

**ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ**  
**Федеральное государственное автономное образовательное**  
**учреждение высшего образования**

**Национальный исследовательский университет**  
**«Высшая школа экономики»**

**Факультет гуманитарных наук**  
**Образовательная программа**  
**«Фундаментальная и компьютерная лингвистика»**

**КУРСОВАЯ РАБОТА**

На тему «Какие темы всегда с нами? - исследование изменения новостных тематик с течением времени»

*Тема на английском «What Topics are Always with us? - Study of Changes in News Topics Over Time»*

Студентка 3 курса  
группы №212  
Куканова Абина Денисовна

Научный руководитель  
Клышинский Эдуард  
Станиславович  
Доцент, Школа лингвистики

Москва, 2024 г.

## **Содержание**

<b>1 Аннотация</b>	<b>3</b>
1.1 Аннотация . . . . .	3
1.2 Abstract . . . . .	3
1.3 Список ключевых слов . . . . .	4
<b>2 Введение</b>	<b>4</b>
2.1 Описание проблемы и актуальности темы . . . . .	4
2.2 Цели и задачи работы . . . . .	5
<b>3 Теоретические основы</b>	<b>5</b>
3.1 Аддитивная Регуляризация . . . . .	5
3.2 BERTopic . . . . .	8
3.3 Dynamic Topic Modeling . . . . .	9
<b>4 Метрики для оценки модели ARTM</b>	<b>10</b>
<b>5 Эксперименты</b>	<b>11</b>
5.1 Исходные данные . . . . .	11
5.2 Подготовка данных . . . . .	13
5.3 Lenta.ru . . . . .	14
5.4 РИА Новости . . . . .	19
5.5 Газета.Ru . . . . .	26
<b>6 Результаты</b>	<b>32</b>
<b>7 Заключение</b>	<b>34</b>
<b>8 Литература</b>	<b>34</b>

## **1. Аннотация**

### *1.1. Аннотация*

Настоящая работа посвящена методам анализа тематики большой текстовой коллекции и ее динамики во времени. В работе используется два подхода к моделированию тем. Первый подход включает в себя построение тематической модели, основанной на аддитивной регуляризации и учитывающей метки времени, чтобы отследить событийность тем. Второй подход использует языковую модель из семейства BERT для построения тематической модели и анализирует, как темы изменяются с течением времени. Основная задача данной работы заключается в построении тематических моделей для коллекции новостных статей с нескольких российских новостных порталов за 10 лет, используя выше упомянутые методы. Цель данного исследования заключается в анализе долговременной динамики новостных тематик и выявлении основных тем, которые постоянно присутствуют в информационном пространстве. Дальнейшие исследования в этой области могут включать улучшение качества обучения моделей, применение других подходов к построению темпоральной тематической модели.

### *1.2. Abstract*

This work is devoted to methods of analyzing the subject matter of a large text collection and its dynamics over time. The paper uses two approaches to topic modeling. The first approach involves building a topic model based on additive regularization and taking into account timestamps in order to track the eventfulness of topics. The second approach uses a language model from the BERT family to build a topic model and analyzes how topics change over time. The main task of this work is to build topic models for a collection of news articles from several Russian news agencies over 10 years using the above-mentioned methods. The purpose of this study is to analyze the long-term dynamics of news topics and identify the main topics that are constantly present in the information space. Further research in this area may include improving the quality of model learning, applying other approaches to building a temporal topic model.

### *1.3. Список ключевых слов*

тематическое моделирование, аддитвная регуляризация, модальности, bigartm, bertopic, dynamic topic modeling

## **2. Введение**

### *2.1. Описание проблемы и актуальности темы*

Классические тематические модели разбивают коллекцию текстовых документов на некоторое количество тем и определяют, к каким темам относятся документы и какие слова составляют каждую тему. Однако такие модели не учитывают динамику текстовой структуры во времени, а только определяют темы в текстовой коллекции, не отражая их изменение с течением времени. Это ограничивает возможность анализа различных событий, трендов, изменений в обществе.

