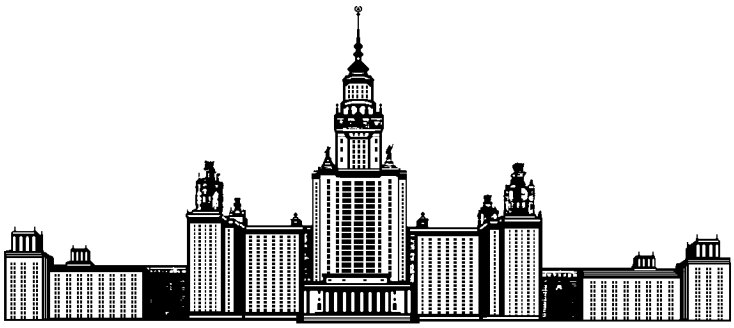
Московский государственный университет имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математических методов прогнозирования



Магистерская программа «Логические и комбинаторные методы анализа данных»

Магистерская диссертация

«Параллельная реализация аддитивно регуляризованного тематического моделирования и её применение для поиска этно-релевантного контента в социальных медиа»

Работу выполнил
Апишев Мурат Азаматович

Научный руководитель:
д.ф-м.н., доцент
Воронцов Константин Вячеславович

Москва, 2017

Содержание

Аннотация

В современных исследованиях Интернета часто используются различные методы анализа текстов для обучения без учителя с целью извлечения информации, релевантной различным тематикам. Разработанный ранее подход аддитивной регуляризации тематических моделей (APTM) предоставляет возможность более гибкого вывода и контроля над темами, чем различные расширения LDA. В данной работе подход APTM был применён в задаче извлечения этносоциального контента из текстов русскоязычного медиапространства. В рамках работы были представлены более совершенный онлайновый параллельный EM-алгоритм для обучения модели, новые регуляризаторы и проведено сравнение моделей APTM и LDA. С помощью экспертных оценок показано, что подход APTM лучше подходит для поиска релевантных и интерпретируемых тем.

1 Введение

Тематические модели стали одним из стандартных инструментов для извлечения данных из больших текстовых коллекций. По сути, тематические модели производят разложение разреженной матрицы «слова-документы» в произведение матриц «слова-темы» и «темы-документы». Впервые эта идея появилась в модели вероятностного латентного семантического анализа (PLSA) [9], сейчас же основным инструментом стала модель латентного размещения Дирихле (LDA), которая является байесовской версией PLSA с априорными распределениями Дирихле для распределения слов в темах и тем в документах [3, 6].

В течение долгого времени LDA находится в центре внимания, было опубликовано множество работ, предлагающий различные её модификации для решения конкретных задач, однако такие модификации являются отдельными инструментами, которые нельзя легко комбинировать друг с другом. Разработка каждого подобного расширения — это большая работа для исследователя в области анализа данных. Специалист из другой предметной области, например, социологии, не будет заниматься разработкой новой модели LDA для каждой конкретной возникающей задачи. Более того, даже небольшая модификация уже существующей реализации модели может оказаться слишком сложной с технической точки зрения задачей.

В данной работе для решения задач из различных предметных предлагается использовать подход аддитивной регуляризации тематических моделей (APTM) [25] и реализующий его программный продукт с открытым кодом BigARTM [24]. APTM обобщает базовую модель PLSA при помощи механизма регуляризации, который можно использовать для оказания прямого влияния на те или иные аспекты модели, которые важно учесть при её построении. По факту, модель LDA [23] является частным случаем APTM с регуляризатором сглаживания.

Гибкость является огромным преимуществом APTM на практике. Обучив базовую модель LDA или APTM без регуляризаторов, исследователь может понять, чего ему не хватает, и сформулировать свои предпочтения в терминах регуляризаторов. В большинстве случаев BigARTM позволяет исследователям комбинировать регуляризаторы из встроенной библиотеки регуляризаторов для получения необходимого качества модели по заданным метрикам.

Помимо гибкой настройки моделей библиотека BigARTM позволяет производить их обучение быстрым онлайновым параллельным EM-алгоритмом. Версия алгоритма, реализованная на момент написания этой работы, уже превосходила существующие аналоги в скорости работы [24]. Тем не менее, она имеет ряд недостатков:

недетерминированность, неочевидность подбора параметров обучения для достижения высокой производительности.

Эта работа состоит из двух связанных логических блоков. Первый заключается в разработке и внедрении нового онлайнового параллельного EM-алгоритма для модели APTM в библиотеку BigARTM с целью его использования в дальнейшем моделировании. Во втором блоке показано применения APTM в решении проблемы извлечения тем, связанных с этно-социальным дискурсом из большой текстовой коллекции (постов блогов). Входными данными, помимо самой текстовой коллекции, является словарь предметных терминов (этнонимов). Для получения хорошей тематической модели множество всех тем делится на два подмножества: предметных (или этнических) и фоновых. Разработан новый регуляризатор частичного обучения, работающий со словарём этнонимов и информацией о разбиении тем, а также регуляризатор учёта этнонимов в виде отдельной модальности. Построена комбинация регуляризаторов для получения более интерпретируемых, разреженных и разнообразных тем. Для текстов, содержащих информацию дате и регионе публикации, построены модели, учитывающие эти метаданные. APTM позволяет делать подобные вещи довольно легко, без сложного вывода и разработки новых алгоритмов.

Для демонстрации результатов применения описанного подхода использовались экспертные оценки качества тем. Показано, что слабо регуляризованный APTM и LDA дают примерно одинаковые по качеству модели, в то время как модель APTM с правильно подобранным набором регуляризаторов даёт более хороший результат [13].

Работа имеет следующую структуру. В разделе 2 представлены базовая модель PLSA, её байесовская модификация LDA и общие понятия подхода APTM. Раздел 3 посвящён описанию регуляризаторов, использованных в этой работе и комментированию эффектов их воздействия на итоговую модель. В разделе 4 рассказывается о библиотеке тематического моделирования BigARTM, реализованных в ней вариантах EM-алгоритмов и новом алгоритме `DetAsync`. Раздел 5 описывает проведённые эксперименты: в первой его части показано тестирование нового алгоритма, вторая же часть посвящена списку различных моделей, которые были обучены, а также результатам оценивания качества полученных ими тем. В разделе 6 описаны основные результаты, полученные при выполнении данной работы и выносимые на защиту.

2 Тематическое моделирование

Пусть D обозначает конечное множество (коллекцию) документов (текстов) и пусть W — конечное множество (словарь) всех *терминов*, из которых состоят эти документы. Под термином подразумевается либо слово, либо целая фраза. В соответствии с гипотезой «мешка слов» каждый документ d from D представляется в виде подмножества словаря W, где каждому слову w ставится в соответствие число n_{dw} раз, которое он встретился в документе d. Предположим, что появления каждого термина в каждом документе связано с некоторой латентной темой из конечного множества тем T. Текстовая коллекция представляется в виде набора троек (d_i, w_i, t_i), $i = 1, \ldots, n$, выбранных независимо из дискретного распределения $p(d, w, t)$ над конечным вероятностным пространством $D \times W \times T$. Термины w_i и документы d_i — это наблюдаемые переменные, а темы t_i — скрытые.

Вероятностная тематическая модель описывает вероятности $p(w \mid d)$ появления терминов в документах как смеси распределений слов в темах $\phi_{wt} = p(w \mid t)$ и тем в документах $\theta_{td} = p(t \mid d)$:

$$p(w \mid d) = \sum_{t \in T} p(w \mid t)\, p(t \mid d) = \sum_{t \in T} \phi_{wt} \theta_{td}. \tag{1}$$

Эта смесь напрямую соответствует генеративному процессу, в процессе которого модель порождает документы d: для каждой позиции слова i происходит генерация индекса темы t_i из распределения $p(t \mid d)$, после чего сэмплируется слово w_i из распределения $p(w \mid t_i)$.

Параметры вероятностной тематической модели часто представляются в виде матриц $\Phi = \big(\phi_{wt}\big)_{W \times T}$ и $\Theta = \big(\theta_{td}\big)_{T \times D}$ с неотрицательными и нормированными столбцами ϕ_t и θ_d, представляющими собой мультиномиальные распределения слов в темах и тем в документах.

2.1 PLSA

В *вероятностном латентном семантическом анализе* (PLSA) [9], тематическая модель (1) обучается путём максимизации логарифма правдоподобия с линейными

ограничениями неотрицательности и нормировки:

$$L(\Phi,\Theta)=\sum_{d\in D}\sum_{w\in d}n_{dw}\ln\sum_{t\in T}\phi_{wt}\theta_{td}\ \to\ \max_{\Phi,\Theta}$$
$$\sum_{w\in W}\phi_{wt}=1,\ \phi_{wt}\geq 0,\quad \sum_{t\in T}\theta_{td}=1,\ \theta_{td}\geq 0.$$

где n_{dw}, как было отмечено ранее, — абсолютная частота слова w в документе d.

Решение этой оптимизационной задачи удовлетворяет условиям Каруша-Куна-Такера со вспомогательными переменными p_{tdw}, n_{wt}, n_{td}:

$$p_{tdw}=\underset{t\in T}{\mathrm{norm}}\big(\phi_{wt}\theta_{td}\big) \tag{2}$$
$$\phi_{wt}=\underset{w\in W}{\mathrm{norm}}(n_{wt}),\quad n_{wt}=\sum_{d\in D}n_{dw}p_{tdw}, \tag{3}$$
$$\theta_{td}=\underset{t\in T}{\mathrm{norm}}(n_{td}),\quad n_{td}=\sum_{w\in d}n_{dw}p_{tdw}, \tag{4}$$

где оператор «norm» преобразует вещественный вектор $(x_t)_{t\in T}$ в вектор $(\tilde{x}_t)_{t\in T}$, представляющий собой дискретное распределение:

$$\tilde{x}_t=\underset{t\in T}{\mathrm{norm}}(x_t)=\frac{\max\{x_t,0\}}{\sum_{s\in T}\max\{x_s,0\}}.$$

Метод простых итерация для решения этой системы уравнений эквивалентен EM-алгоритму и обычно на практике используется именно он.

E-шаг (2) может рассматриваться как применение формулы Байеса для получения вероятностей $p_{tdw}=p(t\,|\,d,w)$ для каждого термина w и документа d. M-шаг (3)-(4) интерпретируется как частотная оценка условных вероятностей ϕ_{wt} и θ_{td}. Итеративный процесс обычно начинается со случайных начальных приближений Φ и Θ.

2.2 LDA

Модель латентного размещения Дирихле (LDA) [3, 6] вводит априорные распределения Дирихле для векторов вероятностей слов в темах $\phi_t\sim\mathrm{Dir}(\beta)$ и для векторов вероятностей тем в документах $\theta_d\sim\mathrm{Dir}(\alpha)$ с векторами параметров $\beta=(\beta_w)_{w\in W}$ и $\alpha=(\alpha_t)_{t\in T}$ соответственно.

