Правительство Российской Федерации

Федеральное государственное автономное образовательное

учреждение высшего образования

«Национальный исследовательский университет

«Высшая школа экономики»

Факультет компьютерных наук

Основная образовательная программа

Прикладная математика и информатика

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

на тему

Темпоральное моделирование новостных потоков

Выполнил студент группы 141, 4 курса,

Фадеева Анастасия Вадимовна

Научный руководитель:

доктор физ-мат наук, профессор,

Воронцов Константин Вячеславович

Москва 2018

Содержание

1 Ключевые слова

Topic Detection and Tracking, Topic Models, Probabilistic Latent Semantic Analysis, Additive Regularization of Topic Models

2 Введение

В данный момент обработка новостного потока является очень важной задачей, поскольку с увеличением информации, производимой медиа, возрастает необходимость в автоматической агрегации новостей. Особенностью работы с новостными источниками является то, что критически важно при агрегации учитывать время создания новости. Область машинного обучения, которая занимается этой проблемой, называется детекция и отслеживание тем (в иностранной литературе Topic Detection and Tracking).
В данной работе решается практическая задача мониторинга сообщений СМИ о компаниях. Этот инструмент может быть полезен как для самой компании, так и для ее партнеров, банков, рейтинговых агенств и тд. Так, например, публикации в СМИ о жалобах пользователей скорее всего отрицательно скажутся на привлекательности услуг компании.
В области обнаружения и отслеживания тем применяются тематические модели, которые по коллекции документов строят распределения слов в темах и распределения тем в документах. Особенностью онлайн моделей является то, что они строятся последовательно по периодам, то есть на каждой итерации на вход получают лишь часть данных. Целью моего исследования было построить тематическую онлайн модель, которая работает с качеством, сопоставимым с тематической моделью на всех данных. Задача детекции заключается в том, чтобы обнаруживать темы в новостном потоке по мере поступления новых документов. Такое обнаружение возможно лишь с некоторым запаздыванием. Поэтому для оценивания качества онлайн модели ставится задача сравнения тем со стандартной моделью, которая строится ретроспективно по историческим данным. Второй целью моего исследования была разработка методики оценивания отслеживания тем в темпоральных тематических моделях. Качество отслеживания тем является очень важной характеристикой темпоральной тематической модели, потому что период обсуждения события может быть значительно дольше одного периода времени, который обрабатывается в модели. Например, судебные разбирательства освещаются на протяжении всего процесса. В задачи работы входило изучение существующих тематических моделей в области выявления и отслеживания тем, предложение улучшенной модели и оценка ее качества.

Объектом исследования являются темпоральные тематические модели, а предметом – применение онлайн тематических моделей для задачи обнаружения и отслеживания тем. В исследовании применялись следующие методы исследования: моделирование новостного потока и проведение экспериментов на реальных данных.
В данной работе реализована модель с динамически увеличивающимся словарем, автоматическим детектированием новых тем и отслеживанием уже существующих. Основные алгоритмы в этой области (Topic Detection and Tracking) исходят из предпосылки фиксированного словаря и не обрабатывают появления новых слов, однако, эта информация может быть очень важной для обнаружения новых событий. Например, фамилия нового генерального директора компании является хорошим признаком для отнесения новости к новому событию.
Работа организована следующим образом: сначала будут описаны работы в области обнаружения и отслеживания тем и связанные с ними исследования, далее будут представлены методы исследования с подробным описанием предложенной модели, а также метрика для оценки качества тематических моделей. В разделе с экспериментами будет приведено сравнение с офлайн тематической моделью на всем периоде и оценка качества решения задачи по отслеживанию тем во времени.

3 Обзор литературы

Для задачи отслеживания и детектирования тем часто применяется тематическое моделирование. Тематическое моделирование – это метод мягкой кластеризации (документ может относиться к нескольким классам), в котором по коллекции документов строятся распределения слов в темах и распределения тем в документах. Наиболее популярные модели тематического моделирования – это Latent Dirichlet Allocation [4], Probabilistic latent semantic analysis [10] и Non-negative Matrix Factorization [6]. Далее будут описаны две модели для решения задачи обнаружения и отслеживания тем на основе тематических моделей: Dynamic Latent Dirichlet Allocation и Online Latent Dirichlet Allocation.
Первый подход был предложен в статье [3]. В нем заранее фиксируется количество тем, и эти темы развиваются во времени. Сначала рассматриваемый временной отрезок делится на периоды (в статье на недели), на каждом отрезке строится LDA модель, и темы за соседние периоды связываются лог-нормальным распределением, благодаря этому темы не сильно изменяются во времени.
В статье [1] был предложен другой подход для моделирования новостного потока. В Online-LDA на первом периоде строится обычная LDA

модель, а потом полученные распределения слов в темах используются как априорные распределения для тем на следующем периоде. Также на каждом периоде к модели могут быть добавлены новые темы, которые отвечают за новые события для этого периода. Для таких тем априорное распределение не вводится.
Динамическая модель лучше подходит для ретроспективного поиска событий и анализа трендов, потому что она обрабатывает сразу всю коллекцию. Онлайн модель нацелена на выявление новых событий, а также прослеживает темы во времени и обрабатывает новости итерационно по периодам.
В статье [17] был предложен метод для динамического моделирования, в котором тематическая модель строится на каждом периоде отдельно, а потом темы сопоставляются по похожести наиболее вероятных слов с помощью венгерского алгоритма [12].
В статье [7] вводится модель с распределением времени в теме, которое строится на основе меток времени документов. В этой модели применяется l1-регуляризация, чтобы вероятности соседних периодов времени не сильно различались. Этот подход также используется для ретроспективного анализа событий.
Исследования в области ретроспективного поиска событий нацелены на выявление значимых событий в теме. Для этого на данных запускаются темпоральные тематические модели и ищутся значительно изменившиеся темы. Предполагается, что значительное изменение темы было вызвано каким-то событием. В статье [5] вероятностные распределения тем были выделены с помощью динамической модели, а изменение темы измерялось с помощью косинусной близости между распределениями тем за соседние периоды времени. В других статьях, например [11], консистентность тем измерялась с помощью дивергенции Кульбака-Лейблера, похожести Хеллингера и других.
Мотивацией для исследования было улучшение качества онлайн модели и разработка методов оценивания качества отслеживания тем. В следующей главе более подробно будет описан алгоритм тематического моделирования PLSA и его модификация для решения задачи отслеживания и детекции тем.

4 Методы

В данной секции будут введены основные понятия, стандартная модель тематического моделирования PLSA, а также ее модификации с использованием регуляризации. Потом будет описана модель Онлайн PLSA и регуляризаторы, которые в ней используются. В последнем разделе бу-

дет описан стандартный метод оценки качества тематической модели.

4.1 Тематическое моделирование

Тематическое моделирование — это алгоритм мягкой кластеризации, который группирует неразмеченные документы в кластеры на основе их похожести между собой. Тема в тематической модели — это вероятностное распределение над словами. Для каждого текста в тематической модели вычисляется вероятность принадлежности к каждой теме.

4.2 PLSA

Probabilistic latent semantic analysis — это вероятностная модель, которая принимает на вход коллекцию текстов в формате матрицы: n_{dw} сколько раз термин w встречается в документе d. В модели вводятся темы t как распределения над словами. Основное предположение модели в условной независимости слова от конкретного документа: $p(w|d,t) = p(w|t)$. Из этого следует выражение:

$$p(w|d) = \sum_t p(w|t,d)p(t|d) = \sum_t p(w|t)p(t|d). \tag{4.2.1}$$

Введем следующую параметризацию: $\phi_{wt} = p(w|t)$ – вероятность слова w в теме t и $\theta_{td} = p(t|d)$ – вероятность темы t в документе d. В таком случае правдоподобие выборки D можно записать следующим образом:

$$\begin{aligned} p(D) = & \\ &= \prod_{d\in D}\prod_{w\in d} p(w,d)^{n_{wd}} \\ &= \prod_{d\in D}\prod_{w\in d} p(w|d)^{n_{wd}} p(d)^{n_{wd}} \\ &= \prod_{d\in D}\prod_{w\in d} \phi_{wt}^{n_{wd}}\ \theta_{td}^{n_{wd}}\ p(d)^{n_{wd}}. \end{aligned} \tag{4.2.2}$$

Взяв логарифм и избавившись от константы относительно параметров модели, по принципу максимального правдоподобия получим следующую оптимизационную задачу с ограничениями:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} \to \max_{\Phi,\Theta} \tag{4.2.3}$$

$$\phi_{wt} \geq 0 \quad \sum_w \phi_{wt} = 1 \quad \theta_{td} \geq 0 \quad \sum_t \theta_{td} = 1$$

Ограничения появляются из-за того, что ϕ_t и θ_d должны быть вероятностными распределениями.