Темпоральная тематическая модель является важным инструментом для анализа изменения тем во времени. С постоянно растущим объемом данных, возникает необходимость в разработке методов, способных учитывать не только тематическую структуру документов, но и их эволюцию во времени. Интерес к темпоральным тематическим моделям обусловлен необходимостью понимания динамики информации, поиска трендов и изменений в поведении пользователей на основе анализа контента. Это актуально как для исследования социальных сетей, так и для прогнозирования событий в различных областях, таких как финансы, медицина, политика и многое другое.

В отличие от привычных нам тематических моделей, темпоральные тематические модели не только разбивают коллекцию текстовых документов на некоторое количество тем, но и учитывает метки времени каждого документа. Такие модели позволяют отслеживать, как темы появляются, развиваются и исчезают во времени. Модель показывает, как темы становятся более популярными и менее популярными со временем, а какие и вовсе постепенно исчезают из коллекции текстовых документов. Также модель позволяет классифицировать темы по мере их событийности: постоянные темы и темы-события. Постоянные темы проходят «красной нитью» по всему исследуемому временному интервалу, а событийные темы внезапно появля-

ются во временном промежутке и постепенно исчезают.

В этой работе с помощью использования библиотек ARTM и BERTopic предполагается построить тематические модели, которые будут обучены на текстовых коллекциях статей с российских новостных порталов за последние 10 лет.

## 2.2. Цели и задачи работы

Целью данного исследования является построение темпоральной тематической модели для коллекции новостных постов, которая будет учитывать временную принадлежность каждого документа. В работе используется подход аддитивной регуляризации при построении модели в одном случае, а в другом, чтобы построить тематическую модель, используется языковая модель из семейства BERT. Чтобы определить, какие темы являются событийными, а какие относятся к классу постоянных, необходимо построить тематическую модель, используя аддитивную регуляризацию. Кроме того, сравнить классическую модель PLSA и модель ARTM, к которой добавили регуляризаторы и модальности. Другой задачей данной работы является рассмотреть эволюцию тем с течением временем, и для этого построить динамическую тематическую модель.

## 3. Теоретические основы

### 3.1. Аддитивная Регуляризация

Тематическое моделирование коллекции текстовых документов определяет, к каким темам относится каждый документ коллекции. Алгоритм модели на входе получает коллекцию текстовых документов, а на выходе каждого документа выводит числовой вектор, который состоит из оценок степени принадлежности документа каждой из тем. Каждый документ в корпусе представлен как мешок слов, который получается некоторым множеством тем. Исходя из этих данных, алгоритм восстанавливает вероятностные распределения тем в корпусе и определяет, к какому подмножеству относится каждый документ. Для этой задачи не требуется ручная разметка данных, обучение модели происходит без учителя (unsupervised learning). Похоже на задачу кластеризации, но тематическая кластеризация допускает, чтобы один и тот же до-

кумент мог относится к нескольким темам.

Тематическое моделирование используется не только для выявления множества латентных тем, но и для решения следующих задач: разведочный информационный поиск, выявление тематических сообществ в социальных сетях, определение тематики различных сущностей в текстовых коллекциях, аннотирование изображений, а также обнаружение и отслеживание событий в новостных потоках.

Latent Dirichlet Allocation (LDA) является самым известным и одним из популярных методов тематического моделирования. Модель основана на гипотезе, что каждый документ может быть представлен как набор тем, а каждая тема состоит из распределения слов. LDA выявляет скрытые темы в текстовой коллекции, которые можно не заметить при прочтении. К тому же, результаты такой модели легко интерпретировать, так как каждая тема представлена набором слов. Однако, есть проблема в том, что у тематического моделирования существует множество решений, и LDA выбирает только одно из них, не давая возможности указать ограничений на модель. Эту проблему решает аддитивная регуляризация, которая позволяет задавать несколько регуляризаторов за раз. Например, в библиотеке реализованы регуляризаторы, которые сглаживают и разреживают подмножества тем в  $\Theta$  и  $\Phi$  матрицах, используя любое заданное распределение, или же с помощью другого регуляризатора можно декоррелировать столбцы в  $\Phi$  матрице, чтобы повысить интерпретируемость тем. Таким образом, множество тем разбивается на два подмножества, предметные и фоновые темы,  $T = S \sqcup B$ . Предметные темы  $t \in S$  содержат термины предметных областей, их распределения  $p(w|t)$  и  $p(t|d)$  разрежены и существенно различны. Фоновые темы  $t \in B$  содержат слова общей лексики, их не должно быть в предметных темах. Их распределения  $p(w|t)$  могут быть менее разрежены, чем у предметных тем, и совсем на них не похожи. Слова общей лексики большую часть в любых документах, поэтому распределения  $p(t|d)$  должны быть сильно сглажены. Матрицы  $\Theta$  и  $\Phi$  должны иметь такую же разреженность, что и на рис.1.