Вывод в LDA обычно производится либо с помощью вариационного приближения, либо с помощью сэмплирования Гиббса. Обычно используется *свёрнутая схема*

Гиббса, где тема t_i для каждой позиции слова (d_i, w_i) итеративно сэмплируется из распределения $p(t \mid d, w)$, такого же, как в PLSA, но со со сглаженными байесовскими оценками условных переменных:

$$\phi_{wt} = \underset{w \in W}{\text{norm}}(n_{wt} + \beta_w), \quad \theta_{td} = \underset{t \in T}{\text{norm}}(n_{td} + \alpha_t),$$

где n_{wt} — число раз, которое термин w был сгенерирован из темы t и n_{td} это число раз, которое термины из документы d были сгенерированы из темы t, исключая текущую тройку (d_i, w_i, t_i).

В последние годы было опубликовано много работ с различными расширениями LDA. Для описываемой здесь задачи извлечения пользовательской информации по некоторой специфичной тематике (здесь — связанной с этничностями) наиболее релевантными являются модель *Topic-in-Set knowledge* и её расширение [2, 1], где слова, связаны с «z-метками» (z-метка описывает тему, к которой должно быть отнесено слово), а также модель Interval Semi-Supervised LDA (ISLDA) [4, 15], где выделенным темам присваиваются заданные слова, и сэмплирование распределений проецируется на это множество.

2.3 ARTM

Тематическое моделирование может быть рассмотрено как специальный случай матричного разложения, где задача состоит в том, чтобы найти низкоранговую аппроксимацию $\Phi\Theta$ данной разреженной матрицы счётчиков терминов-документов. Следует отметить, что произведение $\Phi\Theta$ определено с точностью до невырожденного линейного преобразования: $\Phi\Theta = (\Phi S)(S^{-1}\Theta)$. Таким образом, задача является некорректно поставленной и имеет бесконечное множество решений. Прошлые эксперименты на модельных [23] и реальных [4] показали, что ни PLSA, ни LDA не удаётся достигнуть устойчивого решения. Для увеличения стабильности обучения следует добавить дополнительные оптимизационные ограничения, обычно называемые *регуляризаторами* [21].

В *аддитивной регуляризации тематических моделей* (APTM) [25] модель обучается путём максимизации линейной комбинации логарифма правдоподобия $L(\Phi, \Theta)$ и r регуляризаторов $R_i(\Phi, \Theta)$, $i = 1, \ldots, r$ с *коэффициентами регуляризации* τ_i:

$$R(\Phi, \Theta) = \sum_{i=1}^{r} \tau_i R_i(\Phi, \Theta), \quad L(\Phi, \Theta) + R(\Phi, \Theta) \;\to\; \max_{\Phi, \Theta}.$$

Условия Каруша-Куна-Такера для этой нелинейной оптимизационной задачи дают (с учётом некоторых технических ограничений) необходимые условия локального максимума как решения следующей системы уравнений [23]:

$$p_{tdw} = \underset{t \in T}{\text{norm}}\big(\phi_{wt}\theta_{td}\big); \tag{5}$$

$$\phi_{wt} = \underset{w \in W}{\text{norm}}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right); \quad n_{wt} = \sum_{d \in D} n_{dw} p_{tdw}; \tag{6}$$

$$\theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \quad n_{td} = \sum_{w \in d} n_{dw} p_{tdw}. \tag{7}$$

Как и в случае PLSA, для решения этой системы может быть использован EM-алгоритм. Преимущество АРТМ заключается в том, что каждый аддитивный регуляризатор превращается в простую модификацию М-шага. Многие модели, разработанные прежде в рамках байесовского подхода, могут быть несложно интерпретированы, обучены и скомбинированы в рамках теории АРТМ [22, 23]. Например, PLSA не использует никакой регуляризации, $R = 0$, а LDA с априорными распределениями Дирихле $\phi_t \sim \text{Dir}(\beta)$ и $\theta_d \sim \text{Dir}(\alpha)$ and оценками максимума апостериорных вероятностей Φ, Θ соответствует модели со сглаживающим регуляризатором, который интепретируется как минимизатор KL-дивергенций между столбцами Φ, Θ и заданными распределениями β, α соответственно.

3 Аддитивная регуляризация

3.1 Общий подход

В этом разделе рассматривается задача разведочного поиска всех этнических в большой коллекции постов блогов. Пусть дан набор этнонимов $Q \subset W$, который может быть слишком большим для традиционных поисковых систем. Для извлечения этнических тем используется тематическая модель частичного обучения с заданной априорной информацией. Похожая техника использовалась ранее в задаче кластеризации новостей [11], поиска тем, связанных со здоровьем, в социальных медиа [16] и задаче поиска этно-релевантных тем в постах блогов [4, 15]. Во всех этих исследованиях каждой теме задавался предопределённый набор ключевых слов, часто очень маленький, т.е. категория новости или этничность. Это означает, что информация о числе тем и их примерном содержимом известна заранее.

Модель *interval semi-supervised LDA* (ISLDA) позволяет присвоить каждой этнич-

ности больше, чем одну тему, однако, довольно сложно определить, сколько реально тем соответствует каждой из этничностей. И если исследователь не задаст ключевые слова для каждой из тем, обучение модели невозможно.

Например, в [4, 15], где цель исследования была схожа с описываемой в данной работе, ISLDA использовался для поиска этнического контента, но, поскольку этнонимы были соотнесены с различными темами, появление мульти-этничных тем было невозможно.

Предлагаемое решение описанной проблемы состоит в задании общей для всех этнических тем лексической априорной информации Q. Модели предоставляется самостоятельно определять распределение этничностей и их комбинаций по темам.

Используется аддитивная комбинация регуляризаторов сглаживания, разреживания и декорреляции тем для того, чтобы сделать темы более интерпретируемыми, разреженными и разнообразными [23]. АРТМ позволяет делать всё это легко, без разработки новых алгоритмов и сложного вывода. Более того, все эти регуляризаторы уже реализованы в библиотеке с открытым кодом BigARTM[1].

Прежде всего необходимо произвести разбиение всех тем T на два подмножества: *предметные* темы S и *фоновые* темы S. Регуляризаторы воздействуют на B и S поразному. Относительные размеры S и B могут варьироваться. Идея использования фоновых тем состоит в сборе всех неинтересных слов, как это было продемонстрировано в [5]. Отличие от описанной работы состоит том, что используется не одна, а много фоновых тем, затем, чтобы как можно лучше очистить предметные темы, сделать их более этно-релевантными и повысить общее качество модели

3.2 Сглаживание и разреживание

Наиболее естественный способ внедрения априорной информации в модель состоит в использовании регуляризаторов сглаживания и разреживания с равномерным распределением β следующего вида:

$$\beta_w = \frac{1}{|Q|}[w \in Q].$$

Основным регуляризатором является LDA-подобный регуляризатор сглаживания, поощряющий появление этнонимов $w \in Q$ в этнических темах S. Также полезен регуляризатор разреживания с противоположным знаком, действующий на

[1] `http://bigartm.org`

фоновых темах, т.е. предотвращающий появление этнонимов в них:

$$R(\Phi) = \tau_1 \sum_{t \in S} \sum_{w \in Q} \ln \phi_{wt} - \tau_2 \sum_{t \in B} \sum_{w \in Q} \ln \phi_{wt}.$$

В задаче разведочного поиска предполагается, что доля релевантного содержимого в коллекции незначительна. В описываемой задаче ситуация именно такова, весь этно-социальный дискурс содержится в не более чем одном проценте от общего объёма коллекции. Задача состоит в том, чтобы хорошо описать тематическую структуру релевантного контента большим числом небольших, но качественных тем S. В то же время, тематическая модель должна описывать гораздо больший по объёму контент меньшим числом фоновых тем B.

Эти требования формализуются в терминах сглаживающего регуляризатора матрицы Θ, работающего только с фоновыми темами, и регуляризатора, равномерно разреживающего этническими темы в этой же матрице:

$$R(\Theta) = \tau_3 \sum_{d \in D} \sum_{t \in B} \ln \theta_{td} - \tau_4 \sum_{d \in D} \sum_{t \in S} \ln \theta_{td}.$$

Идея этого регуляризатора Θ заключается в сглаживании фоновых тем для того, чтобы они оттянули на себя как можно больше нерелевантных слов, и разреживании этнических тем в надежде на то, что они станут более непохожими друг на друга.

3.3 Декорреляция тем

Повышение различности распределений слов в темах приводит к росту интерпретируемости тем [20].

Для того, чтобы сделать темы настолько различными, насколько это возможно, используется регуляризатор максимизации ковариаций между столбцами ϕ_t для всех этнических тем t:

$$R(\Phi) = -\tau_5 \sum_{t \in S} \sum_{s \in S \setminus t} \sum_{w \in W} \phi_{wt} \phi_{ws} + \tau_6 \sum_{t \in B} \sum_{w \in W} \ln \phi_{wt}.$$

Декоррелятор также стимулирует разреженность и имеет тенденцию к группировке общих слов в отдельные темы [20]. Для того, чтобы эти темы образовались среди фоновых, в не предметных, применяется ещё один аддитивный регуляризатор, равномерно сглаживающий все фоновые темы B .

3.4 Модальность этнонимов

В качестве альтернативного метода регуляризации на основе априорной лексической информации предлагается использование этнонимов в качестве отдельной модальности. В общем случае, модальность — это тип терминов в документе. Примерами модальностей могут служить именованные сущности, теги, иностранные слова, n-граммы, авторы, категории, метки времени, ссылки и т.п. Каждая модальность имеет свой собственный словарь и свою матрицу Φ, нормализуемую отдельно от матриц Φ других модальностей. Мультимодальное расширение APTM было предложено в [24] и уже реализовано в BigARTM.

Используются две модальности: обычные слова и этнонимы. Модальность этнонимов определяется словарём Q и матрицей $\tilde{\Phi}$ размера $|Q| \times |T|$. В APTM логарифм правдоподобия модальности рассматривается как регуляризатор:

$$R(\tilde{\Phi}, \Theta) = \tau_7 \sum_{d \in D} \sum_{w \in Q} n_{dw} \ln \sum_{t \in T} \tilde{\phi}_{wt} \theta_{td},$$

где коэффициент регуляризации τ_7 по сути является множителем счётчиков «слово-документ» n_{dw} второй модальности.

Для того, чтобы сделать этно-релевентные темы более различными с точки зрения отражаемых в них подмножеств этнонимов, задействован дополнительный регуляризатор декорреляции тем, действующий на модальности этнонимов:

$$R(\tilde{\Phi}) = -\tau_8 \sum_{t \in S} \sum_{s \in S \setminus t} \sum_{w \in Q} \tilde{\phi}_{wt} \tilde{\phi}_{ws}.$$

Следует отметить, что декорреляция предметных тем S вводится отдельно для слов матрицы Φ и модальности этнонимов с матрицей $\tilde{\Phi}$.

3.5 Модальности меток времени и геотегов

Наличие информации о привязке текстов к меткам времени и геотегам может быть также использовано при обучении моделей с помощью описанного выше механизма мультимодальности. Это приносит двойную пользу: во-первых, дополнительная информация может быть использована алгоритмом для построения более качественной модели; во-вторых, в результате моделирования пользователь получает не только информацию о составе тем, но и том, как они изменяются в пространстве и во времени. Последнее свойство особенно ценно в рамках проводимого исследования.