4.3 Аддитивная регуляризация PLSA

Задача матричного разложения, которая решается в модели PLSA, имеет бесконечно много решений, и для корректной работы алгоритма применяют регуляризацию:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + \sum_{i=1}^{k} \tau_i R_i(\Phi, \Theta) \to \max_{\Phi,\Theta}, \tag{4.2.4}$$

где $R_i(\Phi, \Theta)$ – это регуляризатор с неотрицательным коэффициентом τ_i. Впервые такой подход был предложен в [19]. Данный функционал оптимизируется с помощью EM-алгорима, подробнее описанного в [8]. Далее будут приведены два регуляризатора из статьи [18], которые применяются для улучшения интерпретируемости тем в модели. Про другие регуляризаторы, которые применяются в тематическом моделировании, подробно написано в статье [7].

4.3.1 Регуляризатор сглаживания/разреживания

Введем регуляризатор сглаживания/разреживания распределений из матриц Φ и Θ следующим образом:

$$R(\Phi, \Theta) = \sum_{t\in T} \sum_{w\in W} \beta_{wt} \ln \phi_{wt} + \sum_{d\in D} \sum_{t\in T} \alpha_{td} \ln \theta_{td} \tag{4.2.5}$$

Если параметры β_{wt} и α_{td} отрицательные, то этот регуляризатор будет обладать разреживающим эффектом, то есть обнулять маленькие вероятности. При положительных значениях параметров он действует наоборот и сглаживает распределения. Этот регуляризатор реализован в пакете BigARTM [2] в модулях Smooth Sparse Phi Regularizer и Smooth Sparse Theta Regularizer.

4.3.2 Декоррелирование тем в модели

Для того чтобы темы в матрице Φ были как можно более различны, вводится регуляризатор декоррелирования, который штрафует за похожие темы в модели. Похожесть определяется как скалярное произведение распределений тем над словами. Этот регуляризатор представляет

собой сумму всех возможных попарных скалярных произведений распределений тем:

$$R(\Phi) = -\sum_{t \in T} \sum_{s \in T/t} \sum_{w \in W} \phi_{wt}\phi_{ws}. \tag{4.2.6}$$

Этот регуляризатор реализован в пакете BigARTM [2] в модуле Decorrelator Phi Regularizer.

4.4 Онлайн PLSA для темпорального моделирования

В данном разделе будет представлена модель на основе PLSA, которая решает задачу обнаружения и отслеживания тем во времени. Она называется Онлайн PLSA, так как работает по принципу онлайн алгоритма, то есть в ней, так же как и в модели Онлайн LDA, новые документы добавляются итерационно по периодам. Предложенная модель избавляется от предположения о фиксированном словаре. Чтобы это сделать, априорные распределения вводятся не для тем, как сделано в модели Онлайн LDA, а для уже обработанных моделью документов, и они равны распределениям с предыдущего шага. Для новых документов вводится априорное распределение для улучшения детекции. Также темы декоррелируются между собой, и распределения в матрицах Φ и Θ разреживаются с помощью инструментов, описанных в разделе 4.3. Далее более подробно будут описаны регуляризаторы, предложенные для этой модели.

4.4.1 Регуляризация для старых документов

Для документов, которые уже были обработаны моделью, и их разложения по темам известны, вводятся следующие априорные распределения:

$$\theta_k^d \sim Dir(\cdot|\theta_{k-1}^d). \tag{4.2.6}$$

Таким образом, в качестве априорного распределения для уже обработанного документа берется его распределение по темам с предыдущего шага, что и записано выше. Это сделано для того, чтобы распределение документа по темам не сильно изменялось со временем. В виде регуляризатора модели это записывается следующим образом:

$$R_{old}(\Theta) = \ln \prod_{d \in D_{old}} Dir(\theta_k^d, \theta_{k-1}^d) = \sum_{d \in D_{old}} \sum_{t \in T} (\theta_{k-1}^{dt} - 1) \ln \theta_k^{dt} \tag{4.2.7}$$

Вероятности отнесения старого документа к новым темам строго равны нулю. Результаты работы этого регуляризатора представлены в разделе 5.

4.4.2 Регуляризация для новых документов

Для новых документов вводится единое априорное распределение, которое поощряет отнесение новых документов к новым темам. Это распределение пропорционально времени создания темы:

$$\theta_k^d \sim Dir(\cdot|\beta) \quad \beta_t \propto \text{период времени } t, \tag{4.2.8}$$

где период времени t – это номер периода по порядку добавления. В виде регуляризатора модели это записывается следующим образом:

$$R_{new}(\Theta) = \ln \prod_{d \in D_{new}} Dir(\theta_d, \beta) = \sum_{d \in D_{new}} \sum_{t \in T} (\beta_t - 1) \ln \theta_{dt} \tag{4.2.9}$$

Подобный подход для другой модели описан в статье [14].

4.4.3 Онлайн PLSA алгоритм

Выход: Распределения ϕ_t и θ_d
для k – период времени **сделать**
- Добавляются новые темы
- Априорные распределения для существующих документов: $\theta_k^d \sim Dir(\cdot|\theta_{k-1}^d)$
- Априорное распределение для новых документов: $\theta_k^d \sim Dir(\cdot|\beta) \quad \beta_t \propto$ период времени t

конец

Алгоритм 1. Онлайн PLSA

Алгоритм Онлайн PLSA последовательно обрабатывает периоды времени: на каждом шаге добавляются новые темы и применяются два типа регуляризации описанные ваше. Алгоритм PLSA представляет собой разложение матрицы частот слов в документах на матрицы Φ и Θ. На рисунке представлена визуализация одного шага онлайн алгоритма:

- в матрице Φ добавились новые темы T' и новые слова W'

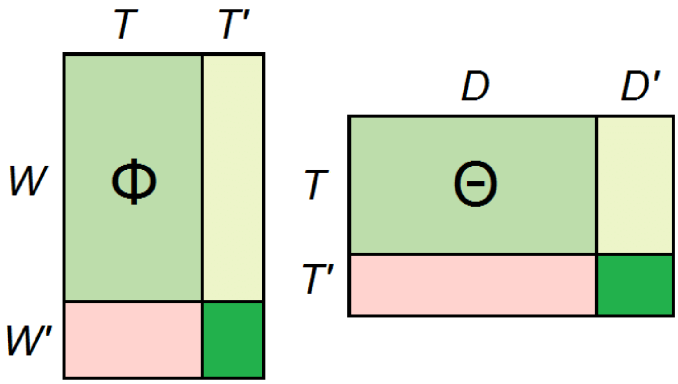

Рис. 1: Шаг онлайн алгоритма

- в матрице Θ добавились новые темы T' и новые документы D'

На рисунке 2 приведена визуализация регуляризации Онлайн PLSA модели. В модели регуляризуется только матрица Θ, а матрица Φ строится по методу максимального правдоподобия. Уже обработанные документы D регуляризуются в соответствии с их распределениями с предыдущего шага и не могут относиться к новым темам T', поэтому там стоят нулевые значения. Для новых документов D' более темный цвет на рисунке 2 обозначает большую априорную вероятность, что соответствует априорному распределению 4.2.8. Таким образом новые документы более вероятно относятся к новым темам.

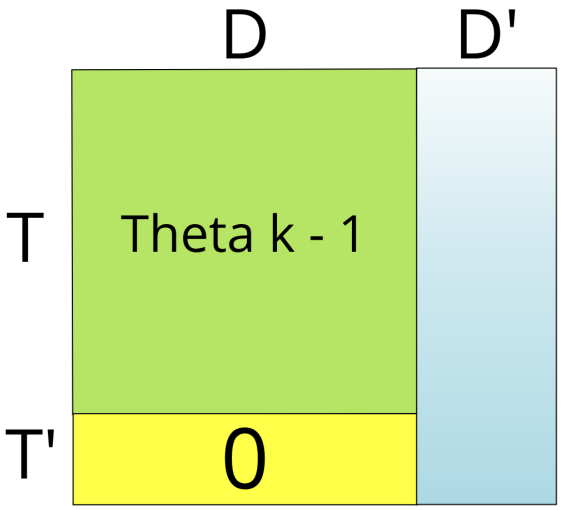

Рис. 2: Матрица Θ

4.5 Оценка качества тематической модели

Для оценки качества тематической модели используется метрика когерентности, которая оценивает интерпретируемость тем. Для каждой темы выбирается m слов с самыми большими вероятностями $p(w|t)$, обозначим это список слов W_{top}. Когерентность темы вычисляется как средняя pointwise mutual information списка W_{top}:

$$\text{Coherence} = \frac{1}{|W_{top}|(|W_{top}|-1)} \sum_{w_j \in W_{top}} \sum_{w_i \in W_{top}/w_j} \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}, \quad (4.2.10)$$

где $p(w)$ – это вероятность встретить слово w в случайном документе, а $p(w_i, w_j)$ – это вероятность встретить оба слова в окне заранее заданного размера. Эта метрика может вычисляться по обрабатываемым данным или по сторонним источникам. Более подробно про меру когерентности написано в статье [13]. В этой статье также было показано, что такая метрика лучше всего коррелирует с ассесорскими оценками интерпретируемости тем.