Регуляризаторы улучшают качество модели: повышают различность тем, точность и полноту поиска, позволяют учитывать дополнительные нетекстовые данные. В целом, регуляризаторы служат как оптимизационные критерии для задания желаемых свойств в тематической модели. В отличие от простых моделей PLSA и

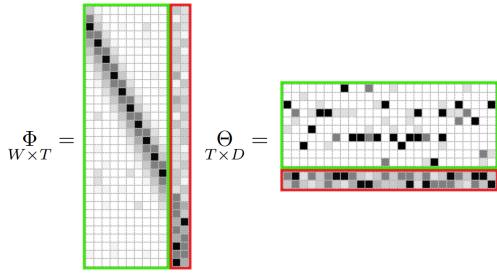


Рис. 1: Структура разреженности  $\Theta$  и  $\Phi$  с предметными и фоновыми темами

LDA, библиотека BigARTM реализует не только регуляризацию, но и такие механизмы, как модальности, тематические иерархии, обработку текста как последовательности тематических векторов слов, использование данных о совместной встречаемости слов, тематизация сложных транзакционных данных. Модальности позволяют не только описывать слова в документах, но и нетекстовые объекты, т.е. токены других модальностей такие, как авторы, временные метки документов, источники, рубрики, именованные сущности и т.д.

При добавлении временных меток к классической тематической модели, информация о времени никак не используется при обучении модели и не влияет на образование тем. Однако эта проблема решается в ARTM путем добавления модальности времени  $I$ . Искомое распределение  $p(i|t) = \varphi_{it}$  получается в таблице матрицы  $\Phi$ . Многие темы являются событийными и недолго задерживаются во временном промежутке, поэтому необходимо провести разреживание распределений  $p(t|i)$  с помощью регуляризатора:

$$R_{sparse}(\Theta/\Phi) = -\tau_{sparse} \sum_{i \in I} \sum_{t \in T} \ln p(t|i)$$

Еще распределения  $p(i|t)$  как функции времени могут меняться не слишком быстро, и для этого вводится регуляризатор сглаживания, минимизирующий модели разностей  $p(i|t)$  в соседних интервалах времени:

$$R_{smooth}(\Theta/\Phi) = -\tau_{smooth} \sum_{i \in I} \sum_{t \in T} |p(i|t) - p(i-1|t)|$$

Оба регуляризатора можно записать и как функции от  $\Phi$ , так и от  $\Theta$ . Вводится модальность времени интервалов в одном из случаев, а в другом - приходится

обеспечивать доступ к вектор-столбцам  $\Theta_d$  документов, относящихся к соседним интервалам  $i \pm 1$ .

### 3.2. *BERTopic*

BERTopic – это метод тематического моделирования, который использует Transformers и c-TF-IDF для создания кластеров, позволяющих легко интерпретировать темы, сохраняя при этом важные слова в описаниях тем. BERTopic был разработан как член семейства BERT специально для моделирования темы. Сначала на вход модели BERT подаются документы, представленные в виде вложений с использованием преобразователей предложений. Далее происходит процесс уменьшения размерности и кластеризации. На шаге кластеризации вычисляется сходство между различными документами для определения принадлежности к набору-теме. Разделив документы на темы, алгоритм c-TF-IDF выделяет наиболее релевантные слова для каждой темы. Наконец, алгоритм MMR улучшает согласованность терминов по одной и той же теме.

