Московский государственный университет имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математических методов прогнозирования

Исмагилов Тимур Ниязович

Частично обучаемые вероятностные тематические модели коллекций научных текстов

ДИПЛОМНАЯ РАБОТА

Научный руководитель:
д.ф-м.н., доцент
К. В. Воронцов

Москва, 2016

Содержание

Аннотация

Один из наиболее распространенных вариантов использования тематических моделей — анализ текстовых коллекций с целью выявить их внутреннюю семантическую структуру и предоставить пользователю функции визуализации, навигации и тематического поиска внутри коллекции. В данной работе рассматриваются методы решения задачи анализа текстовых коллекций с помощью аддитивной регуляризации тематических моделей (ARTM) и частичного обучения. В ходе работы создан инструмент для интерактивного тематического моделирования, позволяющий эксперту вводить обучающую разметку тем. Разработан набор регуляризаторов частичного обучения, позволяющий использовать данных, полученных от эксперта, при построении тематической модели. На основе этих регуляризаторов разработан эффективный метод тематического поиска, который также может быть использован экспертом для улучшения качества тематической модели. Практическим результатом работы является программа с веб-интерфейсом, которая для произвольной текстовой коллекции позволяет эксперту при небольшом приложении усилий построить максимально качественную тематическую модель с функцией тематического поиска.

1 Введение

1.1 Задача построения тематической модели с частичным привлечением учителя

Тематическая модель коллекции текстовых документов представляет каждый документ как вероятностную смесь тем, каждая из которых представляет собой дискретное распределение на множестве терминов. Таким образом, тематическая модель выступает как средство понимания, систематизации и смыслового поиска в больших текстовых коллекциях. Тематическое моделирование активно развивается в последние годы, пополняясь структурами и новыми алгоритмами построения тематических моделей.

Данная работа рассматривает возможность улучшения качества тематической модели за счет использования знаний экспертов в той предметной области, к которой относится коллекция текстов. Конечной целью исследования является создание инструмента для *интерактивного тематического моделирования*, имеющего следующие функции:

- Построение тематической модели;
- Визуализация тематической модели;
- Навигация по темам, терминам и документам внутри тематической модели;
- Оценка интерпретируемости тематической модели (метрика качества, оценивающая, насколько хорошо выделяемые ей темы интерпретируются экспертами);
- Тематический поиск документов, ближайших к заданному в семантическом смысле;
- Частичное обучение тематической модели, дающее возможность применить знания экспертов для улучшения ее качества.

Таким образом, задача исследования распадается на следующие подзадачи:

- Создание алгоритма частичного обучения тематической модели на основе знаний экспертов;
- Разработка метрики интерпретируемости тематической модели;
- Исследование задачи тематического поиска и разработка оптимального метода поиска;

- Создание инструмента с веб-интерфейсом, воплощающего разработанные алгоритмы и предполагающего использование экспертами.

1.2 Структура работы

Раздел 2 посвящен описанию алгоритма ARTM, который был использован в качестве основы для исследования. Приводятся аргументы в пользу этого выбора.

Раздел 3 описывает визуальный интерфейс, разработанный для использования экспертом при обучении тематической модели.

Раздел 4 посвящен задаче тематического поиска как ключевой для тематической модели коллекции научных текстов.

Раздел 5 описывает разработку регуляризаторов частичного обучения, которые позволяют применить экспертные знания, полученные через визуальный интерфейс, к построению тематической модели.

Раздел 6 продолжает рассмотрение вопроса тематического поиска. Описываются эксперименты по улучшению его качества.

Раздел 7 подводит итоги работы. Формулируются возможности для дальнейших исследований в данной области.

2 Аддитивная регуляризация тематических моделей

Аддитивная регуляризация тематических моделей (ARTM) [1] — подход к построению тематических моделей, основанный на максимизации взвешенной суммы логарифма правдоподобия и дополнительных критериев — регуляризаторов. Преимущество этого подхода в том, что регуляризаторы позволяют формализовать произвольное количество требований к тематической модели, причем необязательно имеющих вероятностную интерпретацию. В частности, многие известные тематические модели могут быть рассмотрены как регуляризаторы в терминах модели ARTM [2]. Таким образом, механизм регуляризаторов модели ARTM представляется наилучшим вариантом интеграции обучающих данных, полученных от экспертов, в построение тематической модели.

2.1 Обозначения

Пусть D — множество (коллекция) текстовых документов, W — множество (словарь) всех употребляемых в них терминов. Терминами могут быть как отдельные слова, так и ключевые фразы. Каждый документ $d \in D$ представляет собой последовательность n_d терминов $(w_1, \ldots, w_{nd})$ из словаря W. Термин может повторяться в документе много раз.

Предполагается, что существует конечное множество тем T, и каждое употребление термина w в каждом документе d связано с некоторой темой $t \in T$, которая не известна. Коллекция документов рассматривается как случайная и независимая выборка троек $(w_i, d_i, t_i), i = 1, \ldots, n$ из дискретного распределения $p(w, d, t)$ на конечном множестве $W \times D \times T$. Термины w и документы d являются наблюдаемыми переменными, тема $t \in T$ является латентной (скрытой) переменной.

Определим матрицы Φ и Θ следующим образом:

$$\Phi = (\varphi_{wt})_{W \times T}, \qquad \varphi_{wt} = \mathsf{p}(w \mid t)$$
$$\Theta = (\theta_{td})_{T \times D}, \qquad \theta_{td} = \mathsf{p}(t \mid d)$$

2.2 Алгоритм

В основе ARTM лежит алгоритм вероятностного латентного семантического анализа (PLSA) [3], в котором матрицы распределения тем для документов Θ и распределения терминов для темы φ вычисляются с помощью EM-алгоритма. Алгоритм основан на формулах для стационарной точки задачи оптимизации логарифма правдоподобия.

Меняя задачу с максимизации правдоподобия на максимизацию суммы критерия правдоподобия $L(\Phi, \Theta)$ и критерия регуляризации $R(\Phi, \Theta)$, получим алгоритм, в котором формулы E и M -шагов меняются на следующие:

$$p_{tdw} = \frac{\varphi_{wt}\theta_{td}}{\sum_{s \in T} \varphi_{ws}\theta_{sd}}$$
$$\varphi_{wt} = \operatorname{norm}_w \left(n_{wt} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}} \right), \qquad n_{wt} = \sum_{d \in D} n_{dw} p_{tdw}$$
$$\theta_{td} = \operatorname{norm}_t \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right), \qquad n_{td} = \sum_{w \in d} n_{dw} p_{tdw}$$

Здесь norm — оператор неотрицательного нормирования, который преобразует произвольный вектор $(x_i)_{i \in I}$ в вектор вероятностей $(p_i)_{i \in I}$ дискретного рас-

пределения путем обнуления отрицательных элементов с последующей нормировкой:

$$p_i = \text{norm}_{i \in I}(x_i) = \frac{\max\{x_i, 0\}}{\sum\limits_{j \in I} \max\{x_j, 0\}}, \text{ для всех } i \in I.$$

В результате получаем алгоритм 2.1.