Алгоритм 4.1. ProcessDocument(d, Φ)

Входные данные: документ $d \in D$, марица $\Phi = (\phi_{wt})$;
Выходные данные: матрица $(\tilde{n}_{wt})$, вектор (θ_{td}) для документа d;

1 инициализировать $\theta_{td} := \frac{1}{|T|}$ для всех $t \in T$;
2 **повторять**
3 $\quad p_{tdw} := \underset{t \in T}{\text{norm}}\big(\phi_{wt}\theta_{td}\big)$ для всех $w \in d$ и $t \in T$;
4 $\quad \theta_{td} := \underset{t \in T}{\text{norm}}\big(\sum_{w \in d} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\big)$ для всех $t \in T$;
5 **до тех пор, пока** θ_d не сойдётся;
6 $\tilde{n}_{wt} := n_{dw} p_{tdw}$ для всех $w \in d$ и $t \in T$;

4 Библиотека BigARTM

Данный раздел будет посвящён описанию библиотеки тематического моделирования больших текстовых коллекций BigARTM. Будет подробно описан существующий вариант EM-алгоритма, лежащего в её основе, а также предложен новый, более совершенный алгоритм.

4.1 Обзор

BigARTM — это библиотека с открытым программным кодом для построения регуляризованных мультимодальных тематических моделей больших текстовых коллекций. Полностью поддерживая теорию APTM, библиотека предоставляет предопределённый набор регуляризаторов, а также метрик качества моделирования, оставляя пользователю возможность добавлять собственные. Написана на `C++11`, имеет пользовательский API на `Python`. В BigARTM реализован параллельный онлайновый асинхронный EM-алгоритм, обладающий высокой производительностью в рамках одного вычислительного узле. Данные для библиотеки сохраняются на диске или в памяти в виде специальных пакетов (*батчей*), которые можно генерировать с помощью встроенного парсера. Каждый батч содержит некоторое количество документов, и является атомарной порцией данных для обработки одним потоком.

4.2 Реализации EM-алгоритма

Оффлайновый алгоритм Базовым вариантом EM-алгоритма для модели APTM является оффлайновый алгоритм (4.2). Он основывается на функции `ProcessDocument` (4.1), которая соответствует уравнениям 5, 7 решения задачи APTM. `ProcessDocument`

Алгоритм 4.2. Offline ARTM

Входные данные: коллекция D;
Выходные данные: матрица $\Phi = (\phi_{wt})$;

1 инициализировать (ϕ_{wt});
2 создать батчи $D := D_1 \sqcup D_2 \sqcup \cdots \sqcup D_B$;
3 **повторять**
4 $\quad (n_{wt}) := \sum_{b=1,\ldots,B} \sum_{d \in D_b} \mathsf{ProcessDocument}(d, \Phi)$;
5 $\quad (\phi_{wt}) := \operatorname*{norm}_{w \in W}(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}})$;
6 **до тех пор, пока** (ϕ_{wt}) не сойдётся;

требует на вход фиксированную матрицу Φ и вектор n_{dw} частот слов для заданного документа $d \in D$. Выходными данными являются распределение на темах данного документа θ_{td} и матрица $\hat{n}_{wt}$ размера $|d| \times |T|$, где $|d|$ обозначает число уникальных слов в документе d.

`ProcessDocument` может также быть полезной как отдельная операция, позволяющая получать векторы θ_{td} для новых документов, но в оффлайновом алгоритме она используется в качестве базового блока обработки в EM-алгоритме, и предназначен для вычисления обновлений матрицы Φ.

`Offline ARTM` проходит по всей коллекции текстов, вызывая функцию `ProcessDocument` для каждого документа $d \in D$, а затем агрегирует результирующие матрицы $\hat{n}_{wt}$ в итоговую матрицу n_{wt} размера $|W| \times |T|$.

После каждого прохода по коллекции матрицы Φ обновляется в соответствии с уравнением 6.

На шаге 2 производится разделение коллекции D на батчи D_b[2]. В целях повышения производительности внешний цикл по батчам $b = 1, \ldots, B$ распараллеливается по нескольким потокам, и внутри каждого батча внутренний цикл по документам $d \in D_b$ выполняется в рамках одного потока.

Стоит обратить внимание на то, что значения θ_{td} появляются лишь внутри функции `ProcessDocument`. Это приводит к эффективному использованию памяти, поскольку реализация никогда не хранит целую матрицу Θ. Вместо этого значения θ_{td} пересчитываются с нуля на каждом проходе по коллекции.

На рис. 1 можно увидеть диаграмму Гантта для `Offline ARTM`. В этой и по-

[2]Этот шаг не является обязательным для самого оффлайнового алгоритма, это часть работы библиотеки

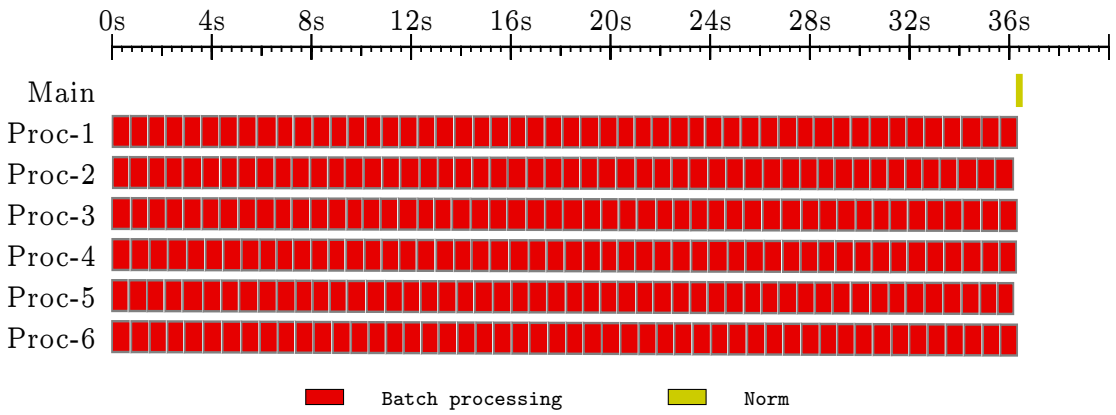


Рис. 1: Диаграмма Гантта `Offline ARTM` (алгоритм 4.2)

следующей диаграммах показана одна итерация EM-алгоритма на данных `NYTimes`[3] ($|D| = 300\text{K}, |W| = 102\text{K}$) в модели с $|T| = 16$ темами. Прямоугольники `ProcessBatch` соответствуют времени, потраченному на обработку одного батча. Финальный прямоугольник `Norm`, выполняющийся в главном потоке, соответствует времени, затраченному на шаг 4 в алг. 4.2, где счётчики n_{wt} нормируются для создания новой матрицы Φ.

Синхронный онлайновый алгоритм `Online ARTM` (алгоритм 4.3) является обобщением онлайнового вариационного EM-алгоритма, предложенного в [8] для модели LDA. Онлайновый алгоритм улучшает сходимости оффлайнового за счёт пересчёта матрицы Φ, производимого не в конце обработки всей коллекции, а в конце обработки некоторой порции батчей. Для упрощения обозначений введём следующую тривиальную функцию:

$$\texttt{ProcessBatches}(\{D_b\}, \Phi) = \sum_{D_b} \sum_{d \in D_b} \texttt{ProcessDocument}(d, \Phi).$$

Она агрегирует результаты `ProcessDocument` для заданного множества батчей при фиксированной матрице Φ.

В онлайновом алгоритме разбиение коллекции $D := D_1 \sqcup D_2 \sqcup \cdots \sqcup D_B$ на батчи играет гораздо более важную роль, чем в алгоритме оффлайновом, поскольку различные разбиения будут приводить к различным результатам.

На шаге 6 новые значения n_{wt}^{i+1} вычисляются как выпуклая комбинация старых

[3] `https://archive.ics.uci.edu/ml/datasets/Bag+of+Words`

Алгоритм 4.3. Online ARTM

Входные данные: коллекция D, гиперпараметры η, τ_0, κ;
Выходные данные: matrix $\Phi = (\phi_{wt})$;

1 создать батчи $D := D_1 \sqcup D_2 \sqcup \cdots \sqcup D_B$;
2 инициализировать (ϕ^0_{wt});
3 **цикл** $i = 1, \ldots, \lfloor B/\eta \rfloor$ **выполнять**
4 $\quad (\hat{n}^i_{wt}) := \mathsf{ProcessBatches}(\{D_{\eta(i-1)+1}, \ldots, D_{\eta i}\}, \Phi^{i-1})$;
5 $\quad \rho_i := (\tau_0 + i)^{-\kappa}$;
6 $\quad (n^i_{wt}) := (1 - \rho_i) \cdot (n^{i-1}_{wt}) + \rho_i \cdot (\hat{n}^i_{wt})$;
7 $\quad (\phi^i_{wt}) := \underset{w \in W}{\mathrm{norm}}(n^i_{wt} + \phi^{i-1}_{wt} \frac{\partial R}{\partial \phi_{wt}})$;

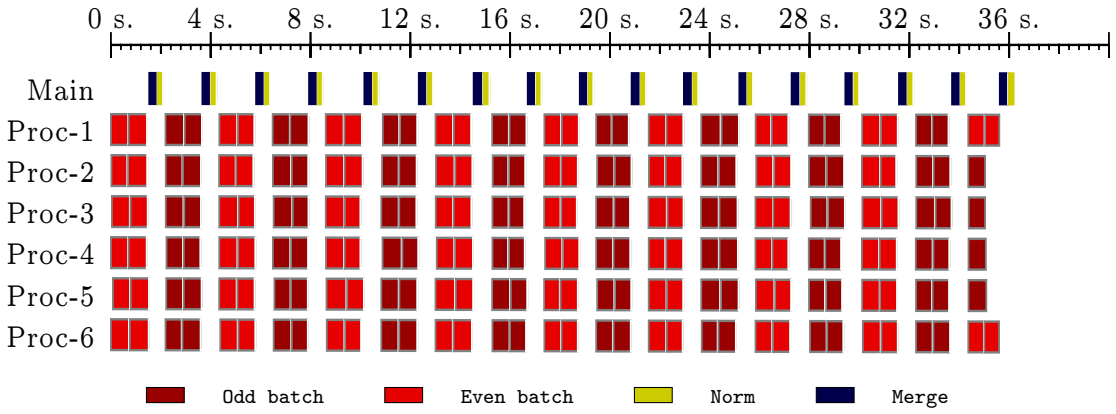


Рис. 2: Диаграмма Гантта `Online ARTM` (алг. 4.3)

значений n^i_{wt} и значения $\hat{n}^i_{wt}$, полученного по только что обработанным батчам. Старые счётчики n^i_{wt} умножаются на множитель $(1-\rho_i)$, зависящий от номера итерации. Общепринятая стратегия состоит в использовании $\rho_i = (\tau_0 + i)^{-\kappa}$б где стандартные значения τ_0 лежат в диапазоне от 64 до 1024, а κ — от 0.5 до 0.7.

Так же, как и в оффлайновом алгоритме, внешний цикл по батчам $D_{\eta(i-1)+1}, \ldots, D_{\eta i}$ выполняется параллельно в нескольких потоках. Проблема такого подхода в том, что во время шагов 5-7 алг. `Online ARTM` рабочие потоки простаивают.