BERTopic позволяет использовать DTM, вычисляя представление темы на каждом временном шаге без необходимости запуска всей модели несколько раз. Для этого сначала нужно подогнать BERTopic так, как если бы в данных не было временного аспекта. Таким образом, будет создана общая тематическая модель. Используется глобальное представление основных тем, которые, скорее всего, можно найти на разных временных шагах. Для каждой темы и временного шага рассчитывается представление c-TF-IDF. Это приведет к определенному представлению темы на каждом временном шаге без необходимости создавать кластеры из вложений, поскольку они уже созданы.

Далее, есть два основных способа дальнейшей тонкой настройки этих конкретных представлений тем, а именно глобальный и эволюционный. Представление темы на временном шаге  $t$  может быть точно настроено глобально путем усреднения его представления c-TF-IDF с представлением глобального представления. Это позволяет каждому представлению темы немного приблизиться к глобальному представлению, сохраняя при этом некоторые конкретные слова. Представление темы на временном шаге  $t$  может быть точно настроено эволюционно путем усреднения

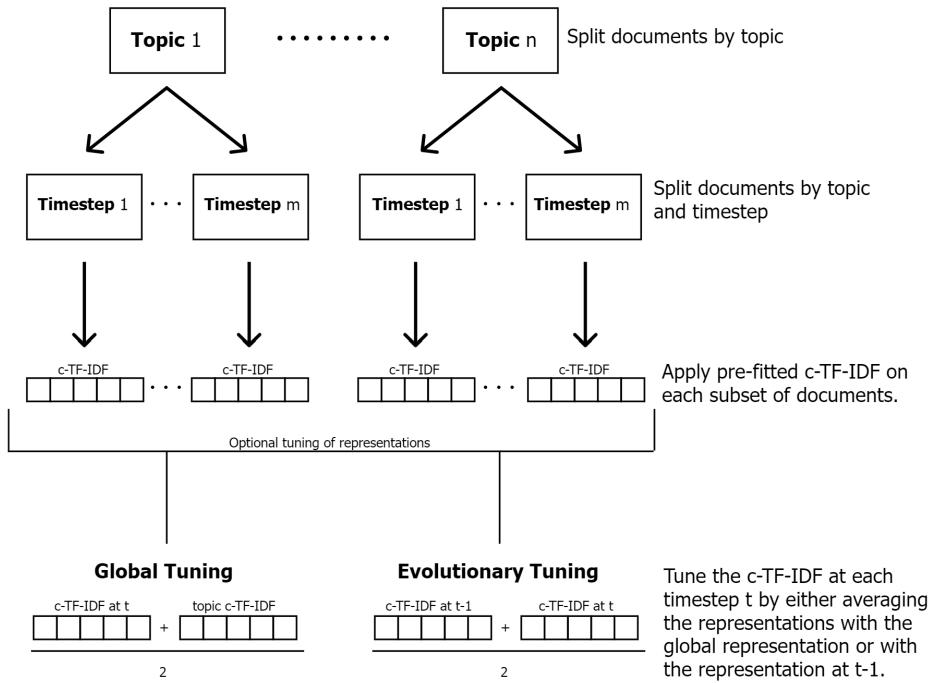


Рис. 2: Схема реализации DTM

его представления с-TF-IDF с представлением с-TF-IDF на временном шаге  $t - 1$  (рис.2). Это делается для каждого представления темы, что позволяет представлениям развиваться с течением времени.

### 3.3. Dynamic Topic Modeling

Dynamic topic models (DTMs) другой вид темпоральных тематических моделей, который моделирует эволюцию тем в текстовой коллекции. Пока другие классические темпоральные модели сосредоточены на непрерывных данных, тематические модели разработаны для категориальных данных. Подход DTM заключается в использовании моделей состояния (state space model) для описания эволюции тем во времени. Вместо того, чтобы моделировать непосредственно сами темы, авторы предлагают моделировать "естественные параметры лежащих в основе многочленов тем".

Этот процесс неявно предполагает, что документы составляются с возможностью обмена по одному и тому же набору тем. Однако для многих коллекций порядок расположения документов отражает меняющийся набор тем. В динамической

тематической модели мы предполагаем, что данные разделены по временным интервалам, например по годам. Мы моделируем документы каждого фрагмента с помощью K-component тематической модели, в которой темы, связанные со срезом  $t$ , развиваются из тем, связанных со срезом  $t-1$ .