4.6 Вывод

В этом разделе был введен основной необходимый инструментарий для решения задачи отслеживания и обнаружения тем. Далее предложенная модель будет тестироваться на реальных данных.

5 Результаты

Для оценки качества работы алгоритма Онлайн PLSA были проведены разнообразные эксперименты. Далее будут описаны данные, которые использовались для обучения, и результаты проведенных экспериментов. Представленные ниже эксперименты универсальны (не используют особенности Онлайн PLSA модели) и подходят для оценки качества любых темпоральных тематических моделей.

5.1 Данные

Для тестирования использовались два датасета, их характеристики представлены в таблице ниже.

	Датасет о компании	**RCVI**
Общее количество документов	857	62 581
Среднее количество документов в день	35	1992
Начало временного промежутка	2017-12-01	1996-08-20
Конец временного промежутка	2018-01-04	1996-09-19
Язык	русский	английский
Средняя длина документа в токенах	500	150
Итоговый размер словаря	12 422	395 155

Датасет о компании представляет из себя автоматически собранную коллекцию новостей из разных источников, где есть упоминания определенной компании. Эта коллекция не находится в открытом доступе. В некоторые дни документов по компании приходило очень мало (менее 20) и онлайн модель не обновлялась, эти документы обрабатывались позже.
Второй датасет содержит новости на английском языке от агентства Reuters и открыт для научных исследований. Он является стандартным датасетом для тестирования решения задачи отслеживания и обнаружения тем. В статье [15] подробно описаны характеристики этого датасета. Эта коллекция размечена ассесорами на четыре темы: рынок, правительство (социальная сфера), экономика, промышленность, и документ может относиться к нескольким темам.
Перед применением моделей текстовые данные нормализовались. Для коллекции о компании использовался пакет pymorphy, а для RCVI – spacy. Также из текстов удалялись стоп-слова.

5.2 Отслеживание тем

В данном разделе представлены эксперименты по оценке качества отслеживания тем в темпоральных тематических моделях на примере Онлайн PLSA. Задача отслеживания тем заключается в том, чтобы определять новости, которые относятся к уже существующим темам. Этот механизм нужен для того, чтобы обрабатывать продолжительные по времени события.

5.2.1 Изменения тем

Гипотеза
В этом эксперименте проверяется гипотеза о том, что распределения тем в модели Онлайн PLSA изменяются не значительно. Если темы остаются консистентными во времени (то есть не сильно меняются), то в модели в продолжения тем будут попадать документы по тому же событию, что

является желаемым поведением для темпоральной тематической модели.

Архитектура эксперимента

В каждый временной период k созданные в модели темы имеют вероятностные распределения над словарем. Для того чтобы оценить изменение темы при переходе от периода k к периоду $k+1$, нужно оценить изменение вероятностных распределений ϕ_k^t и ϕ_{k+1}^t. Для этого была использована косинусная мера близости распределений.

Реализация

Была запущена модель Онлайн PLSA на датасетах на русском и на английском языках и на каждой итерации алгоритма по двум матрицам: Φ_k и Φ_{k+1} считалась следующая метрика:

$$\text{Consistence} = \frac{1}{|T_k|} \sum_{t \in T_k} \frac{\phi_k^t \ \phi_{k+1}^t}{||\phi_k^t||_2 \ ||\phi_{k+1}^t||_2},$$

где T_k – это список тем на шаге k. Также на датасете на английском языке была запущена модель предложенная в статье [17]. В этом алгоритме тематическая модель строится для каждого периода независимо, а потом темы сопоставляются по похожести с помощью венгерского алгоритма. В качестве меры похожести была также использована косинусная мера. В этой модели на каждом периоде выделялось одинаковое количество тем (параметры приведены ниже). Визуализация изменения тем во времени приведена на рисунках 3-5.

Параметры онлайн модели (датасет о компании)

- Коэффициент при регуляризаторе старых документов: $\max(\#\text{ documents}, 100)$
- Коэффициент при регуляризаторе новых документов: 10
- Количество добавляемых тем на каждом шаге: $\max(\#\text{ new documents}/10, 5)$
- Эпох до разреживания: 20
- Коэффициент разреживания матрицы Θ: -1
- Эпох после разреживания: 10

Параметры онлайн модели (RCVI датасет)

- Коэффициент при регуляризаторе старых документов: $\max(\#\text{ documents}, 100)$

- Коэффициент при регуляризаторе новых документов: 10
- Количество добавляемых тем на каждом шаге: 30
- Эпох до разреживания: 20
- Коэффициент разреживания матрицы Θ: -6
- Коэффициент разреживания матрицы Φ: -3.5
- Эпох после разреживания: 10

Параметры модели с сопоставлением тем (RCVI датасет)

- Количество тем на каждом шаге: 30
- Эпох до разреживания: 20
- Коэффициент разреживания матрицы Θ: -6
- Коэффициент разреживания матрицы Φ: -3.5
- Эпох после разреживания: 10

Вывод
Опираясь на представленные графики (рис 3-4), можно заключить, что темы почти не меняются во времени, что и требовалось от модели. Для сравнения в модели с сопоставлением мера консистентности тем значительно ниже (рис 5).

5.2.2 Частичное обучение

Гипотеза
В этом эксперименте проверяется гипотеза о том, что модель Онлайн PLSA с частичным обучением дает лучшее качество в задаче отслеживания тем, чем бейзлайн методы обучения с учителем. Такой эксперимент является хорошим способом оценки качества отслеживания, однако, требует размеченных данных.
Архитектура эксперимента
Для проведения такого эксперимента нужна коллекция документов с метками времени и ассесорской разметкой на темы. Выбирается период обработки новостей (периодом может быть день). После чего бейзлайн методы обучаются на размеченном датасете, который соответсвует первому периоду. Для новостей за все последующие периоды с помощью этих моделей делается предсказание вероятности каждой из тем. В тематической модели количество тем равно количеству тем из ассесорской

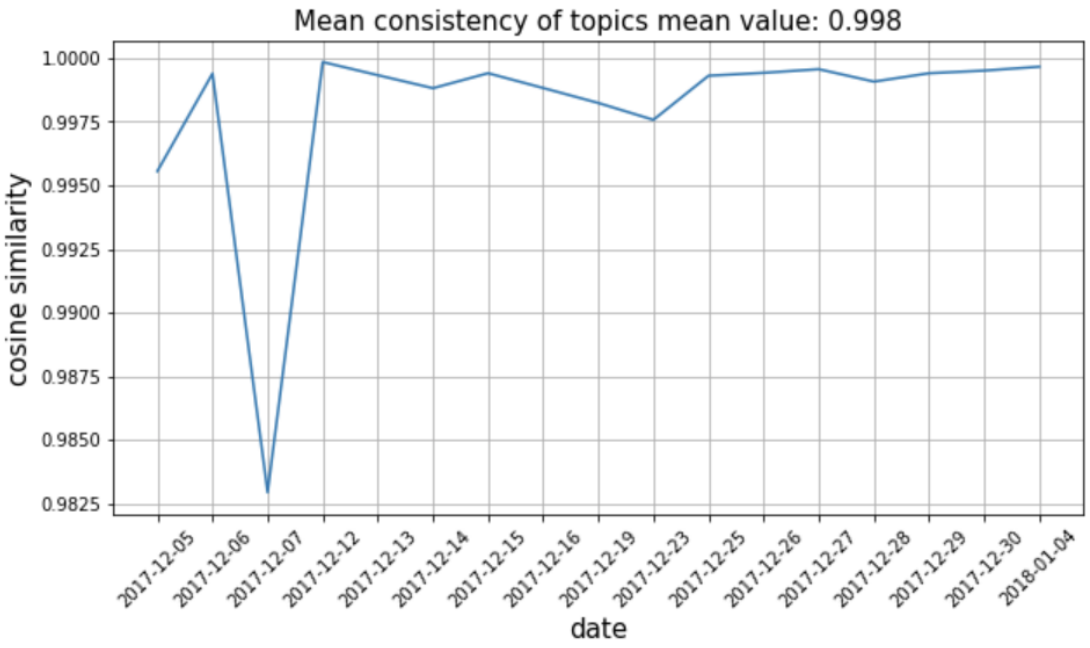

Рис. 3: Изменение тем во времени: новости по компании (выше – лучше)

