	0.1	0.09
Recall@10	0.16	0.14	0.16	0.16	0.16
Recall@15	0.18	0.16	0.17	0.16	0.17
Recall@20	0.23	0.2	0.21	0.21	0.22

Результаты, выносимые на защиту

- Предложено решение задачи информационного поиска по коллекции актов арбитражных судов, с использованием тематического моделирования в качестве ключевой технологии.
- Для решения поставленной задачи построена тематическая модель коллекции актов арбитражных судов с помощью открытой библиотеки BigARTM. При построении модели учтена специфика предметной области путем добавления в модель модальностей.
- Построенная модель показывает высокую интерпретируемость тем и точность поиска.
- По теме диссертации были опубликованы тезисы в сборнике трудов IV Международной научно-технологической конференции студентов и молодых ученых «Молодежь. Инновации. Технологии» (стр. 5-6).