Потоки не могут начать обработку следующей порции батчей, поскольку новая версия матрицы Φ ещё не готова. Результатом этого является неэффективное использование процессорных ресурсов, обычная диаграмма Гантта для алгоритма `Online ARTM` показана на рис. 2.

Прямоугольники `Even batch` и `Odd batch` оба соответствуют шагу 4 и обозначают версию матрицы Φ^i (чётное i или нечётное i). Прямоугольник `Merge` соответствует времени, затраченному на слияние n_{wt} с $\hat{n}_{wt}$. `Norm`, как и выше, обозначает

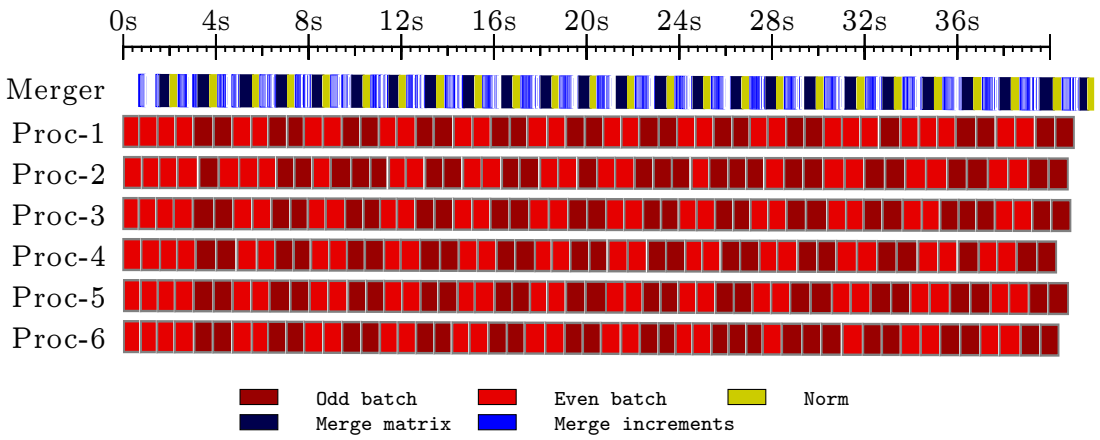


Рис. 3: Диаграмма Гантта `Async ARTM` — нормальная ситуация

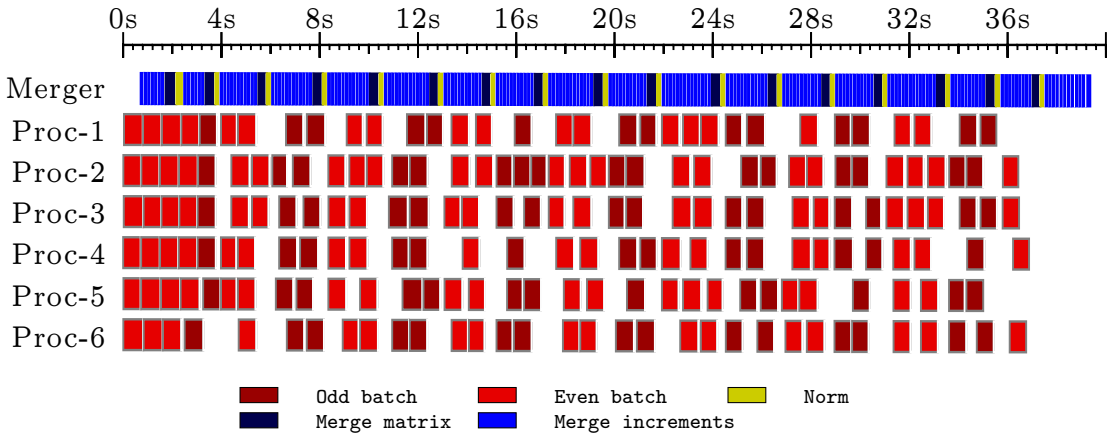


Рис. 4: Диаграмма Гантта `Async ARTM` — проблемы производительности

время, затраченное на нормализацию счётчиков n_{wt} для получения новой матрицы Φ, которая будет использована во время следующей итерации.

Алгоритм Async Алгоритм `Async ARTM` [24] был призван решить проблемы обычного онлайнового синхронного алгоритма, описанные выше. Идея заключается в организации асинхронной работы `Offline ARTM` и сохранении результирующих матриц $\hat{n}_{wt}$ в очередь. Затем, в тот момент, когда количество матриц в очереди достигает заданного η, алгоритм производит шаги 5-7 алгоритма `Online ARTM` (алг. 4.3).

Из соображений производительности слияние счётчиков $\hat{n}_{wt}$ производится в фоновом режиме выделенным потоком слияния `Merger`.

Как было сказано выше, данный алгоритм показал высокую производительность и масштабируемость в сравнении с аналогичными инструментами [24], но у него

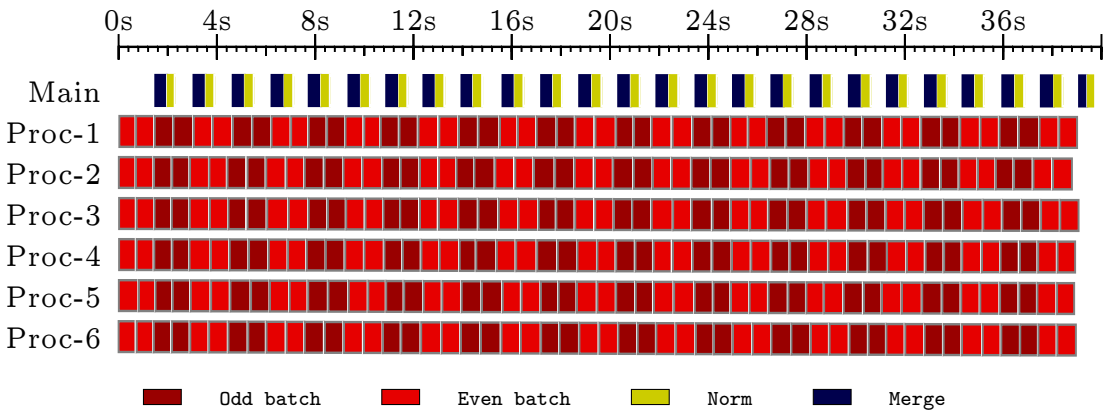


Рис. 5: Диаграмма Гантта для `DetAsync ARTM` (алг. 4.4)

имеется ряд недостатков.

Первая проблема состоит в том, что алгоритм не детерминирует порядка слияния счётчиков $\hat{n}_{wt}$. Этот порядок обычно отличается от порядка обработки батчей и меняется от запуска к запуску. Это приводит к тому, что и результирующая матрица Φ от запуска к запуску может быть различной.

Другая проблема, связанная с `Async ARTM`, состоит в том, что хранение счётчиков $\hat{n}_{wt}$ в очереди может существенно увеличить потребление памяти и привести к тому, что поток `Merger` станет узким местом алгоритма с точки зрения производительности.

При правильном подборе параметров эффективность подобного алгоритма будет высокой, что можно видеть на диаграмме 3. Однако несложно подобрать и такой набор параметров (например, слишком малый размер батча или маленькое число внутренних итераций в `ProcessDocument`), который приведёт к перегрузке потока слияния. В такой ситуации диаграмма Гантта примет вид, показанный на рис. 4: большинство потоков простаивают, поскольку в очереди нет места для новых счётчиков n_{wt}.

В следующем разделе данные проблемы будут решены с помощью предлагаемого алгоритма `DetAsync`, который является полностью детерминированным и позволяет производить обучение в онлайновом режиме с высокой производительностью без необходимости тонкой настройки параметров пользователем.

4.3 Онлайновый алгоритм DetAsync

Описание `DetAsync ARTM` [10] (алг. 4.4) основан на двух новых функциях, `Await` и `AsyncProcessBatches`.

Алгоритм 4.4. DetAsync ARTM

Входные данные: коллекция D, параметры η, τ_0, κ;
Выходные данные: matrix $\Phi = (\phi_{wt})$;

1 создать батчи $D := D_1 \sqcup D_2 \sqcup \cdots \sqcup D_B$;
2 инициализировать (ϕ^0_{wt});
3 $F^1 := \mathsf{AsyncProcessBatches}(\{D_1, \ldots, D_\eta\}, \Phi^0)$;
4 **цикл** $i = 1, \ldots, \lfloor B/\eta \rfloor$ **выполнять**
5 **если** $i \neq \lfloor B/\eta \rfloor$ **тогда**
6 $F^{i+1} := \mathsf{AsyncProcessBatches}(\{D_{\eta i+1}, \ldots, D_{\eta i+\eta}\}, \Phi^{i-1})$;
7 $(\hat{n}^i_{wt}) := \mathsf{Await}(F^i)$;
8 $\rho_i := (\tau_0 + i)^{-\kappa}$;
9 $(n^i_{wt}) := (1 - \rho_i) \cdot (n^{i-1}_{wt}) + \rho_i \cdot (\hat{n}^i_{wt})$;
10 $(\phi^i_{wt}) := \underset{w \in W}{\mathrm{norm}}(n^i_{wt} + \phi^{i-1}_{wt} \frac{\partial R}{\partial \phi_{wt}})$;

Вторая во всём эквивалентна описанной ранее `ProcessBatches`, за исключением того, что она берёт задачу на асинхронную обработку и немедленно возвращает управление в вызвавший поток. Её результатом является future-объект (например, `std::future` из стандарта `C++11`), который может быть затем передан в вызов `Await` для получения вычисленного результата, в нашем случае счётчиков $\hat{n}_{wt}$.

Между вызовами `AsyncProcessBatches` and `Await` алгоритм может производить различную вспомогательную работу, пока рабочие потоки в фоновом режиме производят вычисление матрицы $\hat{n}_{wt}$.

Для вычисления $\hat{n}^{i+1}_{wt}$ `DetAsync ARTM` использует матрицу Φ^{i-1} с предыдущего обновления. Это добавляет некоторое запаздывание между моментом вычисления очередной версии матрицы Φ и моментом её использования, что даёт алгоритму в результате дополнительную гибкость в распределении нагрузки на рабочие потоки. Шаги 3 и 5 — это технический трюк, направленный на реализацию описанной идеи с запаздыванием.

Добавление запаздывания может негативно сказаться на сходимости алгоритма в сравнении с `Online Async`. Например, в `AsyncProcessBatches` начальная матрица Φ^0 используется дважды, в то время как последние две матрицы $\Phi^{\lfloor B/\eta \rfloor - 1}$ и $\Phi^{\lfloor B/\eta \rfloor}$ вообще не будут использованы.

С другой стороны, асинхронный алгоритм позволет добиться более высокой степени загрузки ядер, что наглядно продемонстрировано на диаграмме 5.

В этом состоит некоторый компромисс между сходимостью и загрузкой CPU, и он будет рассмотрен подробнее в разделе 5.1.