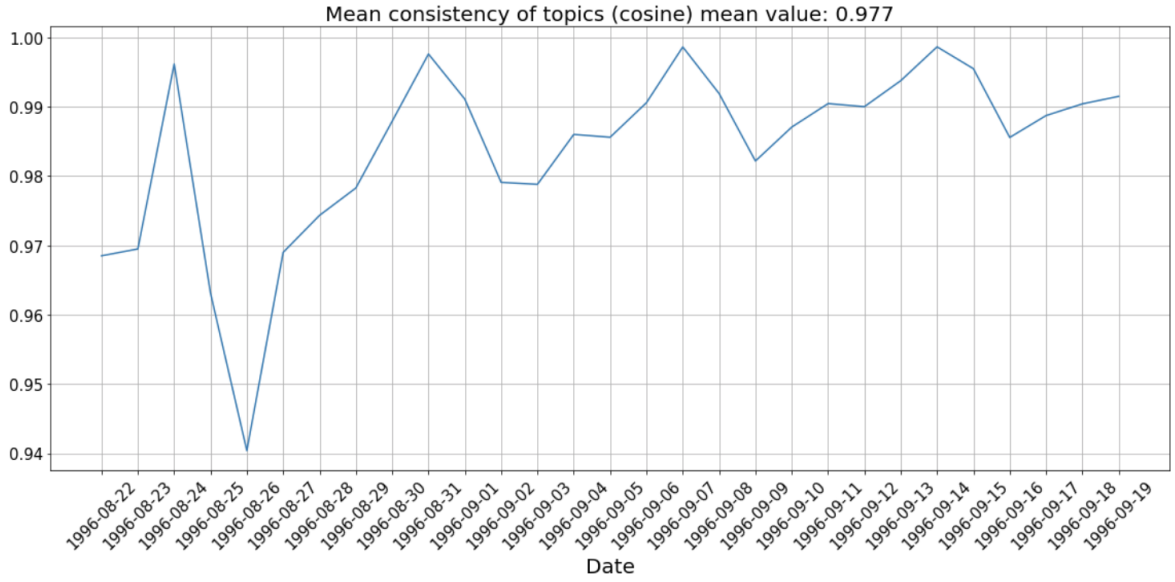

Рис. 4: Изменение тем во времени: RCVI коллекция, Онлайн PLSA модель (выше лучше)

разметки. Для документов из первого периода вводятся априорные распределения, которые соответствуют разметке. Например, если документ принадлежит первым двум темам из трех возможных, то его априорное распределение будет равно $(0.5,\ 0.5,\ 0)$.

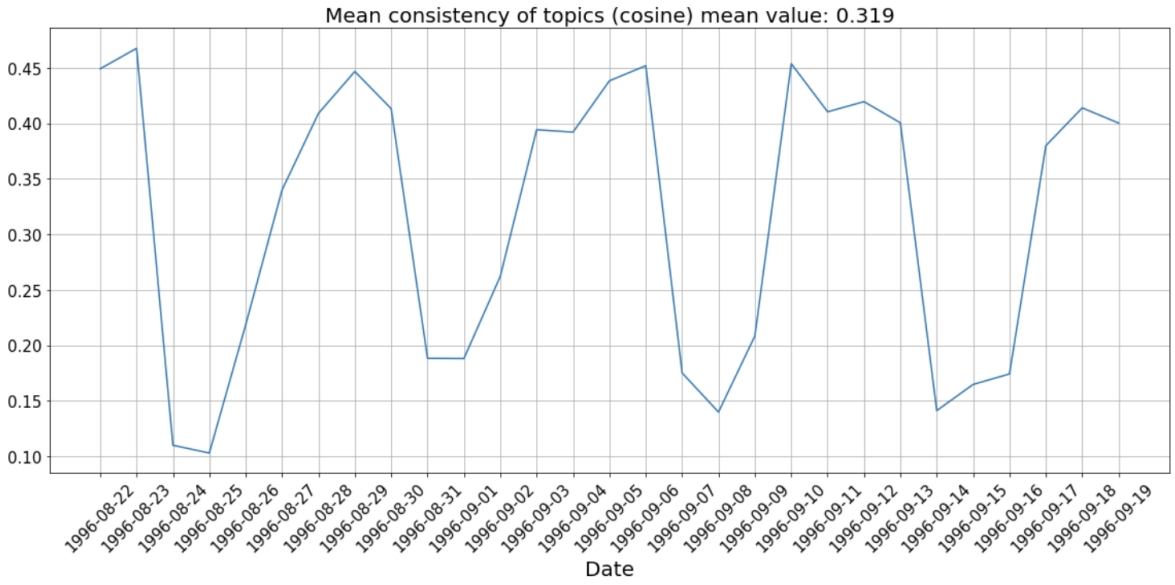

Рис. 5: Изменение тем во времени: RCVI коллекция, модель с сопоставлением тем (выше – лучше)

Реализация

Для тестирования была использована коллекция новостей Reuters. В качестве стандартных методов обучения с учителем были выбраны k ближайших соседей и логистическая регрессия со стандартными параметрами. В модели Онлайн PLSA новые темы не добавлялись, и работал только регуляризатор старых документов. Качество считалось как средняя точность по четырем темам.

Параметры онлайн модели

- Коэффициент при регуляризаторе старых документов – количество документов
- Эпох до разреживания: 20
- Коэффициент разреживания матрицы Φ: -1
- Коэффициент разреживания матрицы Θ: -1.5
- Эпох после разреживания: 10
- Порог отнесения документа к теме: 0.25

Вывод

На рисунке 6 показано среднее качество классификации на всей тестовой выборке (то есть исключая первый день) до определенной даты. Методы

обучения с учителем не переобучались и их качество снижается только за счет увеличения размера тестовой выборки. Модель Онлайн PLSA показывает лучшее качество, чем стандартные модели на коллекции Reuters на всем периоде. Тематическая модель работает лучше, потому что использует не только данные о разметке из тестовой выборки, но и производит кластеризацию тестовых документов

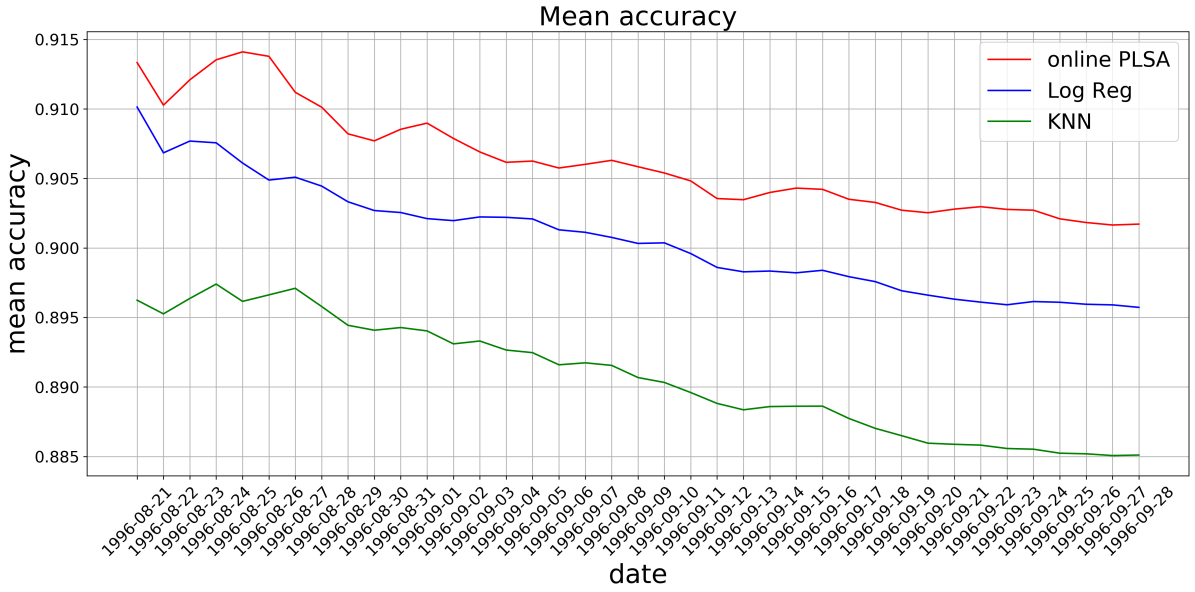

Рис. 6: Средняя точность по четырем темам (выше – лучше)

5.2.3 Полусинтетический датасет

Гипотеза

В случае, когда датасет не размечен на темы и частичное обучение применить нельзя, можно смоделировать ситуацию развития темы. Идея подхода с полусинтетическим датасетом состоит в том, чтобы, изменяя время создания новости, оценивать качество отслеживания тем. В этом эксперименте проверяется гипотеза о том, что модель Онлайн PLSA отслеживает темы с высоким качеством на коллекции русских текстов.