Таким образом, подход DTM заключается в моделировании последовательностей композиционных случайных величин путем объединения гауссовых распределений в динамическую модель и сопоставления полученных значений с симплексом. Это расширение логистического нормального распределения для симплексных данных временных рядов. В LDA тематические пропорции  $\Theta$  для конкретного документа рассчитываются на основе распределения Дирихле. В динамической тематической модели мы используем логистическую нормаль со средним значением  $a$  для выражения неопределенности в отношении пропорций. Последовательная структура между моделями снова фиксируется с помощью простой динамической модели.

#### **4. Метрики для оценки модели ARTM**

Выбор коэффициентов регуляризации в ARTM - это важный шаг, который влияет на качество тематической модели. Правильно подобранные коэффициенты помогают улучшить разреженность матриц  $\Phi$  и  $\Theta$ , что приводит к более интерпретируемым темам. Но избыточная регуляризация может приводить к ухудшению качества модели, поэтому необходимы метрики, которые будут отслеживать качество модели.

Перплексия - популярный критерий, используемый для оценивания вероятностных моделей языка в компьютерной лингвистике. Это величина, выражающаяся через правдоподобие выборки и позволяющая отслеживать сходимость метода оптимизации. Формула перплексии:

$$P(D; \Phi; \Theta) = \exp\left(-\frac{1}{n} \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td}\right) = \exp\left(-\frac{1}{n} L(D; \Phi, \Theta)\right),$$

где  $L(D)$  - логарифмическая вероятность параметров модели в наборе документов  $D$ .

Меньшее значение перплексии означает лучшую сходимость модели и может использоваться для сравнения в случае плотных моделей с одинаковым набором тем, словарей и обучающих документов. Чем меньше перплексия, тем лучше.

Разреженность матриц  $\Phi$  и  $\Theta$  вычисляет соотношение элементов матриц (или ее части), которые меньше заданного  $eps$  порога. Одной из целей регуляризации является достижение разреженной структуры матриц  $\Phi$  и  $\Theta$  с использованием различных разреженных регуляризаторов. Метрики качества позволяет контролировать этот процесс. Используя разные стратегии регуляризации в разных частях матрицы можно создать оценку для каждой части и одну для всей матрицы, чтобы иметь подробные и целые значения.

Семантическое ядро - это множество слов, которое с большой вероятностью употребляется в теме  $t$  и редко употребляется в других темах:

$$W_t = w \in W | p(t|w) > \delta, p(t|w) = \phi_{wt} \frac{n_t}{n_w}$$

Независимо от того, каким образом строится семантическое ядро, определяются следующие показатели:

1.  $pur_t = \sum_{w \in W_t} p(w|t)$  - чистота темы (чем выше, тем лучше)
2.  $con_t = \frac{1}{|W_t|} \sum_{w \in W_t} p(t|w)$  - контрастность темы (чем выше, тем лучше)

Высокая контрастность темы говорит о том, что темы хорошо различимы друг от друга, и слова, характерные для одной темы, редко встречаются в других темах. Чистота показывает, насколько хорошо тема описывается своим ядром.

## 5. Эксперименты

### 5.1. Исходные данные

В качестве исходных данных были использованы статьи с российских новостных порталов за последние 10 лет, а именно: Lenta.ru, РИА Новости и Газета.Ru. Для создания коллекции новостных постов с информационного агентства «РИА Новости» был написан парсер для скачивания статей, используя асинхронное программирование. Для других два источника были взяты готовые датасеты для NLP-задач. Все нужные нам документы хранятся в csv формате. К каждому документу в коллекции соответствует своя временная метка в формате ГГГГ-ММ-ДД. Распределение документов каждой коллекции по годам представлено на рис. 3-5. В табл. 1 представлена общая информация о датасетах новостных статей.

	Количество документов	Начало периода	Конец периода
РИА Новости	80332	1 января 2023г.	31 декабря 2023г.
Газета.Ru	74126	1 июня 2010г.	3 сентября 2021г.
Lenta.ru	39563	1 января 2013г.	31 декабря 2023г.