Алгоритм 2.1 ARTM

Вход: коллекция документов D, число тем $|T|$;
Выход: φ, Θ;

1: инициализировать вектор-столбцы φ_t, θ_d случайным образом;
2: **повторять**
3: обнулить n_{wt}, n_{td}, n_t, n_d для всех $d \in D$, $w \in W$, $t \in T$;
4: **для всех** $d \in D$, $w \in d$
5: $Z := \sum_{t \in T} \varphi_{wt}\theta_{td}$;
6: **для всех** $t \in T : \varphi_{wt}\theta_{td} > 0$
7: увеличить n_{wt}, n_{td}, n_t, n_d на $\delta = \frac{n_{dw}\varphi_{wt}\theta_{td}}{Z}$;
8: $\varphi_{wt} := \text{norm}_w(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}})$ for all $w \in W, t \in T$;
9: $\theta_{td} := \text{norm}_t(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}})$ for all $d \in D, t \in T$;
10: **пока** Θ и Φ не сойдутся

2.3 Мультимодальные тематические модели

Мультимодальная тематическая модель позволяет учитывать наравне со словами другие метки, относящиеся к документу (например, авторы документов или года написания). Каждый класс таких меток (в том числе и слова) называется модальностью тематической модели. Мультимодальное обобщение для алгоритма ARTM вводится в работе [4].

В работе дополнительные модальности использовались для работы с мультиграммами и в качестве вспомогательного инструмента при создании регуляризаторов для документов. Мультиграммами называются группы слов, часто встречающихся вместе. Мультиграммы в исследовании рассматриваются как дополнительная модальность, отличная от модальности слов.

Использование мультиграмм особенно важно при интерактивном тематическом моделировании, поскольку оно существенно повышает интерпретируемость тем, как показано в работе [5].

2.4 Примеры регуляризаторов

Рассмотрим основные регуляризаторы для ARTM (введены в работе [2]).

2.4.1 Сглаживающий регуляризатор

Потребуем, чтобы распределения φ_t и θ_d были близки по дивергенции Кульбака–Лейблера к заданным распределениям β и α соответственно. Используя в качестве $R(\Phi, \Theta)$ дивергенцию Кульбака-Лейблера со знаком минус, получим регуляризатор:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T} \sum_{w \in W} \beta_w \ln \varphi_{wt} + \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_w \ln \theta_{td} \to \max$$

Применение формул для M-шага даст следующий результат:

$$\varphi_{wt} = \mathrm{norm}_w(n_{wt} + \beta_0 \beta_w)$$

$$\theta_{td} = \mathrm{norm}_t(n_{td} + \alpha_0 \alpha_t)$$

Для получения регуляризатора, сглаживающего вероятности тем для документа или терминов для темы, в качестве β и α берутся равномерные распределения.

Такой регуляризатор используется для «фоновых» тем, которые с большей вероятностью содержат все термины коллекции.

2.4.2 Разреживающий регуляризатор

Предположим, что каждый документ и каждый термин связан с небольшим числом тем. Тогда среди вероятностей φ_{wt} и θ_{td} должно быть много нулевых. При построении тематических моделей больших коллекций с большим числом тем сильная разреженность матриц Φ, Θ помогает сократить затраты памяти и времени.

Чем сильнее разрежено распределение, тем меньше его энтропия. Максимальной энтропией обладает равномерное распределение. Поэтому будем максимизировать KL-дивергенцию между модельными распределениями φ_t и θ_d и равномерными распределениями β и α:

Максимизируя KL-дивергенцию между распределениями φ_t, θ_d и равномерными распределениями β, α, получим регуляризатор, разреживающий матрицы Φ и Θ:

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in T} \sum_{w \in W} \beta_w \ln \varphi_{wt} - \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_w \ln \theta_{td} \to \max$$

Формулы для М-шага идентичны формулам из предыдущего подраздела, за исключением знака при $\beta_0\beta_w$ и $\alpha_0\alpha_t$:

$$\varphi_{wt} = \text{norm}_w(n_{wt} - \beta_0\beta_w)$$

$$\theta_{td} = \text{norm}_t(n_{td} - \alpha_0\alpha_t)$$

2.4.3 Сглаживающий регуляризатор для частичного обучения

В данном исследовании поднимается вопрос частичного обучения тематических моделей с помощью экспертов. Предлагается делать это следующим образом: для каждой темы эксперты задают так называемые белые и черные списки терминов и документов (соответственно $W_t^+, W_t^-, D_t^+, D_t^-$). Белые списки содержат те термины и документы, которые, по мнению эксперта, относятся к теме t, а черные — те термины и документы, которые эксперт считает необходимым из темы t исключить.

ARTM предоставляет следующий механизм для реализации задаваемых экспертом связей: используется регуляризатор, минимизирующий сумму KL-дивергенций между φ_{wt} и равномерными распределениями на подмножествах терминов $\beta_{wt} = \frac{1}{|W_t^+|}[w \in W_t^+]$, а также между θ_{td} и равномерными распределениями на подмножествах тем $\alpha_{td} = \frac{1}{|t:\ d \in D_t^+|}[d \in D_t^+]$.

После подстановки этих распределений в формулу сглаживающего регуляризатора, формулы М-шага принимают вид:

$$\varphi_{wt} = \text{norm}_w(n_{wt} + \beta_0\beta_{wt}[w \in W_t^+])$$

$$\theta_{td} = \text{norm}_t(n_{td} + \alpha_0\alpha_{td}[d \in D_t^+])$$

Соответственно, для черных списков следует взять отрицательные коэффициенты β_0 и α_0.

2.4.4 Декоррелирующий регуляризатор для тем

Для повышения различности тем используется регуляризатор, минимизирующий ковариации между вектор-столбцами матрицы Φ:

Считается, что повышение различности тем улучшает интерпретируемость модели [6]. Регуляризатор, минимизирующий ковариации между вектор-столбцами φ_t, φ_s,

$$R(\Phi) = -\gamma \sum_{t \in T} \sum_{s \in T t} \sum_{w \in W} \varphi_{wt}\varphi_{ws} \to \max$$

приводит к формуле M-шага

$$\varphi_{wt} = \text{norm}_w \left(n_{wt} - \gamma \varphi_{wt} \sum_{s \in T} \varphi_{ws} \right)$$

2.5 Библиотека BigARTM

Для проведения экспериментов и создания инструмента для обучения с частичным привлечением учителя была использована библиотека BigARTM [7], в которой алгоритм ARTM реализован в онлайновом виде, поддерживающем параллельную обработку данных.

3 Инструмент для интерактивного тематического моделирования

3.1 Требования к инструменту

Большинство работ в смежной области рассматривают пользовательскую обратную связь как метод оценить качество модели, а не как способ улучшить ее качество. К примеру, в работе [8] рассматриваются методики Word Intrusion и Topic Intrusion, которые позволяют оценить когерентность тематической модели путем опроса групп людей с использованием специально отобранных групп слов. Этот метод можно реорганизовать и для обучения тематических моделей, но он рассчитан на большие группы произвольных людей, а не на одного эксперта.