Архитектура эксперимента

В этом эксперименте сначала строится PLSA модель на первом периоде. Потом из каждой темы (если она достаточно большая) случайным образом выбирается небольшая часть документов и их метка времени меняется на следующую. После чего на измененных данных строится темпоральная модель на два периода и считается средний процент попаданий документов в изначальные темы.

Реализация

Онлайн PLSA модель тестировалась на коллекции текстов о компании. Однако, для правильной работы алгоритма потребовалось увеличить период обработки новостей до 90 дней. Это связано с тем, что коллекция небольшая и при изменении меток времени документов модель слишком сильно менялась. На первом периоде запускалась PLSA модель (параметры указаны ниже) и из 10 самых больших по количеству относящихся документов тем случайно сэмплировалось 20% документов, и их метка времени заменялась на следующую. После этого на измененных данных запускались две модели Онлайн PLSA для двух периодов: в одной на втором периоде добавлялись новые темы, а во второй – нет (параметры указаны ниже). Также на изменных данных запускалась модель, в которой на двух периодах независимо строись PLSA модели, а потом темы сопоставлялись по похожести. Этот же подход был описан в эксперименте 5.2.1. Параметры модели совпадают с параметрами офлайн модели, которые приведены ниже. Качество отслеживания измерялось как доля документов попавших в те же темы, из которых они были удалены.
Сэмплирования и построения моделей производилось 10 раз для получения более точной оценки. Визуализация процесса создания полусинтетического датасета представлена на рисунке 7. Для четырех тем выбирается небольшая часть данных (выделена красным) и переносится во второй период. На втором периоде насыщенным красным показана доля документов из перенесенных, которая правильно попала в соответствующую тему.

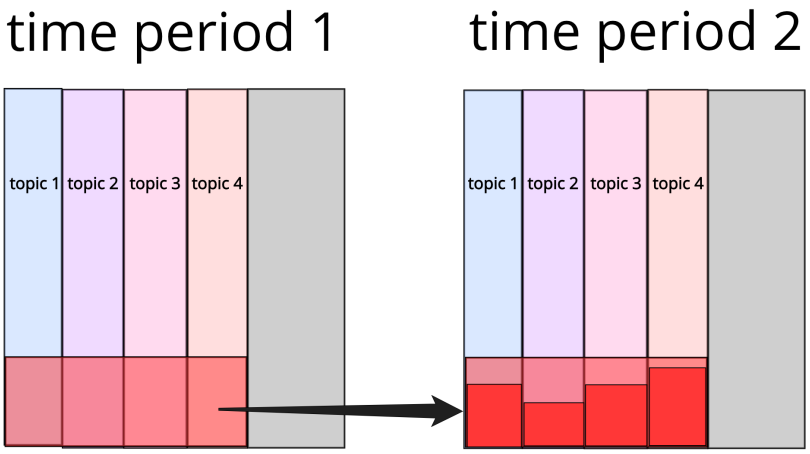

Рис. 7: Построение полусинтетического датасета

Параметры онлайн модели

- Коэффициент при регуляризаторе старых документов – количество документов 10^5

- Количество добавляемых тем: 20
- Эпох до разреживания: 20
- Коэффициент разреживания матрицы Θ: -1
- Эпох после разреживания: 10
- Порог отнесения документа к теме: 0.5

Параметры офлайн модели

- Количество тем: 20
- Коэффициент деккореляции: 10^5
- Эпох до разреживания: 20
- Коэффициент разреживания матрицы Φ: -0.5
- Коэффициент разреживания матрицы Θ: -1.5
- Эпох после разреживания: 10
- Порог отнесения документа к теме: 0.5

Вывод

На рисунке 8 по оси x расположены темы и количество относящихся к ним документов, а по оси y – средняя точность попадания в эту тему и стандартное отклонение. Можно сделать вывод, что с увеличением размера темы вероятность верного отнесения увеличивается, потому что тема менее подвержена изменениям при удалении небольшого количества документов. При добавлении новых тем в онлайн модель качество отслеживания снижается несильно. Метод с сопоставлением тем показывает качетсво сильно ниже онлайн модели везде, кроме темы пять. Тема пять представлена в таблице ниже, и по ее наиболее вероятным словам можно сделать вывод, что она объединяет документы про технику безопастности зимой. Из-за специфичности темы для конкретного времени года с ее отслеживанием отлично справились все три модели. Таким образом гипотеза о хорошем качестве отслеживания у онлайн модели подтвердилась.

	Тема 5
Самые вероятные слова	телефон пожарный спасатель единый up лед набор стационарный мобильный сеть пожар пользователь случай mchs gov ru вызов возникновение амурский
Примеры заголовков	1 Сезон спасения рыбаков 2 В Приамурье открыли еще две ледовые переправы 3 Как уберечь свой автомобиль от пожара 4 Не забывайте уступать дорогу спецавтотранспорту

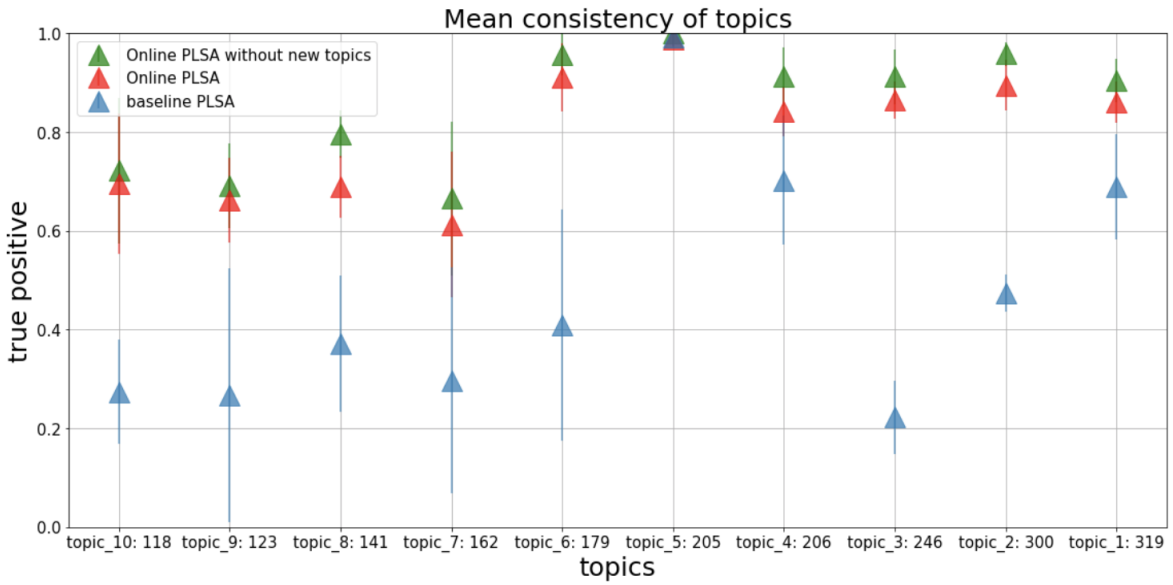

Рис. 8: Точность попадания в изначальную тему (выше – лучше)

5.3 Сравнение с офлайн моделью

В данном разделе производится качественное и количественное сравнение моделей Онлайн PLSA и стандартной PLSA. Онлайн подход более ограничен, потому что у каждой темы есть время создания и документы из периодов раньше времени создания не могут к ней относиться. Обычная PLSA модель строит темы, максимизируя правдоподобие всей коллекции целиком. В статье [1] качество Online LDA модели было хуже, чем у обычной LDA модели.