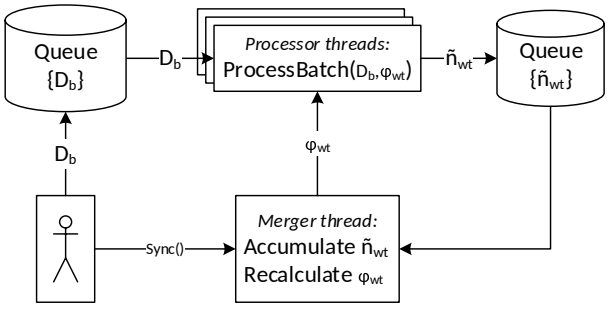


Рис. 6: Схема компонентов BigARTM (Async)

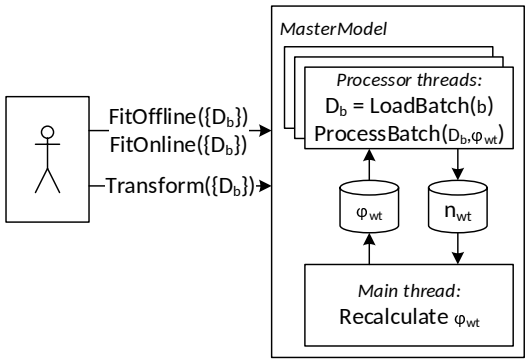


Рис. 7: Схема компонентов BigARTM (DetAsync)

Детали реализации Существенной частью реализации является способ агрегации матриц $\hat{n}_{wt}$ со всех батчей при условии того, что они обрабатываются разными потоками. Со сменой алгоритма в BigARTM этот способ изменился (рис. 6 и 7).

В старой архитектуре счётчики $\hat{n}_{wt}$ сохранялись в очередь, откуда агрегировались выделенным потоком `Merger`. В новой архитектуре этот поток ликвидирован, и счётчики $\hat{n}_{wt}$ пишутся напрямую в результирующую матрицу n_{wt} асинхронно всеми рабочими потоками. Для синхронизации доступа на запись необходимо обеспечить невозможность возникновения ситуации, в которой два потока одновременно производят запись в одну строку матрицы n_{wt}. Это достигается с помощью спин-локов l_w, по одному на каждое слово из словаря W. В конце вызова `ProcessDocument` производится итерирование по всем $w \in d$, для каждого слова производится блокировка соответствующего лока, добавление $\hat{n}_{wt}$ к n_{wt} и разблокировка лока. Этот подход схож с тем, что был предложен в [7], где подобная система была организована в распределённой среде.

В новой архитектуре также был ликвидирован выделенный поток загрузки данных `DataLoader`, который до этого загружал данные с диска в специализированную очередь задач, откуда батчи уже выбирались для обработки рабочими потоками. Теперь процесс загрузки производится непосредственно самим потоком, что упростило

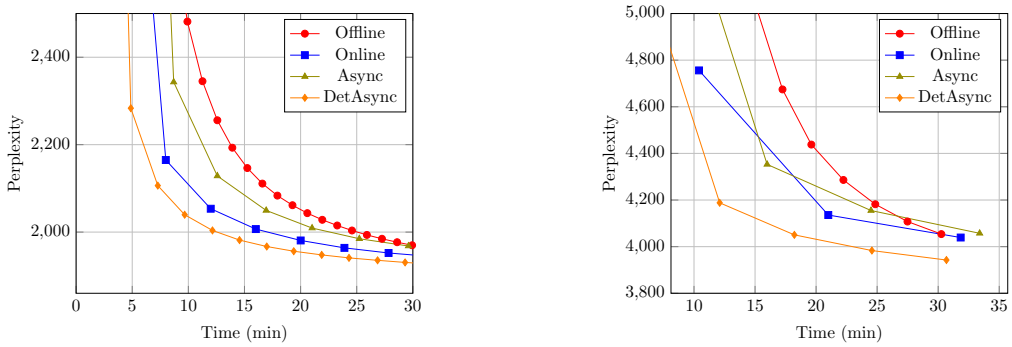


Рис. 8: График перплексии от времени работы для Pubmed (слева) и Wikipedia (справа), $|T| = 100$ topics

Таблица 1: Пиковое потребление памяти BigARTM, Гб

	$\|T\|$	Offline	Online	DetAsync	Async (v0.6)
Pubmed	1000	5.17	4.68	8.18	13.4
Pubmed	100	1.86	1.62	2.17	3.71
Wiki	1000	1.74	2.44	3.93	7.9
Wiki	100	0.54	0.53	0.83	1.28

архитектуру без потери производительности.

5 Эксперименты

5.1 Оценивание реализации DetAsync

В данном разделе производится сравнение эффективности алгоритмов Offline (алг. 4.2), Online (алг. 4.3), Async [24] и DetAsync (алг. 4.4).

Как уже было отмечено выше, алгоритм `Async` уже превосходит по скорости аналоги BigARTM [24]: в однопоточном режиме он почти в 10 раз быстрее Gensim [18] и вдвое быстрее Vowpal Wabbit LDA (VW) [19]; в многопоточном режиме превосходство ещё более выраженное.

В экспериментах этого раздела используются коллекция статей англоязычной Википедии (*Wikipedia*) ($|D| = 3.7$M статей, $|W| = 100$K слов в словаре) и коллекция аннотаций *Pubmed* ($|D| = 8.2$M аннотаций, $|W| = 141$K слов в словаре). Эксперименты производились на системе Intel Xeon CPU E5-2650 v2 с 2 процессорами, 16 физических ядер в совокупности (32 с hyper-threading).

Рис. 8 показывает *перплексию* как функцию от времени, потраченного описанными выше алгоритмами на обучение.

Перплексия определяется как

$$\mathscr{P}(D,p) = \exp\left(-\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw}\ln\sum_{t\in T}\phi_{wt}\theta_{td}\right), \tag{8}$$

где $n = \sum_d n_d$. Более низкое значение перплексии соответствует более хорошему результату. Каждая точка на графике соответствует моменту завершения алгоритмом очередного прохода по коллекции. Каждому алгоритму было выделено на работу 30 минут.

Таблица 1 демонстрирует пиковое потребление памяти каждый алгоритмом при обучении моделей с $|T| = 1000$ и $|T| = 100$ темами на коллекциях Wikipedia и Pubmed.

5.2 Эксперименты на коллекции LiveJournal

С социологической точки зрения, задача проекта заключается в поиске и мониторинге этно-релевантного дискурса в социальных сетях, в частности, в определении степени популярности тем, связанных с теми или иными этническими группами, возможно, в заданных регионах, и выявлении зарождающихся негативных тенденций, могущих повлечь за собой возникновение конфликта на этнической почве. В данном разделе производится построение регуляризованных тематических моделей коллекции постов самой популярной российской блог-платформы LiveJournal [13].

Данные и параметры Коллекция содержит примерно 1.58M лемматизированных постов, написанных топ-2000 блоггерами LiveJournal за годичный период с середины 2013 до середины 2014. Полный словарь коллекции составил примерно 860K слов, но после предобработки, во время которой были сохранены только слова, которые одновременно содержат только символы русского алфавита, с не более, чем одним дефисом; имеют длину не менее 3 символов; встречаются во всей коллекции как минимум 20 раз.

В процессе подбора числа тем были опробованы 100, 300 и 400 тем. В результате работы экспертов по оцениванию качества моделирования было выявлено, что наилучший результат достигается при $|T| = 400$, поэтому именно это число тем использовалось во всех дальнейших экспериментах с этой коллекцией. Это соответствует более ранним экспериментам [4, 15].

Коллекция была разделена на батчи по 10000 документов в каждом. Все модели АРТМ обучались с помощью онлайнового алгоритма с одним проходом по коллек-

ции и 25 проходами по каждому документу; обновления матрицы Φ производились после каждого обработанного батча. Для регуляризатора частичного обучения был подготовлен набор из нескольких сотен этнонимов — существительных, обозначающих различные этнические группы; 249 из этих слов встретились в коллекции.

Этнонимы выглядят лучшими кандидатами на роль средства улучшения качества извлечения тем, связанных с этничностями и межэтничными отношениями. Участниками таких отношений являются конкретные люди или группы людей. Нужно отличать их от отношений международных, в которых основную роль играют государства, их правительства или официальные представители, а затрагиваемые вопросы далеко не всегда касаются этничностей. Межэтничные и международные отношения тесно связаны и, в некоторых ситуациях, пересекаются, однако, интуитивно ясно, что для мониторинга и предотвращения конфликтов на этнической почве (связанных, например, с мигрантами) логичнее анализировать блогосферу, чем новости официальных источников.

Предполагается, что в этнических темах будут превалировать этнонимы (турки), в то время как прилагательные (турецкий) и названия стран (Турция) более связаны с международными отношениями. В русском языке эти три категории, как правило, представляют собой различные слова, что позволяет проще классифицировать темы по рассматриваемым отношениям на международные и межэтничные.

Модели В экспериментах с использованием BigARTM были обучены наборы тематических моделей. Во всех моделях с гиперпараметрами коэффициенты регуляризации подбирались вручную в ходе многократных запусков обучения. Во всех моделях с регуляризацией темы были разделены на $|S| = 250$ предметных и $|B| = 150$ фоновых.

Далее приведён список различных моделей, которые были настроены и сравнены:

1. **plsa**: базовая модель вероятностного латентного семантического анализа (PLSA) без регуляризаторов;

2. **lda**: базовая модель латентного размещения Дирихле (LDA), реализованная в BigARTM как модель с регуляризаторами сглаживания Φ и Θ равномерными распределениями α и β с гиперпараметрами $\alpha_0 = \beta_0 = 10^{-4}$;

3. **smooth**: модель APTM со сглаживанием и разреживанием по этнонимам, с коэффициентами регуляризации $\tau_1 = 10^{-5}$ and $\tau_2 = 100$; кроме того, в этом и всех последующих экспериментах использовался описанный ранее регуляризатор матрицы Θ с коэффициентами $\tau_3 = 0.05$ and $\tau_4 = 1$;

4. **decorrelated**: модель APTM, обобщающая предыдущую путём добавления декорреляции с параметрами $\tau_5 = 5 \times 10^4$ and $\tau_6 = 10^{-8}$; коэффициент сглаживания этнических тем $\tau_1 = 10^{-6}$;

5. **restricted dictionary**: модель APTM, обобщающая предыдущую путём добавления декоррелируемой модальности этнонимов с коэффициентами $\tau_7 = 100$ and $\tau_8 = 2 \times 10^4$; Другие коэффициенты приняли следующие значения: $\tau_5 = 1.5 \times 10^6$, $\tau_6 = 10^{-7}$ и $\tau_1 = 1.1 \times 10^{-4}$; в этой модели использовался словарь из $|Q| = 249$ этнонимов;

6. **extended dictionary**: модель APTM, идентичная предыдущей, в которой использовался расширенный словарь: помимо этнонимов, в него были добавлены прилагательные и названия стран для тех этничностей, для которых соответствующего этнонима в коллекции не нашлось;

7. **recursive**: базовая модель PLSA, обученная на специальном подмножестве документов, полученных из тем модели 5, которые были сочтены этническими экспертами: использовались все документы, которые в данных темах в матрице Θ имели вероятность выше порога 10^{-6};

8. **keyword documents**: модель PLSA, идентичная предыдущей, но обученная на подмножестве документов всей коллекции, содержащих хоть один этноним из Q.

Модели 7 и 8 были обучены для сравнения двух методов обогащения исходной коллекции. Модель 8 использовалась в качестве базовой при проверке предположения о том, что циклическое использование тематических моделей может дать лучший результат, чем извлечение текстов по ключевым словам.