Таблица 1: Информация о текстовых коллекциях

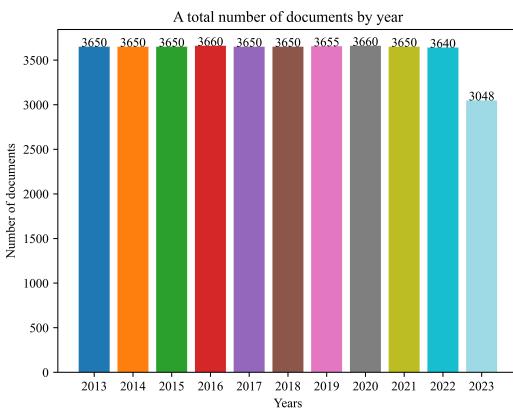


Рис. 3: Распределение документов коллекции Lenta.Ru

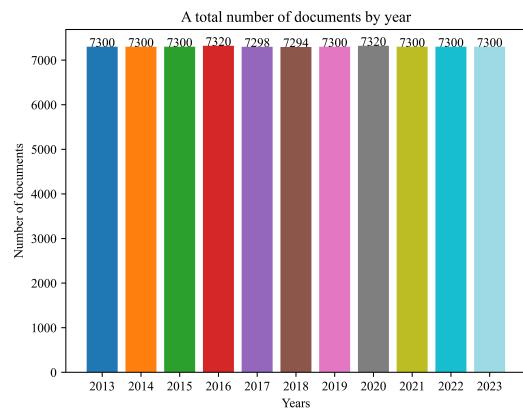


Рис. 4: Распределение документов коллекции РИА Новости

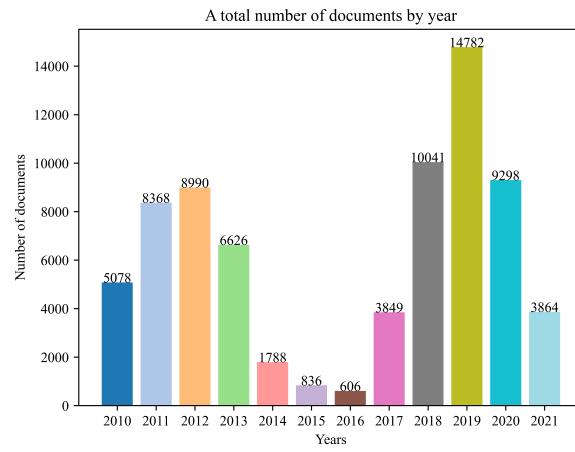


Рис. 5: Распределение документов коллекции Газета.Ru

## *5.2. Подготовка данных*

Для того, что можно быть приступить к построению тематической модели, необходимо сначала провести предварительную обработку текста. Исходные текстовые файлы содержат шум и нерелевантные данные: HTML-теги, числа, знаки пунктуации, стоп-слова и т.д. Необходимо очистить текст от всего лишнего и в дальнейшем работать уже с подготовленными данными. Для каждого текстового документа выполняются следующие действия:

1. Все слова приводятся к нижнему регистру для того, чтобы слова «привет», «Привет» и «ПРИВЕТ» считались одним и тем же словом.
2. В тексте остаются только кириллические символы, убираем специальные символы: избавляемся от всего, что не является "словами"
3. Удаляются стоп-слова – слова, встречающиеся почти в каждом документе. Для получения списка стоп-слов используется библиотека nltk.
4. Удаляются все HTML-теги, все URL и ссылки
5. Приводим словоформы к лемме – ее нормальной (словарной) форме. Для лемматизации текста была выбрана библиотека pymystem3

В модель на вход подаются документы в формате UCI Bag-of-words, который предполагает создание vocab.\*.txt (словарь) и docword.\*.txt (частоты слов в документах). Чтобы при обучении модели учитывались временные метки, то к vocab.\*.txt и docword.\*.txt добавляются метки времени (месяц публикации). В vocab.\*.txt файле указывается модальности меток: @timestamps\_class.