Оптимальным способом взаимодействия эксперта с инструментом в данном исследовании был выбран следующий:

1. Предположим, что тематическая модель для коллекции уже построена. Теперь эксперт может видеть выделенные в коллекции темы, а также какие термины и документы к ним относятся.

2. Теперь эксперт может оценить построенные темы, для каждой темы указав, какие термины действительно являются в этой теме ключевыми, а какие, наоборот, следует из темы исключить. При необходимости можно добавить в модель новые темы, либо удалить старые.

3. После того, как экспертная разметка была произведена, обучение тематической модели запускается заново, при этом используется информация, полученная от эксперта (каким именно образом она используется, описано в разделе 5).

Получаем, что основным требованием к визуализации тематической модели является возможность для эксперта быстро оценить, для какой темы какие термины и документы являются основными, и при необходимости подтвердить или опровергнуть наличие взаимосвязи между термином и темой или документом и темой.

3.2 Обзор средств визуализации тематических моделей

Вопрос визуализации тематических моделей активно изучается и описан в работах [9, 10]. Подробному обзору методов визуализации посвящена работа [11].

В данном исследовании для визуализации был выбран подход, аналогичный использованному в работе [9], как наиболее удовлетворяющий требованиям предыдущего подраздела. Подход заключается в том, что каждой теме, каждому документу и каждому термину программе-визуализаторе сопоставляется свое «страница». Для темы она содержит список терминов и документов, отранжированных по вероятностям $\mathsf{p}(w|t)$ и $\mathsf{p}(t|d)$. Аналогично, для термина страница содержит список тем, отранжированных по вероятностям $\mathsf{p}(w|t)$ и для документов — список тем, отранжированных по вероятностям $\mathsf{p}(t|d)$. При этом термины, темы и документы на таких «страницах» являются гиперссылками, что позволяет пользователю осуществлять тематическую навигацию по коллекции.

3.3 Пользовательский интерфейс инструмента

Инструмент для интерактивного тематического моделирования позволяет загрузить произвольную текстовую коллекцию с возможностью разбития на модальности.

После загрузки коллекции становится доступна опция обучения тематической модели (см. рис. 1). Есть возможность выбрать, какие регуляризаторы тем следует оставить в модели, а какие нет. Также есть возможность вручную задать параметры регуляризаторов тематической модели.

После обучения тематической модели предоставляется возможность просмотра тем (см. рис. 2 и 3), терминов и документов (см. рис. 4). Окна тем, документов и терминов связаны друг с другом гиперссылками, а также содержат кнопки, позволяющие эксперту создать набор входных данных для регуляризаторов черных и белых списков (эти регуляризаторы будут доступны при следующем обучении тематической модели). Окна содержат также дополнительные опции для повышения наглядности структуры созданной тематической модели.