Результаты В этом разделе обсуждаются количественные и качественные результаты обучения. Сперва будет описана методология экспертного оценивания, затем будет проведено обсуждение полученных оценок. Кроме того, результаты оценивания людьми будут сравнены со значениями tf-idf когерентности, предложенной ранее в [15, 4]. Было показано, что такая метрика лучше коррелирует с человеческими оценками, чем традиционная когерентность [14].

Результаты измерения средних когерентности и tf-idf когерентности для каждой модели показаны в таблице 2; представлены две версии метрик когерентности, посчитанные на топ-10 и топ-20 словах в каждой теме.

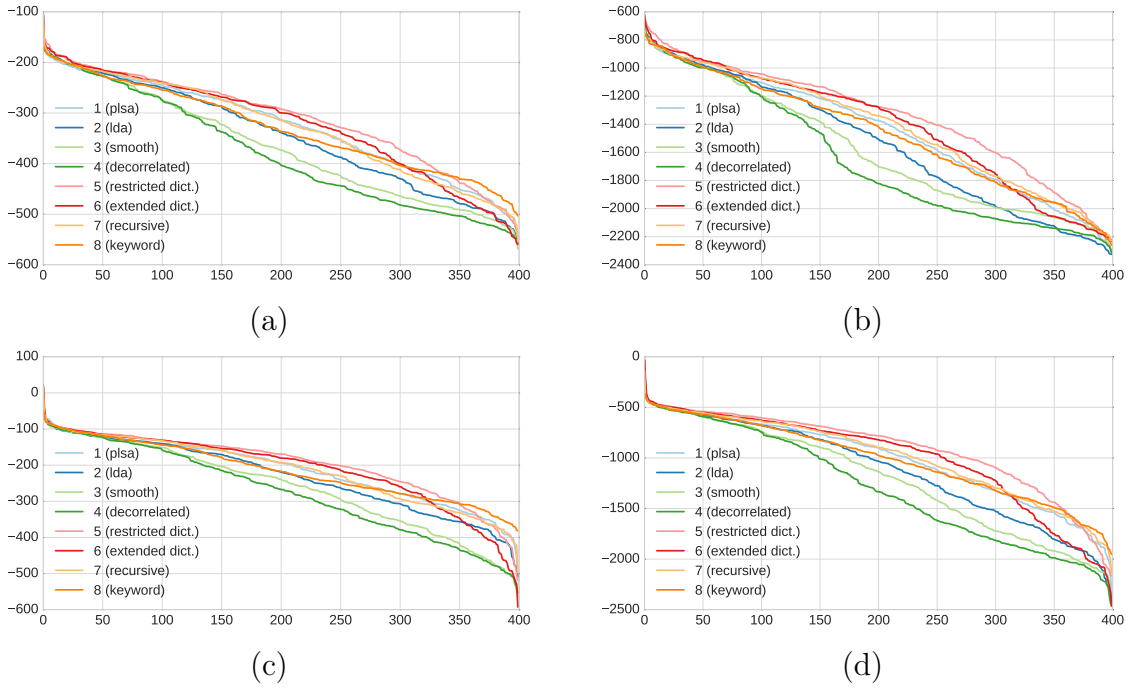


Рис. 9: Сортированные метрики качества тем: (a) coh_{10}; (b) tfidf_{10}; (c) coh_{20}; (d) tfidf_{20}.

Модель	T	coh_{10}	tfidf_{10}	coh_{20}	tfidf_{20}
1 (plsa)	400	-325.3	-212.0	-1447.0	-1011.6
2 (lda)	400	-344.2	-230.9	-1539.8	-1121.2
3 (smooth)	400	-367.1	-261.2	-1583.9	-1210.2
4 (decorr)	400	-378.9	-274.0	-1651.2	-1296.1
5 (restr. dict.)	400	**-310.0**	**-196.4**	**-1341.9**	**-908.4**
6 (ext. dict.)	400	-321.7	-209.6	-1409.1	-995.3
7 (recursive)	400	-326.5	-212.1	-1415.6	-982.5
8 (keyword)	400	-328.8	-214.4	-1463.6	-1014.5

Таблица 2: Средние когерентность и tf-idf когерентность для всех обученных моделей.

Вопрос	Разница
1 (general understanding)	0.28
2 (event/phenomenon)	0.30
3 (ethnonyms)	0.07
4 (ethnic issues)	0.06
5 (international relations)	0.08
6 (other)	0.25

Таблица 3: Согласованность кодировщиков: общая доля различных ответов.

Распределения всех четырёх метрик также показаны в деталях на рис. 9, демонстрирующем отсортированные метрики (coh_{10}, tfidf_{10}, coh_{20}, и tfidf_{20}) для каждой из моделей, и график, идущий выше всех остальных, соответствует лучшей модели. Таблица 2 и рис. 9 показывают, что хотя модели 5 (restricted dictionary) и 6 (extended dictionary) побеждают во всех четырёх случаях, все остальные модели имеют сопоставимые результаты, кроме моделей 3 (smooth) и 4 (decorrelated). Это было подтверждено предварительными оценками людей, поэтому было принято решение исключить эти две модели из дальнейшего рассмотрения для более полезного использования ограниченного количества человеческих ресурсов.

Для всех моделей экспертам было предложено интерпретировать каждую тему в каждой модели по топ-20 словам этой темы. Для каждой темы два эксперта отвечали на следующие вопросы, связанные с качеством и степенью этничности темы; на

	#	coh_{10}	$tfidf_{10}$	coh_{20}	$tfidf_{20}$
	Частично интерпретируемые темы				
1 (plsa)	139	-258.7	-145.3	-1145.9	-696.9
2 (lda)	192	-274.9	-163.3	-1224.1	-777.5
5 (restricted dict.)	237	-284.6	-163.0	-1247.9	-768.8
6 (extended dict.)	146	-258.6	-141.2	-1156.0	-686.1
7 (recursive)	239	-281.9	-166.3	-1235.7	-788.1
8 (keyword)	114	-256.3	-140.2	-1141.4	-682.8
	Хорошо интерпретируемые темы				
1 (plsa)	119	-318.0	-206.6	-1414.7	-982.5
2 (lda)	120	-389.5	-273.1	-1743.7	-1324.6
5 (restricted dict.)	87	-330.7	-227.0	-1410.7	-1028.2
6 (extended dict.)	103	-313.8	-199.9	-1372.6	-936.4
7 (recursive)	58	-349.2	-241.1	-1498.1	-1086.1
8 (keyword)	106	-310.0	-198.9	-1354.3	-914.8
	Обе группы тем вместе				
1 (plsa)	258	-286.0	-173.6	-1269.9	-828.7
2 (lda)	312	-319.0	-205.5	-1424.0	-988.0
5 (restricted dict.)	324	-297.0	-180.2	-1291.6	-838.5
6 (extended dict.)	249	-281.5	-165.5	-1245.6	-789.6
7 (recursive)	297	-295.1	-180.9	-1287.0	-846.3
8 (keyword)	220	-282.2	-168.5	-1244.0	-794.6

Таблица 4: Экспериментальные результаты: средние интерпретируемости и когерентности разных групп тем.

каждый вопрос требовалось дать один из трёх ответов: «нет», «частично» и «да»:

1. Понятно ли Вам, почему эти слова собрались вместе в данной теме?
2. Если в вопросе 1 Вы дали ответ «частично» или «да»: понятно ли Вам, какое явление или события может описываться в текстах, связанных с этой темой?
3. Есть ли среди топ-слов этнонимы? Укажите количество.
4. Если в вопросе 2 Вы дали ответ «частично» или «да»: связано ли это событие с этничностями?
5. Если в вопросе 2 Вы дали ответ «частично» или «да»: связано ли это событие с международными отношениями?
6. Если в вопросе 2 Вы дали ответ «частично» или «да»: связано ли это явление или событие с другой темой, не имеющей отношения к этничностям?

Эксперты были проинструктированы по всем вопросам, включая различия между межэтничными и межнациональными отношениями. Были собраны ответы семи экспертов; таблица 3 суммирует значения общего согласия экспертов, демонстрируя

Темы	Релевантные темы														
	частично					хорошо					обе группы				
	#	coh_{10}	$tfidf_{10}$	coh_{20}	$tfidf_{20}$	#	coh_{10}	$tfidf_{10}$	coh_{20}	$tfidf_{20}$	#	coh_{10}	$tfidf_{10}$	coh_{20}	$tfidf_{20}$
1 (plsa)															
ethnic	5	-313.2	-190.2	-1399.2	-904.8	12	-334.0	-207.1	-1480.9	-996.3	17	-327.9	-202.1	-1456.9	-969.4
IR	20	-279.1	-150.7	-1227.0	-733.8	19	-315.3	-194.0	-1410.7	-946.8	39	-296.8	-171.8	-1316.5	-837.6
all relev.	20	-289.6	-163.0	-1271.2	-784.9	25	-315.9	-194.3	-1408.0	-938.7	45	-304.2	-180.4	-1347.2	-870.3
2 (lda)															
ethnic	2	-239.7	-124.4	-1158.5	-646.0	13	-306.8	-190.0	-1369.1	-927.9	15	-297.9	-181.3	-1341.0	-890.3
IR	21	-285.1	-158.9	-1266.2	-763.1	29	-353.3	-225.7	-1580.6	-1097.5	50	-324.7	-197.7	-1448.6	-957.1
all relev.	18	-289.4	-162.3	-1287.3	-777.7	37	-336.3	-212.2	-1496.3	-1023.0	55	-320.9	-195.9	-1427.9	-942.7
5 (restricted dictionary)															
ethnic	18	-288.7	-164.7	-1264.2	-798.5	30	-331.6	-222.3	-1419.0	-1015.8	48	-315.5	-200.7	-1360.9	-934.3
IR	33	-269.1	-142.5	-1190.8	-707.7	26	-323.1	-207.4	-1358.1	-917.3	59	-292.9	-171.1	-1264.5	-800.1
all relev.	36	-267.2	-142.0	-1177.6	-695.1	47	-322.7	-211.1	-1374.5	-958.4	83	-298.7	-181.1	-1289.1	-844.2
6 (extended dictionary)															
ethnic	8	-288.4	-160.5	-1315.2	-805.1	22	-280.7	-150.0	-1226.8	-713.8	30	-282.8	-152.8	-1250.4	-738.2
IR	18	-250.0	-126.3	-1130.6	-641.1	29	-287.4	-156.3	-1240.9	-740.8	47	-273.1	-144.8	-1198.7	-702.6
all relev.	22	-261.2	-136.5	-1199.9	-707.7	37	-285.5	-158.3	-1234.6	-741.8	59	-276.4	-150.2	-1221.7	-729.1
7 (recursive)															
ethnic	18	-308.2	-181.3	-1418.7	-952.6	22	-320.1	-201.8	-1431.0	-971.4	40	-314.7	-192.6	-1425.5	-962.9
IR	30	-283.3	-161.6	-1236.8	-780.4	30	-291.4	-171.4	-1292.9	-827.3	60	-287.4	-166.5	-1264.9	-803.9
all relev.	34	-285.4	-161.3	-1269.0	-810.6	47	-299.0	-180.1	-1331.3	-869.8	81	-293.3	-172.2	-1305.1	-844.9
8 (keyword)															
ethnic	5	-289.7	-161.1	-1315.9	-805.0	37	-297.9	-175.6	-1318.9	-834.7	42	-297.0	-173.9	-1318.6	-831.1
IR	18	-264.7	-138.4	-1168.7	-670.7	32	-278.5	-164.3	-1240.7	-782.9	50	-273.5	-155.0	-1214.8	-742.5
all relev.	17	-279.5	-154.3	-1230.7	-741.3	52	-282.5	-165.5	-1260.1	-793.1	69	-281.8	-162.8	-1252.8	-780.4

Таблица 5: Релевантность и когерентность тем.