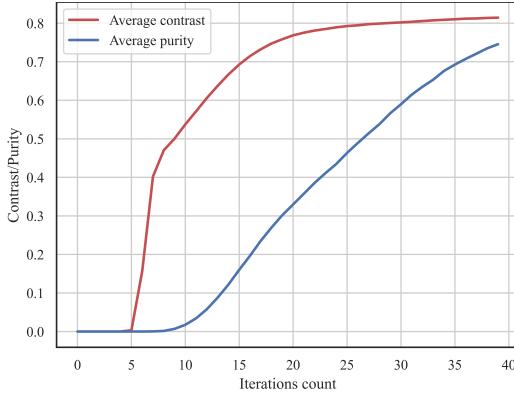
Создается векторизатор по данным из каждой коллекции в формате UCI Bag-of-words, а после разбиения коллекции на пакеты создается вручную словарь BigARTM, который хранит данные о словах и используется в некоторых регуляризаторах и метриках качества.

### 5.3. Lenta.ru

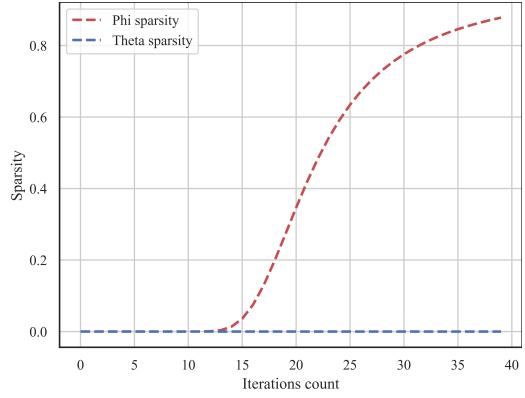
В качестве первичного эксперимента была построена тематическая модель без регуляризатора или же классическая модель PLSA. Было задано 100 тем для эксперимента. Для оценки качества модели были добавлены следующие метрики: перплексия, разреженность матриц, контрастность, чистота ядра темы. Сначала модель обучается с 15 проходами по коллекции. После добавления метрики Top Tokens, которая возвращает наиболее вероятные токены к запрашиваемым темам, происходит обучение уже с 25 проходами по коллекции. Получаем следующие результаты (рис.6). Разреженность матрицы  $\Phi$  - 87,8%, разреженность матрицы  $\Theta$  - 0,0%, средняя контрастность ядра по темам - 0,814, средняя чистота ядра - 0,745, перплексия - 1848.700. Так как была добавлена метрика качества Top Tokens, то можно представить наиболее вероятные токены к темам (табл.2). Видно, что в темах присутствуют слова из фоновой лексики корпуса.

Далее была построена модель ARTM на этой же коллекции новостных потоков. Были добавлены регуляризаторы декорреляции, разреживания и сглаживания матриц  $\Theta$  и  $\Phi$ . Добавили к модели дату публикации новостной статьи. Метрики качества рассматривались таки же, что и у классической модели PLSA. Результаты этой модели показаны на рис.7. Разреженность матрицы  $\Phi$  - 89,3%, разреженность матрицы  $\Theta$  - 19,9%, средняя контрастность ядра по темам - 0,782, средняя чистота ядра - 0,414, перплексия - 1881.285. С точки зрения метрики Delta-AUC были приведены примеры постоянных тем (чем тема постояннее, тем Delta-AUC больше), табл. 3. На рис.8 показаны распределения некоторых событийных тем. На рис.9 показана матрица, отсортированная по метрике Delta-AUC.

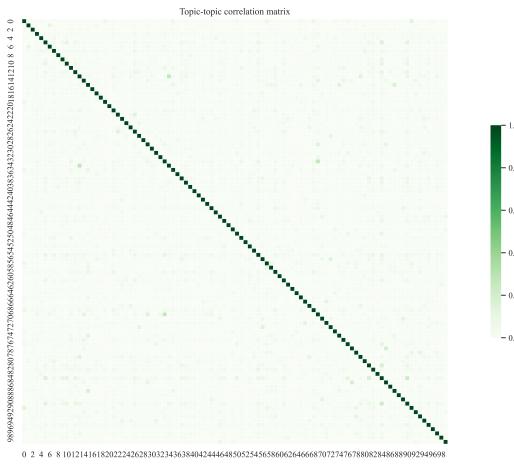
Кроме того, была построена модель BERTopic, в которой реализован подход DTM. На рис.10 изображен график изменения топ-20 самых популярных тем с течением времени