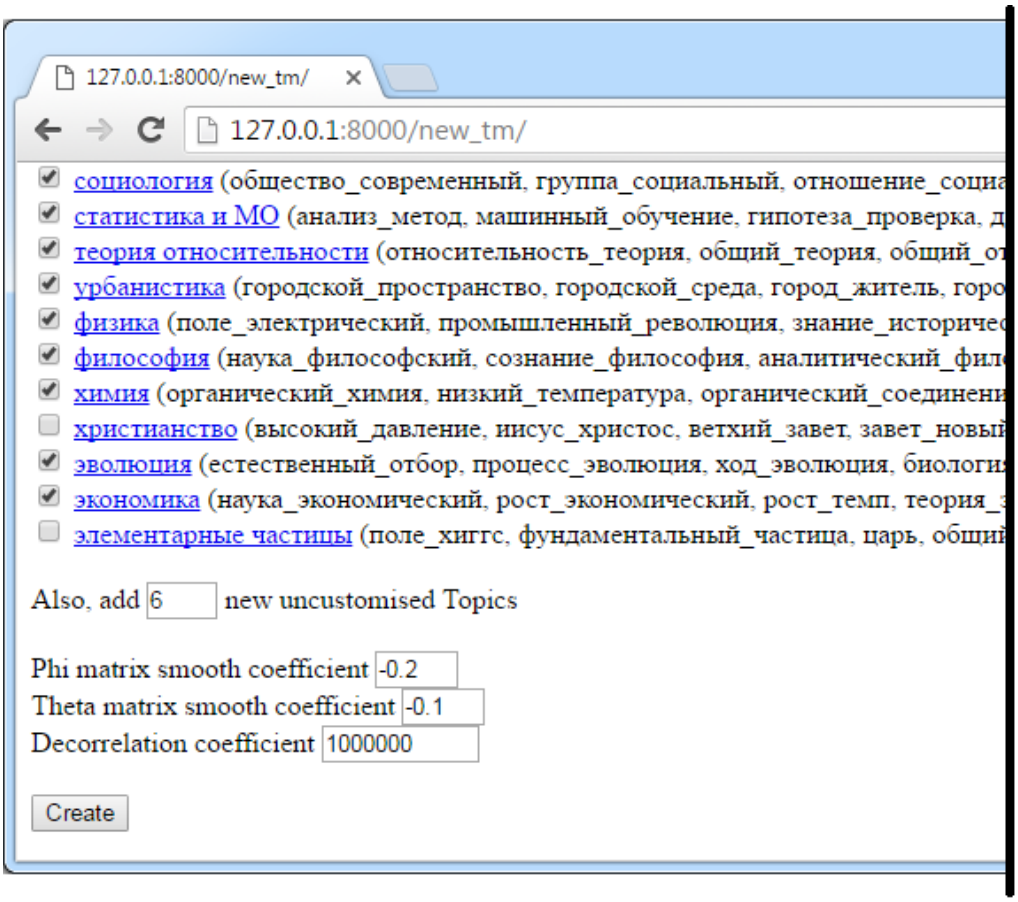

Рис. 1: Обучение новой тематической модели

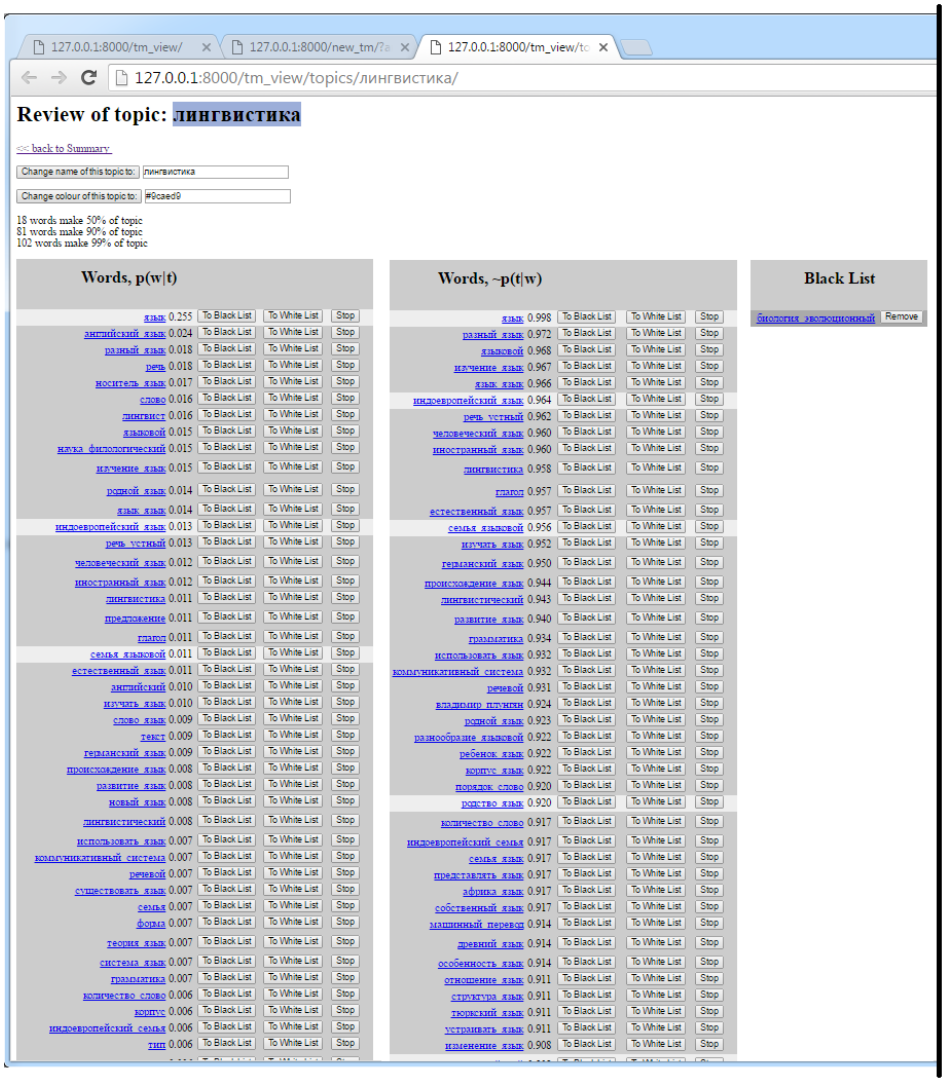

Рис. 2: Окно темы (часть 1)

Окно темы содержит следующие функции:

- Название темы, кнопки, позволяющие переименовать тему и сменить ее цвет;
- Количество терминов, которые составляют $k\%$ темы, где $k = 50, 90, 99$;
- Списки терминов, упорядоченные в соответствии с $\mathsf{p}(w|t)$ и $\mathsf{p}(t|w)$;
- Списки терминов, находящихся в черном и белом списках темы;

- Кнопки, позволяющие заносить термины в черный и белый списки, а также удалять их оттуда;
- Кнопки, позволяющие заносить термины в список стоп-слов;
- Список документов, упорядоченный в соответствии с $\mathsf{p}(t|d)$;
- Кнопки, позволяющие заносить документы в белый список темы, а также удалять их оттуда;
- Список тем, отсортированный в соответствии с косинусной мерой близости (каждая тема рассматривается как вектор ее терминов).

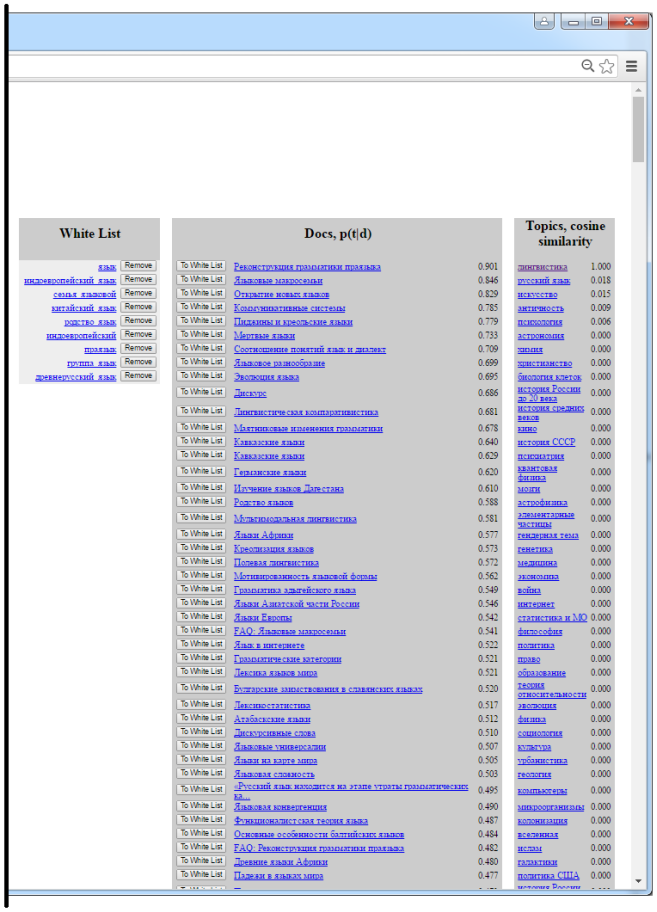

Рис. 3: Окно темы (часть 2)