	Вопрос 1			Вопрос 2			Вопрос 1			Вопрос 2		
Релев. темы	част.	хор.	все	част.	хор.	все	част.	хор.	все	част.	хор.	все
	1 (plsa)						6 (extended dictionary)					
ethnic	1.80	1.75	1.76	1.20	1.50	1.41	2.00	1.73	1.80	1.75	1.27	1.40
IR	1.90	1.68	1.79	1.75	1.26	1.51	1.94	1.72	1.81	1.72	1.17	1.38
all relev.	1.85	1.72	1.78	1.65	1.36	1.49	1.95	1.62	1.75	1.68	1.16	1.36
	2 (lda)						7 (recursive)					
ethnic	2.00	1.92	1.93	2.00	1.62	1.67	1.78	1.59	1.68	1.00	0.95	0.97
IR	2.00	1.69	1.82	1.86	1.21	1.48	1.87	1.87	1.87	1.43	1.20	1.32
all relev.	2.00	1.76	1.84	1.83	1.32	1.49	1.94	1.72	1.81	1.35	1.09	1.20
	5 (restricted dictionary)						8 (keyword)					
ethnic	2.00	1.40	1.62	1.89	1.27	1.50	2.00	1.76	1.79	1.20	0.89	0.93
IR	1.85	1.42	1.66	1.85	1.35	1.63	1.94	1.91	1.92	1.33	1.16	1.22
all relev.	1.89	1.45	1.64	1.86	1.32	1.55	1.94	1.83	1.86	1.41	1.08	1.16

Таблица 6: Результаты интерпретации тем, связанных с межэтничными и международными отношениями.

доли различающихся ответов для каждого вопроса. В целом, эти результаты показывают хорошую согласованность между экспертами, по сравнению с более ранними экспериментами по аналогичной проблеме [17]. В случае разногласия экспертов вместо усреднения их ответов генерировались два набора метрик: с максимальным и с минимальными ответами соответственно. Таким образом определялись верхняя и нижняя границы человеческого восприятия качества моделей.

Для каждой модели таблица 4 также демонстрирует среднее значение tf-idf когерентности. Поскольку тренируемые модели пытались извлечь заданное число тем высокого качества, отодвигая «мусорные» темы в фон, не имеет особого смысла производить сравнение моделей по всем темам. Важнее смотреть на когерентности тех тем, которые были признаны качественными человеческими экспертами.

Таблица 4 суммирует наиболее важные для понимания результаты, таких как интерпретируемость (вопрос 2), и их связь с tf-idf когерентностью. В этой таблице

частично интерпретируемые — это те темы, которые получили «1» хотя бы у одного эксперта; хорошо интерпретируемые темы — те, которым хоть один эксперт поставил «2». Лидерами являются модели 5 и 6 (restricted dictionary и extended dictionary, соотвественно). В таблице можно видеть, что модель 6 превосходит все остальные по общему качеству. Модель 5 же даёт более высокие показатели когерентностей в в группах интерпретируемых тем, но надо учитывать, что число найденных ею интерпретируемых тем меньше. Это значит, что модель 5 находить меньше тем, но эти темы более высокого качества.

Таблица 5 суммирует наиболее важные результаты, определяющие степень релевантности тем нашим целям. Под релевантностью в таблице подразумевается соответствие темы межэтничным или международным отношениям. Средняя интерпретируемость рассчитывалась как среднее арифметическое оценок, поставленных соответствующим темам экспертами при ответе на вопрос 2. Здесь опять видны два лидера — модели 5 и 6. При этом первая превосходит вторую в терминах tf-idf когерентности релевантных тем, а вторая превосходит первую в терминах числа тем, которые эксперты сочли релевантными. Это верно и для межэтничных, и для международных отношений, а также для обоих уровней релевантности. Это означает, что расширения словаря приводит к появлению большего числа полезных тем меньшим качеством.

Таблица 6 показывает экспертные оценки интерпретируемости тем: она показывает среднее значение оценок, выставленных темам из каждого подмножества для двух основных вопросов, т.е. верхний левый угол показывает, что в среднем эксперты выставили оценку 1.80 в вопросе 1 (общая интерпретируемость) темам, высоко релевантным тематике этничности. Стоит отметить, что теперь модель 6 превосходит модель 5 в терминах интерпретируемости: по этой метрике, в модели 6 не только больше релевантных тем, но они ещё и более интерпретируемые, чем в модели 5. Однако лишь часть из них связано со специфичными событиями (вопрос 2). Тем не менее, с социологической точки зрения, модель 6 выглядит более предпочтительной на этой стадии исследования.

В то же время, словарь модели 6 более широк: он подменяет отсутствующие этнонимы соответствующими прилагательными и названиями стран. Такой принцип построения словаря позволяет выявить в модели этничности, которые не были названы напрямую, и такой подход предотвращает переобучение в лучшей модели. Поэтому в будущем планируется использовать именно такой словарь из соображений практичности и надёжности.

Интересные результаты показаны моделями 7 (recursive) и 8 (keyword texts). По

параметрами числа релевантных тем и когернтности модель 7 похожа на модель 5; модель 8 же больше похожа на модель 6. Это означает, что ре-итерирование построения тематической модели на подмножествах текстов не даёт преимущества, или даже приводит к ухудшению качества. В то же время, однократное обучение по подмножеству текстов, содержащих хотя бы один этноним из Q приводит к модели, аналогичной по качеству (или немного худшей) лучшей модели 6. Таким образом можно сделать вывод, что обучение по фильтрованной коллекции может оказаться полезным, особенно в случае обработки больших коллекций.

5.3 Эксперименты на коллекции IQBuzz

Описываемые в этом разделе эксперименты с коллекцией сообщений различных социальных медиа IQBuzz направлены на дальнейшее совершенствование и упрощение тематического моделирования для решения поставленной задачи. В рамках этих экспериментов также рассматривается построение моделей, учитывающих метаданные о метках времени и геотегах, привязанные к сообщениям.

Данные и параметры Коллекция была предоставлена в лемматизированном виде и содержит почти 5.9M текстов. Источников текстов было около 6000, основные: ВКонтакте — 4766761, Twitter — 394060, Google+ — 175213, LiveInternet — 107211, ursa-tm.ru — 58294, Эхо Москвы — 20729. Все прочие источники (более 6000, из каждого меньше 20000 сообщений) были объединены в общий фоновый. Кроме того, все такие сообщения получили общий геотег из-за сложности извлечения геотега из них, либо из-за отсутствия такого геотега. Из этих данных были извлечены этнонимы, которые и составили словарь для последующего обучения, объём этого словаря составил 588 этнонимов (и постсоветстких, и международных). Исходный объём словаря коллекции составил примерно 8.3M слов. Далее были произведены следующие преобразования коллекции:

- все посты с некириллическими геотегами получили общий геотег;
- из меток времени была извлечена только дата в унифицированном формате;
- геотеги были сопоставлены с заранее подготовленным словарём для унификации (объединения в один геотег различных написаний одной и той метки местоположения).

После была проведена фильтрация словаря коллекции, в ходе которой были удалены слова

- встречающиеся в коллекции меньше 150 раз;
- встречающиеся в коллекции чаще 1 млн раз;
- с длиной меньше 4 символов;
- с длиной более 30 символов;
- содержащие что-либо, кроме кириллицы.

В итоге объём словаря составил около 75K слов, число меток времени — 715, число геотегов — 98. Помимо этого словарь был дополнен примерно 10K биграммами, содержащими этнонимы, которые были получены путём выделения всех подобных биграмм из коллекции с последующей частотной фильтрацией. Средняя длина документа после фильтрации коллекции составила 262 слова.

При работе с данной коллекцией для простоты обучения было принято решения использовать оффлайновый алгоритм. Количество тем было выбрано в результате асессорской работы: было выявлено, что начиная с 200 тем модели без регуляризации перестают выявлять новые этнические тем, поэтому именно это количество тем было зафиксировано во всех дальнейших экспериментах с коллекцией IQBuzz.

Для определения оптимального числа итераций по коллекции был произведён следующий эксперимент. Аналогично всем предыдущим экспериментам (по подбору числа тем), строилась модель PLSA с 20 итерациями, поскольку именно при таком количестве проходов по коллекции сходилась метрика перплексии. Для проверки гипотезы об избыточности такого количества проходов, были проанализированы наиболее вероятные слова в темах, получаемые после каждой итерации, начиная с 10 по 20. Было выявлено, что никаких существенных изменений в составе этих слов не происходит уже начиная с 12 итерации, поэтому количество проходов было зафиксировано равным 12. Для 200 тем это привело к тому, что итоговое время обучения модели с распараллеливанием на 10 потоков составило примерно 6700 секунд. Машина core i7, 6 ядер с hyper-threading.

Модели По аналогии с экспериментами, проводившимися ранее, в тематической модели коллекции IQBuzz был опробован ряд описанных инструментов АРТМ. В различных комбинациях и с разными коэффициентами опробованы регуляризаторы для частичного обучения по словарю этнонимов, регуляризаторы сглаживания и разреживания тем, их декорреляции, регуляризация с использованием модальностей этнонимов и биграмм с этнонимами. Оценивание производилось асессорами по

упрощённой схеме: каждая тема оценивалась как этническая или неэтническая по своим 20 наиболее вероятным словам. Выявлено, что наиболее значительный вклад в усовершенствование модель данной коллекции вносят регуляризация с использованием этнонимов и биграмм. Связано это во многом с особенностями коллекции: она относительно велика, содержит большое количество длинных документов и нет хорошо поддаётся более тонким методам регуляризации. В результате перебора по сетке значений оба выбранных регуляризатора получили коэффициенты τ равные 10. По результатам оценивания в наилучшей модели выявлено 87 тем, в той или иной степени относящихся к этническим вопросам, либо касающихся внешней политики. В PLSA таких тем получено 47, и качество их (с точки зрения интерпретируемости) существенно ниже. В итоге эта модель с регуляризаторами модальностей и биграмм была выбрана для дальнейших экспериментов с внедрением модальностей геотегов и меток времени.