5.3.1 Сравнение с офлайн моделью: качество

Гипотеза
В этом эксперименте проверяется гипотеза о том, что модель Онлайн PLSA работает по качеству сравнимо с офлайн PLSA моделью. Качество измеряется с помощью введенной в разделе 4.5 метрики когерентности.
Архитектура эксперимента
На данных с метками времени запускается модель Онлайн PLSA, и на каждом периоде считается средняя когерентность тем в модели. После чего на данных со всех периодов запускается PLSA модель, и считается среднее качество ее тем аналогично тому, как оно считалось для онлайн модели на каждом периоде. Таким образом, можно анализировать изменение качества онлайн модели во времени и сравнивать его с качеством офлайн модели.
Реализация (коллекция на русском)

Модель Онлайн PLSA была запущена на датасете на русском языке, и на каждой итерации алгоритма когерентность считалась по всем обработанным к этому периоду документам с помощью Top Tokens Score из пакета BigARTM по 20 наиболее вероятным токенам в каждой теме. Для офлайн модели когерентность считалась аналогичным образом по всей коллекции. Количество тем в офлайн подходе выбиралось по наибольшей средней когерентности. Параметры онлайн модели совпадают с экспериментом 5.2.1, а параметры офлайн модели приведены ниже.

Параметры офлайн модели

- Количество тем: 38
- Коэффициент деккореляции: 10^5
- Эпох до разреживания: 20
- Коэффициент разреживания матрицы Φ: -0.5
- Коэффициент разреживания матрицы Θ: -1.5
- Эпох после разреживания: 10

Реализация (коллекция RCVI)

Модель Онлайн PLSA была запущена на датасете RCVI, и на каждой итерации алгоритма когерентность считалась с помощью Palmetto библиотеки по 10 наиболее вероятным токенам. В этой библиотеке совстречаемость слов вычисляется по википедии. Было показано, что такая метрика для новостных датесетов среди всех вариантов подсчета когерентности коррелируют с ассесорскими оценками лучше всего [16]. В онлайн подходе на каждом шаге в модель добавляется 30 тем, в офлайн подходе количество тем равно количеству тем в модели Онлайн PLSA в конце работы. На рисунке изображены среднее значение и стандартное отклонение когерентности на каждом периоде для Онлайн PLSA и те же метрики для офлайн модели. Значения для офлайн модели не изменяются во времени, так как она строится один раз по всем документам. Параметры онлайн модели совпадают с экспериментом 5.2.1, а параметры офлайн модели приведены ниже.

Параметры офлайн модели

- Количество тем: 960
- Коэффициент деккореляции: 10^5
- Эпох до разреживания: 20

- Коэффициент разреживания матрицы Φ: -6
- Коэффициент разреживания матрицы Θ: -3.5
- Эпох после разреживания: 10

Вывод
На рисунке 9 изображено изменение качества онлайн модели во времени для датасета на русском языке, и в конце периода оно немного выше, чем у офлайн подхода. Для датасета RCVI средняя когерентность офлайн подхода немного выше, однако разброс качества тем в офлайн подходе больше. Таким образом гипотеза о сравнимом качестве моделей подтвердилась.
Также следует отметить, что на рисунке 10 качество модели не сильно меняется во времени, в отличие от рисунка 9. Это связано с тем, что когерентность, посчитанная по Википедии, является более точной оценкой качества тем и предпочтительно использовать ее. Однако, из-за специфики датасета эта метрика не применима для новостей по компании.

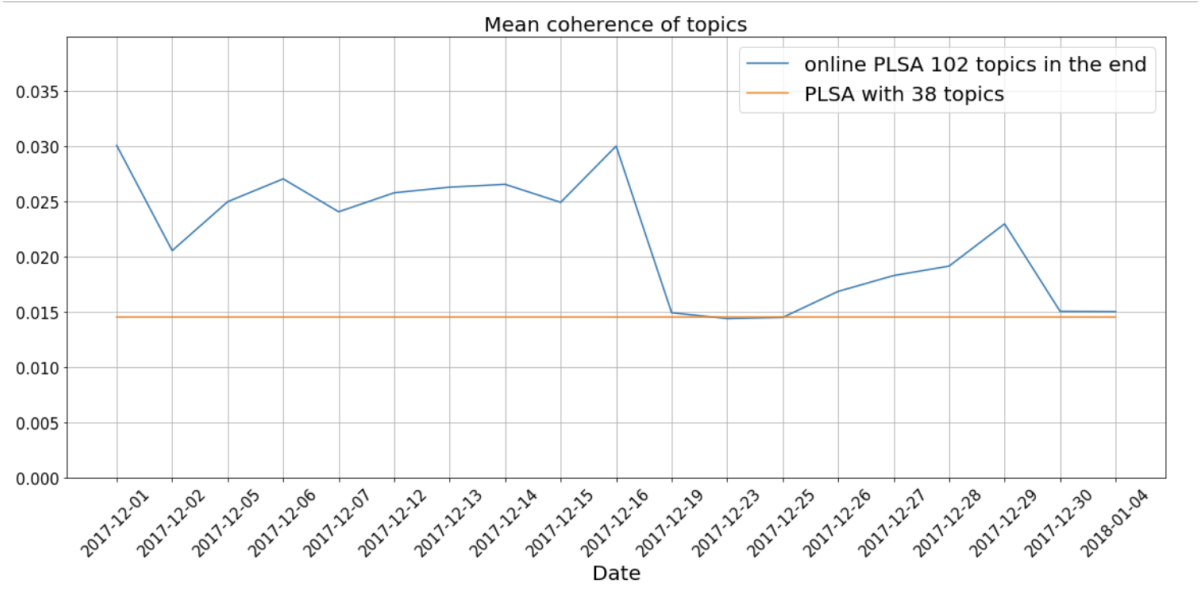

Рис. 9: Когерентность в зависимости от времени: датасет на русском языке (выше – лучше)

5.3.2 Сравнение с офлайн моделью: выделение событий

Гипотеза
В этом эксперименте проверяется гипотеза о том, что самые важные

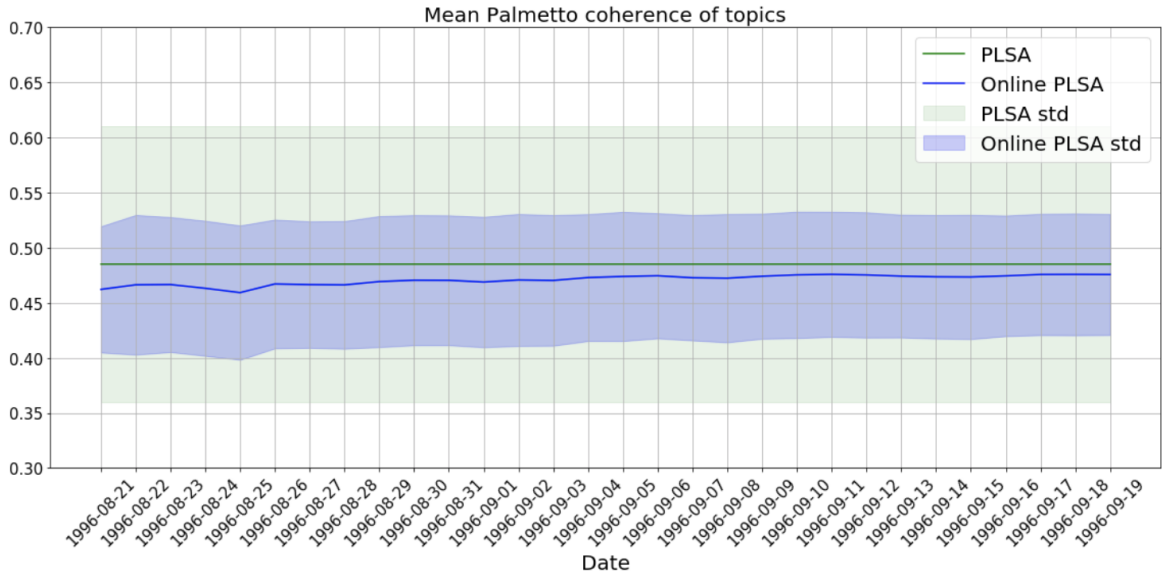

Рис. 10: Когерентность в зависимости от времени: датасет RCVI (выше – лучше)

темы, которые выделяются в офлайн модели, также присутствуют в модели Онлайн PLSA. Этот эксперимент необходим, потому что он показывает, получат ли пользователи онлайн модели информацию о ключевых новостных событиях.