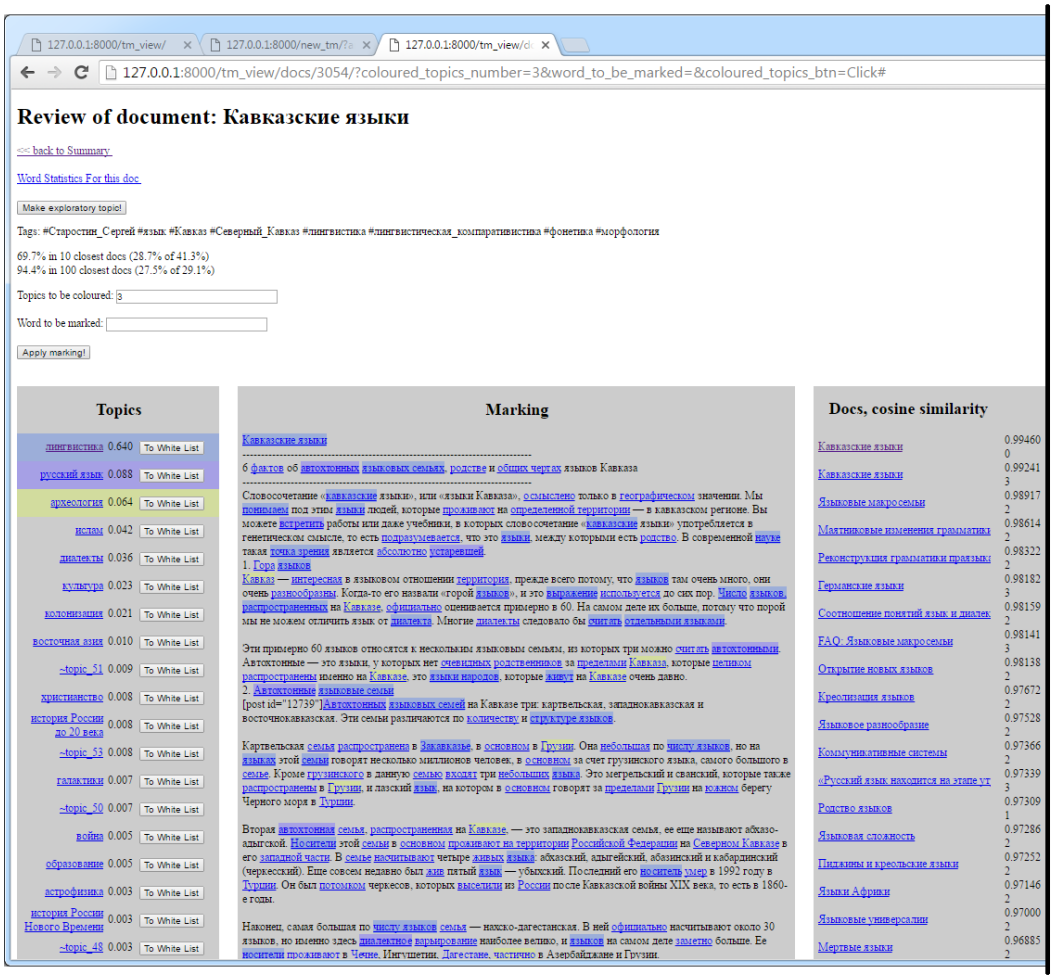

Рис. 4: Окно документа

Окно документа содержит следующие функции:

- Название документа;
- Теги, соответствующие документу, если экспертная разметка была проведена;
- Списки тем, упорядоченные в соответствии с $\mathsf{p}(t|d)$;
- Кнопки, позволяющие заносить документ в белый список той или иной темы, а также удалять его оттуда;
- Разметка документа, которая с помощью цветового кодирования показывает, к какой теме какой термин в документе относится;

- Список документов, отсортированный в соответствии с ранжировкой тематического поиска (см. раздел 4);
- Метрики качества тематического поиска для данного документа (метрики введены в разделе 5).

Окно термина содержит следующие функции:

- Название термина;
- Статистика по количеству вхождений термина во все документы коллекции;
- Класс термина (униграмма или мультиграмма);
- Списки тем, упорядоченные в соответствии с $\mathsf{p}(w|t)$;
- Списки тем, для которых термин находится в черном или белом списке;
- Кнопки, позволяющие заносить термин в белый список той или иной темы, а также удалять его оттуда;
- Кнопка, позволяющая добавить термин в список стоп-слов;
- Список документов, отсортированный в соответствии с количеством вхождения в них данного термина.

Для тематической модели в целом доступна функция оценки качества тематического поиска на ней (см. раздел 4).

Есть возможность заносить термины в список стоп-слов, после чего этот термин не будет учитываться как часть коллекции при обучении тематической модели — это можно сделать как из интерфейса обзора термина, так и централизованно (см. рис. 5).

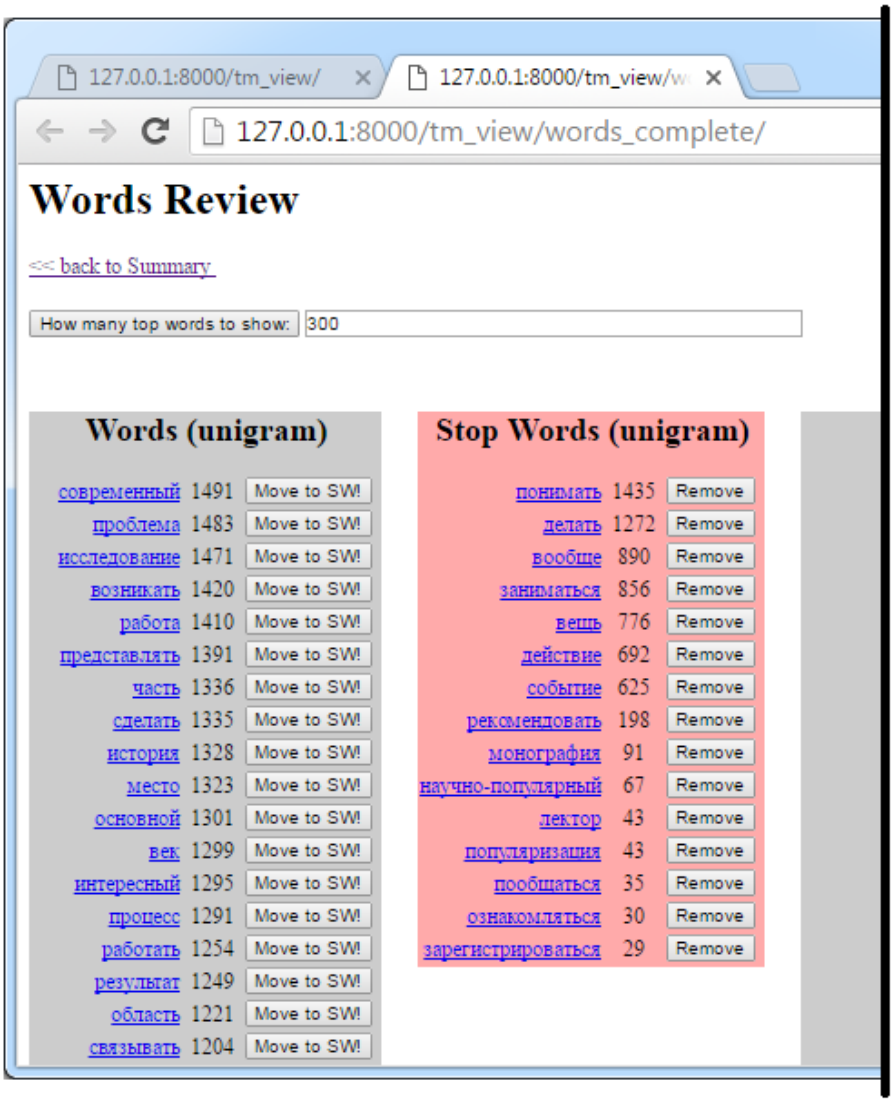

Рис. 5: Разметка стоп-слов

3.4 Технические особенности реализации

Исходя из постановки задачи, наиболее удобен вариант реализации программы, при котором у пользователей инструмента есть доступ к веб-интерфейсу, через который они связываются с приложением на сервере. Поскольку BigARTM имеет API на языке Python, для реализации такого подхода был использован Django — свободный фреймворк для веб-приложений на языке Python.

Версия библиотеки BigARTM, использованная в работе — 0.7.3.

Настройки тематической модели, полученные при работе с экспертом, хранятся на сервере в виде реляционной базы данных SQLite. При желании эти же настройки можно использовать и для работы с другими коллекциями текстов. Также можно модифицировать их, вообще не имея коллекции текстов.

4 Тематический поиск и его качество

4.1 Задача тематического поиска

Задача тематического поиска состоит в том, чтобы по предоставленному пользователем запросу обеспечить выдачу документов, наиболее близких к тексту запроса в тематическом смысле. В качестве запроса обычно выступает документ или несколько документов.

Стандартный метод решения этой задачи заключается в том, что для документа d считается вектор его вероятностей $\mathsf{p}(t \mid d)$, и в качестве выдачи тематического поиска используются документы, ближайшие к вектору d по косинусной мере. Иногда «хвосты» распределения для этого вектора обнуляются.