Коэффициенты регуляризации для данных модальностей были выбраны равными 1 по результатам экспериментов. Причина в том, что величины этих коэффициентов оказывают сглаживающее или, наоборот, разреживающее влияние на соответствующие множества слов. Так, стремление коэффициентов к нулю приводит к получению почти равномерных распределений, что явно не соответствует поставленной задаче. В то же время, значения, большие или равные 2, приводят к распределениям, в которых вероятностная масса сосредоточена в 1-2 значениях, что также является искажением. При значении 1 распределения как геотегов, так и меток времени имеют множество ненулевых вероятностей, но, в то же время, в большинстве тем образуют несколько выраженных пиков, позволяющих оценивать принадлежность подобных тем к различным регионам РФ в разрезе определённых временных интервалов. По этой же причине принято решение отказаться от дополнительных регуляризаторов сглаживания/разреживания модальностей геотегов и меток времени, поскольку сами модальности оказались достаточно сильным регуляризатором, справляющимся с этой задачей.

Результаты Поскольку для данной задачи отсутствуют подходящие автоматические метрики оценивания, было проведено ручное кодирование, с целью поиска тем, для которых найденные геотеги и метки времени могли быть интерпретированы и соотнесены с событиями в реальности. Из 87 отобранных ранее тем были выделены около 20, для которых подобные соответствия удалось установить без особых усилий. В таблицах 7-9 наиболее вероятные слова, геотеги и метки времени для некоторых из них (опущены темы-дубликаты).

	Результаты	Комментарии:
Слова	чеченский, чечня, кадыров, боевик, террорист, убийство, рамзан, грозный, спецназ, наемник, кавказ, погибать, операция, теракт, вооруженный, боевой, заложник, дудаев, лидер, командир	Обсуждение чеченской войны в годовщину её начала.
Геотеги	Москва, Санкт-Петербург, Чечня.	
Метки времени	Начале и конец декабря 2014.	
Слова	украина, олигарх, украинский, хунта, киевский, восточный, режим, юговосток, поддерживать, янукович, переворот, евромайдан, восток, революция, регион, одесса, поддержка, независимость, правый, евросоюз	Обсуждение только что произошедшего государственного переворота на Украине.
Геотеги	Москва, Общий геотег.	
Метки времени	Март и апрель 2014.	
Слова	независимость, кричать, вчера, снова, предок, русский, кацап, олигарх, вспомнить, завидовать, разве, ценность, гордиться, независимый, украина, вера, политик, громко, москаль, отказываться	
Геотеги	Москва, Санкт-Петербург.	
Метки времени	Март и август 2014.	
Слова	армянин, армянский, армения, азербайджан, геноцид, азербайджанский, баку, азербайджанец, карабах, ереван, турция, архив, апрель, кавказ, жертва, грузия, память, убийство, русский, представитель	24 апреля было признано датой памяти жертв геноцида армян в Османской Империи. В теме смешались несколько смежных подтем.
Геотеги	Москва, Краснодарский край, Санкт-Петербург, Ростовская область.	
Метки времени	Двадцатые числа апреля 2015.	

Таблица 7: Результаты интерпретации модели с геотегами и метками времени.

	Результаты	Комментарии:
Слова	переселенец, северянин, мигрант, экстремист, иммигрант, сотрудник, документ, нелегальный, волонтер, сообщать, мятежник, пункт, данные, миграционный, порядок, трудовой, север, приезжать, преступление, флаг	Один из всплесков обсуждения темы прослеживается в июле 2014 в связи с прибытием в Мурманскую область мигрантов из Украины.
Геотеги	Мурманская область, Москва, Воронежская область, Санкт-Петербург	
Метки времени	Тема размазана во времени, есть несколько всплесков.	
Слова	грузинский, грузия, русич, буква, шаман, тува, тбилиси, корень, форма, либо, нибудь, саакашвили, носок, праздничный, южный, гореть, огонь, тувинский, собирать, чулок	Обсуждение войны в Южной Осетии в её годовщину. К теме примешалась неясная подтема, связанная с Тывой.
Геотеги	Москва, Тыва.	
Метки времени	Крупный всплеск 8 августа 2015.	
Слова	украинский, украина, ополченец, донецк, славянск, сбивать, юговосток, дебальцево, самолет, донецкий, хунта, вооружение, новороссия, боинг, боевой, тяжелый, луганск, техника, точка, котел	Обсуждение сбитого над территорией Украины малазийского «Боинга».
Геотеги	Москва, Общий геотег, Ростовская область.	
Метки времени	17-18 июля 2014.	
Слова	татарин, русский, крымский, татарстан, крымскотатарский, депортация, казань, мусульманин, татарский, проживать, родной, меджлис, казанский, родина, мечеть, этнический, тюркский, коренной, ислам, мусульманский	
Геотеги	Татарстан, Пермский край, Москва, Санкт-Петербург.	
Метки времени	Тема размазана во времени.	

Таблица 8: Результаты интерпретации модели с геотегами и метками времени (продолжение).

	Результаты	Комментарии:
Слова	еврей, еврейский, израиль, холокост, гитлер, израильский, фамилия, еврейка, жертва, убийство, комиссар, сионист, уничтожение, газета, раввин, лагерь, начальник, антисемитизм, синагога, палестина	Обсуждение геноцида евреев в фашистской Германии в день памяти жертв Холокоста.
Геотеги	Москва, Санкт-Петербург, Общий геотег.	
Метки времени	27 января 2015.	
Слова	китайский, китаец, сказка, пекин, восток, дальний, восточный, золотой, азия, сибирь, гонконг, желтый, продавать, лиса, аренда, поднебесный, товар, остров, проект, балл	Обсуждение сообщений о сделках Китая и РФ на ПМЭФ 2015 на крупные суммы. ПМЭФ проходил в это время.
Геотеги	Москва, Санкт-Петербург, Пермский край, Свердловская область.	
Метки времени	Июнь и июль 2015.	
Слова	аллах, мусульманин, сирия, сирийский, ислам, пророк, иран, исламский, иранский, саудовский, арабский, наносить, коран, аравия, кожа, мухаммад, мечеть, посланник, имам, мусульманский	Обсуждение сирийского конфликта в момент начала активного вмешательства РФ в него.
Геотеги	Москва, Дагестан, Санкт-Петербург, Чечня, Татарстан.	
Метки времени	Октябрь 2015.	

Таблица 9: Результаты интерпретации модели с геотегами и метками времени (продолжение).

Можно обратить внимание на то, что почти все темы содержат в числе наиболее вероятных геолокаций Москву и Санкт-Петербург, что вполне закономерно, поскольку суммарное число сообщений из этих городов составляет почти 40% от общего числа всех текстов. Под общим геотегом понимается совокупность всех геотегов, которые не были отнесены ни к одному из выделенных геотегов, так же к этой геолокации были отнесены сообщения, у которых геотег изначально отсутствовал.

Таким образом, удалось без дополнительных усилий использовать информацию о метках геолокации и времени в процессе обучения модели, а также на основании полученных результатов выявить изменения полученных тематик в пространствах этих меток.

6 Результаты, выносимые на защиту

На защиту в данной работе выносятся следующие результаты:

1. Новый детерминированный онлайновый асинхронный EM-алгоритм, позволяющий производить эффективное обучение тематических моделей M-APTM.

2. Использование методологии регуляризации для автоматического выявления специфических тематик, обсуждаемых в текстовых данных, с возможностью учёта дополнительной информации и оценивания на её основе.

Список литературы

[1] Andrzejewski, D., Zhu, X.: Latent Dirichlet allocation with topic-in-set knowledge. In: Proc. NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing. pp. 43–48. SemiSupLearn '09, Association for Computational Linguistics, Stroudsburg, PA, USA (2009)

[2] Andrzejewski, D., Zhu, X., Craven, M.: Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In: Proc. 26th Annual International Conference on Machine Learning. pp. 25–32. ICML '09, ACM, New York, NY, USA (2009)

[3] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research 3(4–5), 993–1022 (2003)

[4] Bodrunova, S., Koltsov, S., Koltsova, O., Nikolenko, S.I., Shimorina, A.: Interval semi-supervised lda: Classifying needles in a haystack. In: Proc. MICAI 2013, LNCS vol. 8625, pp. 265–274. Springer (2013)

[5] Chemudugunta, C., Smyth, P., Steyvers, M.: Modeling general and specific aspects of documents with a probabilistic topic model. In: Advances in Neural Information Processing Systems. vol. 19, pp. 241–248. MIT Press (2007)

[6] Griffiths, T., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences 101 (Suppl. 1), 5228–5335 (2004)

[7] A. Smola and S. Narayanamurthy: An architecture for parallel topic models. Proc. VLDB Endow., 3(1-2):703–710, Sept. (2010)

[8] M. D. Hoffman, D. M. Blei, and F. R. Bach.: Online learning for latent dirichlet allocation. In NIPS, pages 856–864. Curran Associates, Inc. (2010)

[9] Hoffmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Machine Learning 42(1), 177–196 (2001)

[10] O. Frei and M. Apishev: Parallel Non-blocking Deterministic Algorithm for Online Topic Modeling. Analysis of Images, Social Networks and Texts. AIST. Communications in Computer and Information Science, vol 661. Springer, pp. 132-144. (2016)

[11] Jagarlamudi, J., Daumé, III, H., Udupa, R.: Incorporating lexical priors into topic models. In: Proc. EACL'12, pp. 204–213 (2012)

[12] Koltcov, S., Koltsova, O., Nikolenko, S.I.: Latent dirichlet allocation: Stability and applications to studies of user-generated content. In: Proc. WebSci 2014, pp. 161–165 (2014)

[13] Apishev M., Koltcov S., Koltsova O., Nikolenko S., Vorontsov K.: Mining Ethnic Content Online with Additively Regularized Topic Models. In: Computacion y Sistemas, Vol. 20, No. 3, pp. 387—403. (2016)

[14] Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proc. EMNLP'11, pp. 262–272 (2011)

[15] Nikolenko, S.I., Koltsova, O., Koltsov, S.: Topic modelling for qualitative studies. Journal of Information Science (2015)

[16] Paul, M.J., Dredze, M.: Discovering health topics in social media using topic models. PLoS ONE 9(8) (2014)

[17] Sociopolitical processes in the internet. Laboratory for Internet Studies. Internal report, National Research University Higher School of Economics, reg. no. 01201362573, Moscow (2013)

[18] R. Rehurek and P. Sojka.: Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, Valletta, Malta (2010)

[19] J. Langford, L. Li, and A. Strehl. Vowpal wabbit open source project. *Technical report*, Yahoo! (2007)

[20] Tan, Y., Ou, Z.: Topic-weak-correlated latent dirichlet allocation. In: 7th International Symposium Chinese Spoken Language Processing (ISCSLP). pp. 224–228 (2010)

[21] Tikhonov, A.N., Arsenin, V.Y.: Solution of ill-posed problems. W. H. Winston, Washington, DC (1977)

[22] Vorontsov, K.V., Potapenko, A.A.: Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. In: AIST'2014, Springer CCIS vol. 436, pp. 29–46 (2014)

[23] Vorontsov, K.V., Potapenko, A.A.: Additive regularization of topic models. Machine Learning, Special Issue on Data Analysis and Intelligent Optimization with Applications 101(1), 303–323 (2015)

[24] Vorontsov, K., Frei, O., Apishev, M., Romov, P., Suvorova, M., Yanina, A.: Non-bayesian additive regularization for multimodal topic modeling of large collections. In: Proc. of TM '15, pp. 29–37, ACM, New York, NY, USA (2015)

[25] Vorontsov, K.: Additive regularization for topic models of text collections. Doklady Mathematics 89(3), 301–304 (2014)