Архитектура эксперимента

На данных с метками времени запускается модель Онлайн PLSA, и на всех данных без учета времени обучается офлайн PLSA модель. После этого вычисляется размер тем в офлайн PLSA модели по формуле:

$$\text{topic size}(t) = \frac{1}{|D|}\sum_{d\in D} p(t|d) = \frac{1}{|D|}\sum_{d\in D}\theta_{td}.$$

Это делается для того, чтобы взять несколько самых значимых тем в модели, то есть тем с большим размером. Для выбранных тем ищется самая близкая тема из онлайн модели и строится соответствие.

Реализация

Были запущены Онлайн PLSA и офлайн модели на датасете на русском языке, параметры моделей аналогичны предыдущему эксперименту. В офлайн модели были взяты десять самых больших тем. Соответствующая тема из онлайн модели для темы t_u из офлайн модели ищется по косинусной мере близости:

$$\text{online topic}(t_u) = \text{argmax}_{t_o\in T_o}\frac{\phi^{t_o}_{\text{online}}\ \phi^{t_u}_{\text{upto}}}{||\phi^{t_o}_{\text{online}}||_2\ ||\phi^{t_u}_{\text{upto}}||_2},$$

где T_o – это список тем в онлайн модели, а $\phi^{t_o}_{\text{online}}$ – это распределение темы t_o из онлайн модели и $\phi^{t_u}_{\text{upto}}$ – это распределение темы t_u из офлайн модели. После этого каждая тема именовалась заголовком самой вероятной новости.

Вывод

Результаты эксперимента приведены в таблице 1. Для всех тем, кроме одной, найденные пары тем очень похожи, из чего можно сделать вывод, что модель Онлайн PLSA хорошо справляется с обнаружением ключевых трендов в новостях. Для седьмой темы сопоставленная тема не соответвует ей по смыслу. Однако, похожая тема про работу государственных учереждений в праздничные дни (имеется в виду Новый год) все-таки присутствует в модели, но она оказалась второй по косинусной близости. В таблице она приведена под номером 2 в той же ячейке.

5.3.3 Сравнение с офлайн моделью: распределения тем во времени

Гипотеза

В этом эксперименте проверяется гипотеза о том, что в Онлайн PLSA модели строятся темы с определенным паттерном во времени: у темы есть начало, когда в нее пришло много документов, а потом на стадии отслеживания в нее приходят еще документы, которые напечатали с задержкой или которые являются развитием событий. Предполагается, что в офлайн модели распределения будут бессистемными, потому что в модели время в явном виде нигде не учитывалось.

Архитектура эксперимента

На данных с метками времени запускается модель Онлайн PLSA, и на всех данных без учета времени обучается офлайн PLSA модель. В обеих моделях ищутся наибольшие темы по формуле из раздела 6.3.2. Для этих тем вычисляется количество документов, у которых $p(t|d) > 0.5$. Далее для каждой темы визуализируется зависимость количества документов от времени, то есть в какой период сколько новостей пришло по этой теме (за время берется время публикации новости). Также для отнормированных на количество новостей в теме распределений считалась энтропия для оценки близости распределения к равномерному.

Реализация

Реализация моделей совпадает с пунктами 6.3.1-6.3.2. Для посчета энтропии использовалась стандартная функция из пакета scipy.

Вывод

Результаты эксперимента приведены на рисунках 11-12 для онлайн и офлайн подходов соответсвенно. На картинках представлены самые круп-

Темы офлайн модели	**Темы Онлайн PLSA модели**
В сети «МегаФона» зарегистрировано более 14 млн мобильных устройств с поддержкой 4G	Высокоскоростной интернет стандарта 4G от «МегаФон»
ФАС завершит анализ данных об исполнении операторами требований по внутрисетевому роумингу к 25 декабря	ФАС разрешила мобильным операторам отложить отмену внутрисетевого роуминга
ПАО «МегаФон» привлечено к административной ответственности за эксплуатацию БС стандарта GSM 900 МГц без соответствующего разрешения на использование частот.	В период с 02.10.2017 по 26.12.2017 завершено мероприятие в отношении ПАО «Мегафон»
"МегаФон"ждет по итогам 2017 года роста OIBDA до 118 млрд рублей	«МегаФон» вырос почти на треть
"МегаФон"обновит совет директоров	Газпромбанк предложил кандидатов в совет директоров «МегаФона»
Российская сборная силы и духа	Спортсменки Приморья готовятся к первому в истории России Чемпионату мира по футболу среди детей-сирот
Режим работы больниц Владимира в новогодние праздники	1 Как найти потерянный телефон 2 Как каникулы изменят режим работы амурских больниц, почты и таможни
"Мегафон"полностью расплатился за долю в Mail.ru Group	«Мегафон» и Mail.ru Group создадут предприятие для совместных проектов
В России на технологии блокчейн в 2018 году могут быть выпущены облигации на 15 млрд рублей	Облигации на блокчейне могут быть выпущены на 10-15 млрд рублей в 2018 году
Подсчитаны потери от «пакета Яровой»	РСПП рекомендует смягчить условия хранения данных по "закону Яровой"

Таблица 1: Таблица соответствия тем офлайн и онлайн моделей

ные темы в каждой из моделей. Гипотеза о том, что в онлайн подходе темы имеют четкое время начала, когда в них добавляется значительное количество документов, подтвердилась. Для сравнения, распределения для офлайн модели таким паттерном не обладают. Этот результат также подтверждается средней энтропией распределений (отнормированных) количества документов от времени, которые представлены в таблице ниже. Средняя энтропия для онлайн модели значительно ниже, чем в офлайн модели, что означает, что эти распределения в случае офлайн модели более равномерные. Также в таблице приведены значения сред-

ней энтропии тем для Онлайн PLSA и Онлайн PLSA без регуляризации новых документов. Введение регуляризации для новых документов снижает среднюю энтропию и стандартное отклонение.

Средняя энтропия	**mean**	**std**
Модель PLSA	1.792	0.566
Онлайн PLSA без рег. новых документов	0.667	0.586
Онлайн PLSA	0.641	0.542

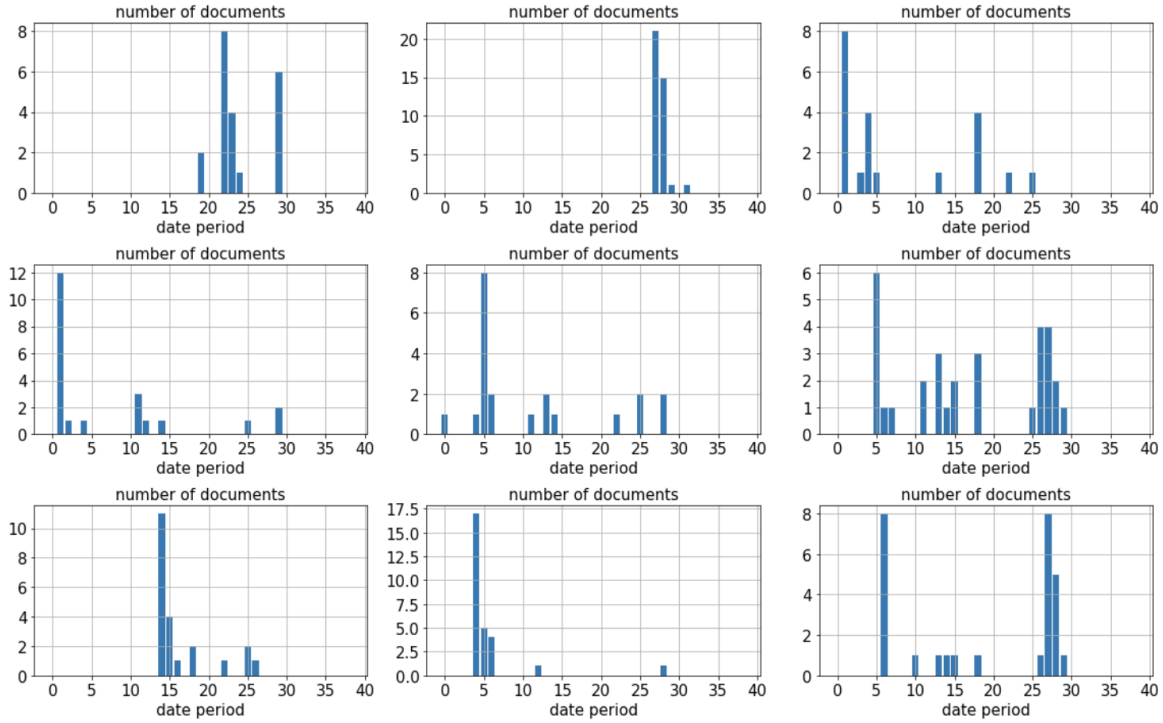

Рис. 11: Распределения новостей для модели Онлайн PLSA

5.4 Вывод

В данном разделе были проведены эксперименты по анализу качества отслеживания тем во времени в модели Онлайн PLSA и по сравнению предложенного подхода со стандарной тематической моделью. В ходе экспериментов гипотезы о хорошем качестве отслеживания подвердились. Также было показано, что по основным критериям, таким как когерентность и выделение ключевых тем, модель не уступает стандарной PLSA.