Тематический поиск является одной из основных функций использования тематических моделей для коллекций текстов. Поисковым намерением пользователя является получение документов, семантически близких к документу-запросу.

4.2 Оценка качества тематического поиска

Наиболее очевидный способ оценить качество тематического поиска — асессорская оценка. Однако этот способ слишком трудоемок для эффективного использования во время исследования. Более того, функционал инструмента для интерактивного тематического моделирования в его конечном виде должен включать оценку качества, что невозможно для данного метода.

В итоге для оценки тематического поиска был использован следующий метод: разрабатывается некоторое множество тегов, и затем каждому документу

экспертами присваиваются те теги, которым этот документ соответствует (примеры тегов: «Первая мировая война», «Вячеслав Молотов», «Буддизм»). Пусть множество тегов документа d_i — $Tag(d_i)$. Пусть $D_N(d_i)$ — множество первых N документов в поисковой выдаче для документа d_i. Для метрики качества тематического поиска для документа d_x будем использовать следующую величину:

$$Q_N(d_x) = \frac{\sum_{d \in D_N(d_x)} \left(\frac{|Tag(d_x) \cap Tag(d)|}{|Tag(d_x)|} \right)}{N}$$

Поскольку коллекция может не содержать достаточное количество документов с нужными тегами, в большинстве случаев более целесообразно использовать относительный эквивалент этой метрики:

$$Q_N^{rel}(d_x) = \frac{Q_N(d_x)}{\max_{D_N}(Q_N(d_x))}$$

В ходе работы в основном использовались метрики $Q_{10}^{rel}(d_x)$ и $Q_{100}^{rel}(d_x)$.

4.3 Использование тематического поиска в качестве метрики качества тематической модели

В ходе рассмотрения различных вариантов метрики качества было принято решение использовать среднее значение Q_N^{rel} для некоторого фиксированного подмножества документов коллекции как метрику качества тематической модели в целом.

Такая метрика качества не имеет прямой зависимости от количества терминов, тем и документов, а также обладает логичной интерпретацией — тематическая модель хороша настолько, насколько правильно она определяет близкие по содержанию документы.

5 Разработка регуляризаторов для обучения с частичным привлечением учителя

5.1 Регуляризаторы черных и белых списков

После того, как эксперты закончили свою работу, в качестве входных данных программа для каждой темы получает черный список терминов, которые не должны относиться к этой теме и белый список терминов, которые к теме относиться должны. Аналогичные списки существуют для документов. Задача заключается в том, чтобы, используя эти знания, построить корректные темы, отвечающие представлениям эксперта.

Задача с черными списками решается просто — достаточно ввести регуляризаторы с коэффициентом $-\infty$ для соответствующих значений φ и θ.

Рассмотрим задачу с белым списком. Использование стандартных регуляризаторов частичного обучения не оправдано, поскольку значения φ получаются слишком большими, и, как правило, φ для остальных терминов темы вырождаются. Это объясняется тем, что такой регуляризатор не делает поправку на частоту терминов в изначальной коллекции, из-за чего полученные темы не имеют статистической интерпретации. Более того, оптимальные значения коэффициентов регуляризации сильно зависят не только от конкретной коллекции, но и от самих множеств белых и черных списков. Для этой задачи был разработан регуляризатор следующего вида:

$$\beta_{wt} = \frac{[w \in W_t^+] \cdot n_w}{\sum_{w' \in W} ([w' \in W_t^+] \cdot n_{w'})}$$
$$\beta_0 = \beta_0^* \cdot \sum_{w' \in W} ([w' \in W_t^+] \cdot n_{w'})$$

(β_0^* — коэффициент, задающий силу регуляризации, n_w — количество раз, которое термин w содержится в коллекции, W_t^+ — множество терминов, находящихся в белом списке темы t).

Этот регуляризатор обладает очень хорошей интерпретируемостью — он увеличивает количество вхождений каждого термина $w \in W_t^+$ в $(1+\beta_0^*)$ раз, причем «добавочная» часть терминов автоматически относится к теме t.

Аналогичный регуляризатор для белого списка документов будет выглядеть следующим образом:

$$\alpha_{td} = \frac{[d \in D_t^+] \cdot n_d}{\sum_{d' \in D} ([d' \in D_t^+] \cdot n_{d'})} \cdot \frac{1}{\sum_{t' \in T} [d \in D_{t'}^+]}$$
$$\alpha_0 = \frac{1}{1 - \alpha_0^*} \cdot \sum_{d' \in D} ([d' \in D_t^+] \cdot n_{d'})$$

(α_0^* — коэффициент, задающий силу регуляризации, n_d — количество терминов в документе d, D_t^+ — множество документов, находящихся в белом списке темы t).

Интерпретация следующая: доля α_0^* от документа d относится к теме t. Если документ d находится в белых списках нескольких тем, доля α_0^* делится между этими темами поровну.

Запишем регуляризаторы «черных списков» для терминов и документов формально:

$$\beta_{wt}^{BL} = \begin{cases} -\infty, & \text{if } w \in W_t^- \\ 0, & \text{if } w \notin W_t^- \end{cases}$$

$$\alpha_{td}^{BL} = \begin{cases} -\infty, & \text{if } d \in D_t^- \\ 0, & \text{if } d \notin D_t^- \end{cases}$$

W_t^- и D_t^- — множества терминов и документов, находящихся в черном списке темы t, соответственно.

Формулы M-шага EM-алгоритма будут иметь вид:

$$\varphi_{wt} = \text{norm}_w(n_{wt} + \beta_0\beta_{wt} + \beta_{wt}^{BL})$$

$$\theta_{td} = \text{norm}_t(n_{td} + \alpha_0\alpha_{td} + \alpha_{td}^{BL})$$

5.2 Выбор текстовой коллекции для экспериментов

Для проведения экспериментов была использована коллекция текстов ПостНауки — научного интернет-журнала статей, авторами которых являются ученые, работающие в той или иной предметной области. Большое преимущество этой коллекции заключается в том, что она имеет научно-популярный характер, а также содержит статьи, лежащие в разных областях науки — таким образом, оценка интерпретируемости тем, выделенных в ней тематической моделью, может быть осуществлена непрофессионалом.

5.3 Выделение мультиграмм и начальных форм слов в текстах коллекции

Для построения тематических моделей для русскоязычных коллекций необходимо решить задачу объединения всех форм одного слова в один класс эквивалентности, чтобы тематическая модель не рассматривала разные формы одного слова как разные термины.

Также для исключения влияния шумовых данных должна быть решена задача о выделении среди мультиграмм текстовой коллекции статистически значимых. Для решения обоих задач в исследовании использовался метод, описанный в работах [12, 13].

5.4 Эксперименты по обучению тематических моделей

Инструмент черных и белых списков удобно использовать даже непрофессионалу благодаря итеративному процессу разметки. Процесс разметки заключается в следующем:

1. Запускаем инструмент для построения тематических моделей с N темами без какой-либо разметки.
2. Смотрим на результат, размечаем те темы, которые оказались удачными, удаляем остальные.
3. Запускаем инструмент с сохранением размеченных тем и добавлением M новых.
4. Смотрим на результат, размечаем новые темы, которые оказались удачными, удаляем новые неудачные темы, пересматриваем разметку старых.
5. Если на этапе (4) были добавлены удачные новые темы, возвращаемся на этап (3). В противном случае, заканчиваем работу.

От параметров N и M результирующая модель зависит слабо, их точный подбор не играет роли.

Эксперименты показали, что оптимальное значение β_0^* для регуляризаторов белых списков составляет 1.0 для первого прохода EM-алгоритма (для оптимизации инициализации) и 0.2 для всех остальных. В отличие от стандартного регуляризатора, нормировочная величина β_0 универсальна и не нуждается в корректировке для разных коллекций. Оптимальное значение α_0^* составляет 0.33.

Сравним тематическую модель, полученную таким образом, с результатами базовых тематических моделей с разным количеством тем. Разметка автора исследования в эксперименте была далека от идеала, которого можно было бы добиться с привлечением экспертов.

Эксперимент	Q_{10}^{rel}	Q_{100}^{rel}
ТМ без белых и черных списков, 20 тем	42.3%	52.0%
ТМ без белых и черных списков, 30 тем	45.4%	55.0%
ТМ без белых и черных списков, 40 тем	48.3%	56.3%
ТМ без белых и черных списков, 50 тем	50.3%	59.4%
ТМ без белых и черных списков, 60 тем	50.7%	57.8%
ТМ без белых и черных списков, 70 тем	50.5%	57.5%
ТМ со стандартными регуляризаторами белых и черных списков ($\beta_0 = 50$), 50 тем	53.4%	60.9%
ТМ с усовершенствованными регуляризаторами белых и черных списков, 50 тем	55.3%	63.9%

Помимо улучшения метрики качества тематического поиска, следует отметить значительное улучшение интерпретируемости. В частности, эксперимент по разметке тем показал, что если при итеративной разметке коллекции не пользоваться регуляризаторами черных списков для терминов, итоговая тематическая модель практически не будет содержать термины, которые потребуют занесения в черные списки.

5.5 Эксперименты по использованию стандартных регуляризаторов

Эксперименты показывают, что использование разреживающего регуляризатора для матрицы φ с коэффициентом -1.0 ведет к улучшению качества, а сглаживающий/разреживающий регуляризатор для Θ и декоррелирующий регуляризатор для φ не помогают. Вероятно, это связано с тем, что большая часть фоновых терминов уже была убрана из коллекции на этапе выделения мультиграмм и начальных форм термина.

6 Улучшение качества тематического поиска

6.1 Методы улучшения качества

Цель данного раздела — разработка метода тематического поиска, позволяющий добиться лучшего качества.

Была принята гипотеза, что для максимально эффективного тематического поиска нужно научиться отличать друг от друга документы, относящиеся к одной теме. Это позволит выбирать среди множества документов, близких к поисковому документу d_s по распределению тем, самые близкие, для отделения которых от остальных одного вектора тем недостаточно. Предполагалось, что эту задачу можно решить за счет модификации регуляризаторов тем, к которым относится d_s таким образом, чтобы тематический поиск по результирующим темам возвращал в первую очередь документы, более близкие к d_s.

6.2 Результаты экспериментов

В соответствии с результатами экспериментов, существенного улучшения качества тематического поиска позволяет добиться следующий метод:

Все регуляризаторы «белых» и «черных» списков для тем остаются неизменными, но в дополнение к ним добавляется новая «разведочная» тема, в белый список которой добавлены «наиболее релевантные» термины документа d_s, а также небольшое количество его наиболее частых терминов (здесь важно отметить, что почти все стоп-слова были исключены из коллекции на этапе статистического анализа для выделения мультиграмм, а остальные, предположительно, были исключены экспертами в процессе разметки).

«Релевантность» в данном случае оценивается с помощью следующей эвристики: «релевантными» признаются те термины, которые имеют суммарное значение φ для трех «верхних» тем документа больше, чем для всех остальных тем.

Документ d_s заносится в белые списки трех наиболее значимых для него тем, а также в белый список «разведочной темы». На этих данных обучается тематическая модель, тематический поиск по которой проводится с использованием не косинусной меры, а функции

$$\text{vect\_min}(d_s, d') = \sum_{t \in T} \min(\theta_{td_s}, \theta_{td'})$$

Из результатов этого тематического поиска исключаются документы, которые были низко (ниже позиции 50) в результатах тематического поиска для немодифицированной модели. Эта эвристика позволяет добиться значительного улучшения метрики качества и имеет простое объяснение: дополнительные

меры по улучшению тематического поиска принимаются прежде всего затем, чтобы выделить лучшие документы из списка отнесенных базовой тематической моделью к одной области (соответственно, документы, изначально к ней не отнесенные, рассматриваться не должны).

Использование описанного алгоритма позволяет существенно улучшить качество тематического поиска. Рассмотрим изменение метрики качества для тематического поиска по зафиксированному случайно выбранному подмножеству документов:

Эксперимент	Q_{10}^{rel}
Стандартный алгоритм	55.3%
Алгоритм с «разведочной темой»	58.9%

Более детальный анализ улучшения метрики качества можно сделать на основании графика, на котором отмечена разница качества между базовым тематическим поиском и новой схемой (см. рис. 6).

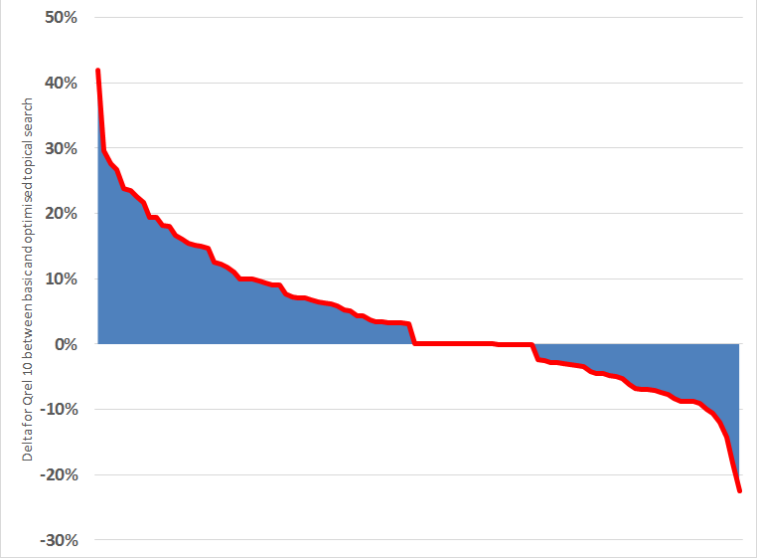

Рис. 6: Улучшение качества тематического поиска

Рассмотрим функцию vect_min(d_s, d'):

Эксперимент	Q_{10}^{rel}
Алгоритм с «разведочной темой» и vect_min	58.9%
Алгоритм с «разведочной темой» и косинусной мерой	56.2%
Алгоритм с «разведочной темой» и косинусной мерой для 3х верхних тем	55.4%
Стандартный алгоритм и vect_min	52.8%
Стандартный алгоритм и косинусная мера	55.3%

Как видно из таблицы, функция vect_min в сочетании с новым методом тематического поиска дает лучшие результаты, чем другие варианты сравнения векторов. В то же время, она работает хуже при стандартном тематическом поиске. Этот факт имеет следующее объяснение: «разведочная тема» в силу алгоритма построения создается не идеально, и многие «пограничные» термины между ней и другими наиболее вероятными темами документа классифицируются неправильно. Из-за этого вероятности $\mathsf{p}(t \mid d)$ оказываются смещены в сторону «разведочной темы», и косинусная мера дает большую погрешность как менее устойчивая к изменениям сравниваемых векторов.