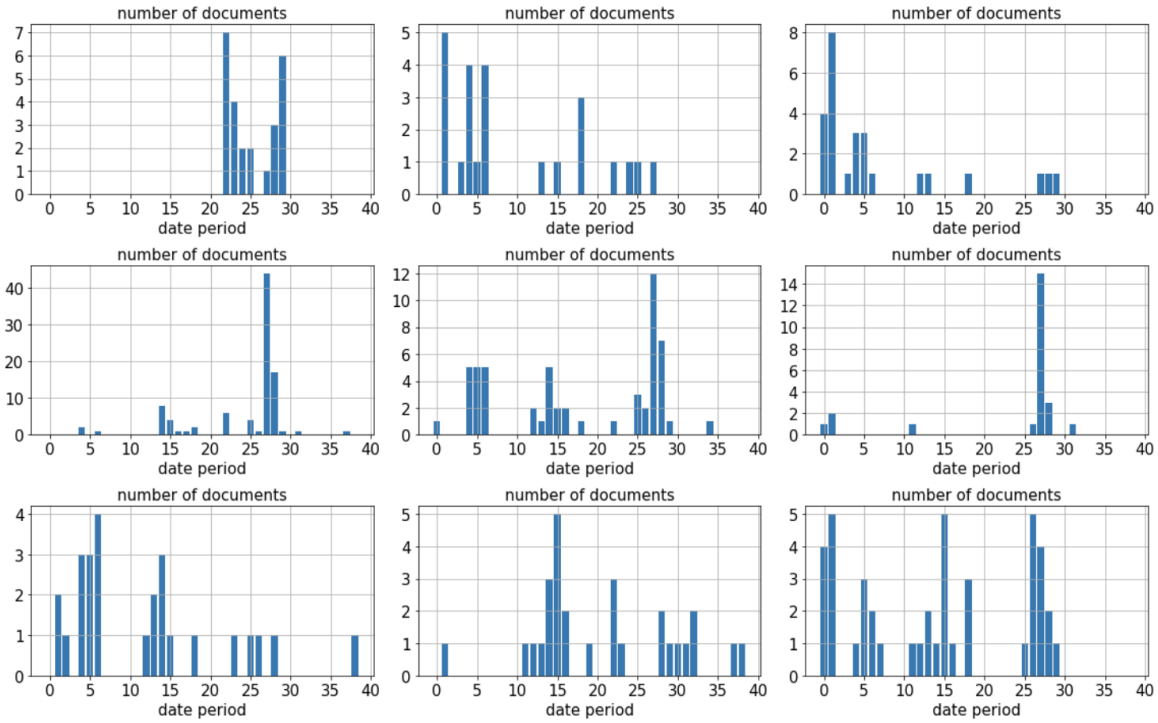

Рис. 12: Распределения новостей для модели PLSA

6 Заключение

В работе был предложен метод решения задачи отслеживания и обнаружения тем с помощью модели PLSA и аддитивной регуляризации. Представленная модель обладает динамически растущим словарем и нацелена на поиск новостных событий. В ходе исследования были достигнуты следующие цели:

- Была построена тематическая онлайн модель, работающая с качеством, сопоставимым с тематической моделью на всех данных
- Была разработана методика оценивания отслеживания тем в темпоральных тематических моделях

Дальнейшие исследования в этой области могут быть связаны с выбором количества тем на каждом шаге онлайн алгоритма. Существуют работы, посвященные алгоритмам выбора количества тем, основанные на анализе стабильности, например [9]. В них используется предположение, что модель с оптимальным количеством тем наиболее устойчива к удалению небольшого количества данных. Однако такие методы не применимы для коллекций с небольшим количеством новостей в период времени, потому что в таком случае удаление даже небольшого количества документов

сильно скажется на любой модели. Еще одной открытой задачей остается удаление устаревших тем из модели. При построении онлайн модели на длительном отрезке времени удаление тем может значительно ускорить скорость работы алгоритма.

7 Библиографический список

1. *AlSumait L.*, *Barbará D.*, *Domeniconi C.* On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking. // ICDM. — IEEE Computer Society, 20 февр. 2009. — С. 3—12.
2. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections / K. Vorontsov [и др.] // AIST'2015, Analysis of Images, Social Networks and Texts. Springer International Publishing Switzerland. — 2015. — С. 370—384.
3. *Blei D. M.*, *Lafferty J. D.* Dynamic Topic Models // Proceedings of the 23rd International Conference on Machine Learning. — Pittsburgh, Pennsylvania, USA : ACM, 2006. — С. 113—120. — (ICML '06).
4. *Blei D. M.*, *Ng A. Y.*, *Jordan M. I.* Latent Dirichlet Allocation // J. Mach. Learn. Res. — 2003. — Март. — Т. 3. — С. 993—1022.
5. *Bruggermann D.*, *Hermey Y.* Storyline detection and tracking using Dynamic Latent Dirichlet Allocation // Proceedings of the 2nd Workshop on Computing News Storylines. — 2016.
6. *Ding C.*, *Li T.*, *Peng W.* On the Equivalence Between Non-negative Matrix Factorization and Probabilistic Latent Semantic Indexing //. Т. 52. — Amsterdam, The Netherlands, The Netherlands : Elsevier Science Publishers B. V., апр. 2008. — С. 3913—3927.
7. Fast and Modular Regularized Topic Modelling / K. Vorontsov [и др.] // Proceedings of the 21st Conference of Open Innovations Association FRUCT. — Helsinki, Finland : FRUCT Oy, 2017. — С. 182—193. — (FRUCT'21).
8. *Frei O.*, *Apishev M.* Parallel non-blocking deterministic algorithm for online topic modeling // 5th International Conference - Analysis of Images, Social networks and Texts. — 2016. — С. 132—144.
9. *Greene D.*, *O'Callaghan D.*, *Cunningham P.* How Many Topics? Stability Analysis for Topic Models // Machine Learning and Knowledge Discovery in Databases / под ред. Т. Calders [и др.]. — Berlin, Heidelberg : Springer Berlin Heidelberg, 2014. — С. 498—513.
10. *Hofmann T.* Probabilistic Latent Semantic Indexing // Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. — Berkeley, California, USA : ACM, 1999. — С. 50—57. — (SIGIR '99).

11. *Keane N.*, *Yee C.*, *Zhou L.* Using Topic Modeling and Similarity Thresholds to Detect Events // EVENTS@HLP-NAACL. — 2015.

12. *Kuhn H. W.*, *Yaw B.* The Hungarian method for the assignment problem //. — 1955. — С. 83—97.

13. *Newman D.*, *Karimi S.*, *Cavedon L.* External evaluation of topic models // in Australasian Doc. Comp. Symp., 2009. — С. 11—18.

14. *Rasmussen C. E.*, *Ghahramani Z.* Occam's Razor // In Advances in Neural Information Processing Systems 13. — MIT Press, 2001. — С. 294—300.

15. RCV1: A New Benchmark Collection for Text Categorization Research / D. D. Lewis [и др.] //. Т. 5. — JMLR.org, дек. 2004. — С. 361—397.

16. *Röder M.*, *Both A.*, *Hinneburg A.* Exploring the Space of Topic Coherence Measures // Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. — Shanghai, China : ACM, 2015. — С. 399—408. — (WSDM '15).

17. Towards a Topic Discovery and Tracking System with Application to News Items / D. Brüggermann [и др.] // Future and Emerging Trends in Language Technology. Machine Learning and Big Data / под ред. J. F. Quesada, F.-J. Martín Mateos, T. López Soto. — Cham : Springer International Publishing, 2017. — С. 183—197.

18. *Vorontsov K.*, *Potapenko A.* Additive regularization of topic models // Machine Learning. — 2015. — Т. 101, 1-3. — С. 303—323.

19. *Воронцов К.* Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН. — 2014. — С. 268—271.