Как следствие, экспертный надзор за «разведочной темой» поможет еще больше улучшить качество тематического поиска.

6.3 Интерпретация результата

Таким образом, логика работы нового метода тематического поиска такова: он создает новую тему в модели, если ее наличие поможет классифицировать документ.

Следовательно, этот метод можно использовать не только для тематического поиска, но и для помощи экспертам в улучшении тематической модели, поскольку разработанный алгоритм помогает создавать новые темы, ориентируясь при этом на конкретный документ. При этом наилучшего качества можно добиться, если использовать «разведочную тему» вместе с экспертным контролем.

6.4 Проверка нового метода разведочного поиска для обучения без учителя

Полученный метод тематического поиска можно использовать и без разметки коллекции экспертами. Поскольку этот метод требует наличия для всех релевантных тем регуляризаторов «белых списков», для этого нужно сэмулировать работу экспертов, записав в «белые списки» каждой темы 5 наиболее вероятных терминов для данной темы. Рассмотрим результаты тематического поиска для такой модели:

Эксперимент	Q_{10}^{rel}
Стандартный алгоритм	50.3%
Стандартный алгоритм с искусственными «белыми списками»	51.8%
Стандартный алгоритм с искусственными «белыми списками» и новым методом тематического поиска	55.1%

Можно отметить, что добавление подобных «очевидных» регуляризаторов само по себе повысило качество модели. Вероятно, это произошло, потому что такой регуляризатор помогает также в начальной инициализации тем, которая важна для создания качественной тематической модели (это показано в работе [14]).

Итак, мы видим, что новый метод тематического поиска демонстрирует значительное улучшение даже для модели, над которой не работали эксперты. Это значительно расширяет его спектр применимости.

7 Заключение

7.1 Результаты, выносимые на защиту

1. Создан инструмент навигации, визуализации, поиска, оценивания и частичного обучения с веб-интерфейсом, позволяющий повысить качество и интерпретируемость тематической модели.

2. Разработаны регуляризаторы частичного обучения на основе черных и белых списков терминов и документов для тем. Показано, что небольшая доля разметки позволяет улучшить качество тематической модели.

3. Разработан новый метод разведочного тематического поиска документов в текстовой коллекции, основанный на введении «разведочной темы».

4. Предложена метрика качества тематического поиска по тегированным коллекциям.

7.2 Выводы и перспективы для продолжения исследования

Задача и подзадачи, поставленные в начале исследования, были выполнены.

Алгоритм ARTM отлично показал себя при решении задачи частичного обучения, что показывает мощность инструмента регуляризации.

Перечислим направления, в которых исследование можно продолжить:

1. В ходе работы тематический поиск рассматривался как задача нахождения документов, ближайших друг к другу по содержанию. Однако имеет смысл и другая трактовка: пользователю следует выдавать документы, которые ему может быть интересно прочитать после данного. Это могут быть, например, документы, в которых раскрываются конкретные термины из документа-запроса. Построение оптимальных «пользовательских рекомендаций» представляет практический интерес, однако требует трудоемкой асессорской оценки при проведении экспериментов.

2. Рассмотрение документов в виде «мешков слов» ограничивает возможности программы. Результаты тематического поиска могут оказаться интереснее, если запускать его по отдельности для разных абзацев текста (или для тематически однородных участков текста).

3. При обучении тематической модели экспертами, а также при изучении коллекции пользователями, может помочь более совершенная графическая визуализация. Поскольку разработок по визуализации тематических моделей довольно много, представляется интересным изучение вопроса, может ли такая визуализация помочь эксперту при построении.

Список литературы

[1] Vorontsov K. V. Additive Regularization for Topic Models of Text Collections // Doklady Mathematics. 2014, Pleiades Publishing, Ltd. — Vol. 89, No. 3, pp. 301–304.

[2] Vorontsov K. V., Potapenko A. A. Additive Regularization of Topic Models // Machine Learning Journal, Special Issue "Data Analysis and Intelligent Optimization", Springer, 2014.

[3] Hofmann, T. Unsupervised Learning by Probabilistic Latent Semantic Analysis // Machine Learning, 42, 177–196, 2001 Kluwer Academic Publishers.

[4] K. Vorontsov, O. Frei, M. Apishev, P. Romov, M. Dudarenko. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. The 4th International Conference on Analysis of Images, Social Networks, and Texts, 2015.

[5] Hanna M. Wallach, Topic Modeling: Beyond Bag-of-words, Proceedings of the 23rd International Conference on Machine Learning, ICML '06, 2006, Pittsburgh, Pennsylvania, USA, 977–984, ACM, New York, NY, USA

[6] Tan Y., Ou Z. Topic-weak-correlated latent dirichlet allocation // 7th International Symposium Chinese Spoken Language Processing (ISCSLP). — 2010. — Pp. 224—228.

[7] Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, Marina Suvorova, Anastasia Yanina, Non-Bayesian Additive Regularization for Multimodal Topic Modeling of Large Collections, Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications, 2015, Melbourne, Australia, 29–37, ACM, New York, NY, USA

[8] Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, David M. Blei. Reading Tea Leaves: How Humans Interpret Topic Models, NIPS 2009.

[9] Allison J. B. Chaney and David M. Blei. Visualizing Topic Models, Association for the Advancement of Artificial Intelligence (www.aaai.org), 2012.

[10] Brynjar Gretarsson, John O'Donovan, Svetlin Bostandjiev, Tobias Hollerer, Arthur Asuncion, DAVID Newman, Padhraic Smyth.

TopicNets: Visual Analysis of Large Text Corpora with Topic Modeling // ACM Transactions on Intelligent Systems and Technology (TIST) archive, Volume 3 Issue 2, February 2012, Article No. 23

[11] Р. М. Айсина, Обзор средств визуализации тематических моделей коллекций текстовых документов, Машинное обучение и анализ данных (http://jmlda.org), 1, 11, 2015, 1584-1618

[12] Царьков С.В., Стенин С.С. Статистические методы сокращения словаря n-грамм для построения вероятностных тематических моделей // Сб. тезисов конференции МФТИ-57. 2014. С.14-15.

[13] Царьков С.В. Автоматическое выделение ключевых фраз для построения словаря терминов в тематических моделях коллекций текстовых документов // Естественные и технические науки. М.: Спутник+, 2012, №6. С. 456–464.

[14] Evgeny Sokolov, Lev Bogolubsky, Topic Models Regularization and Initialization for Regression Problems, Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications, 2015, Melbourne, Australia, 21–27, ACM, New York, NY, USA