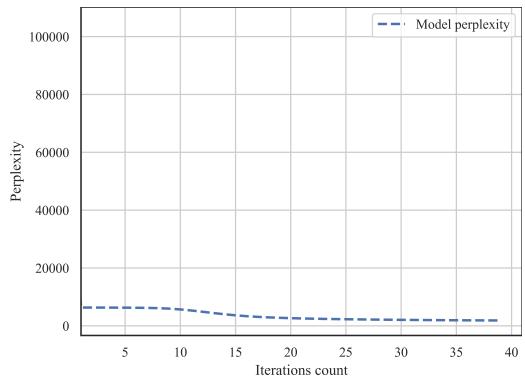
(а) График чистоты и контрастности модели



(б) График разреженности матриц  $\Theta$  и  $\Phi$

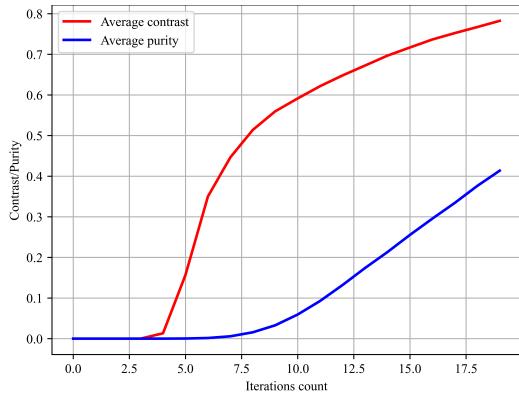


(с) Матрица корреляций между темами

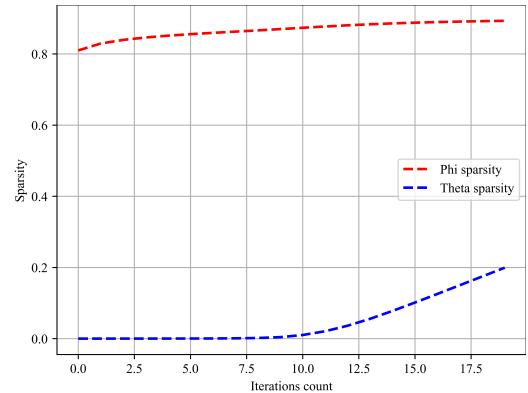


(д) График перплексии

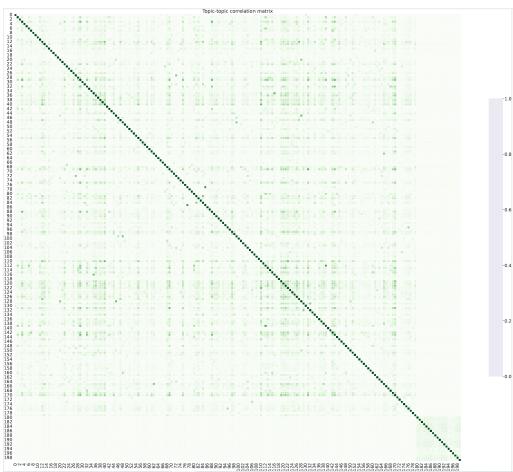
Рис. 6: Графики для модели PLSA коллекции Lenta.ru



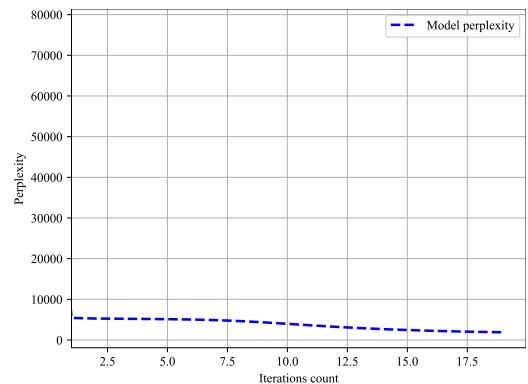
(а) График чистоты и контрастности модели



(б) График разреженности матриц  $\Theta$  и  $\Phi$



(с) Матрица корреляций между темами



(д) График перплексии

Рис. 7: Графики для модели ARTM коллекции Lenta.ru

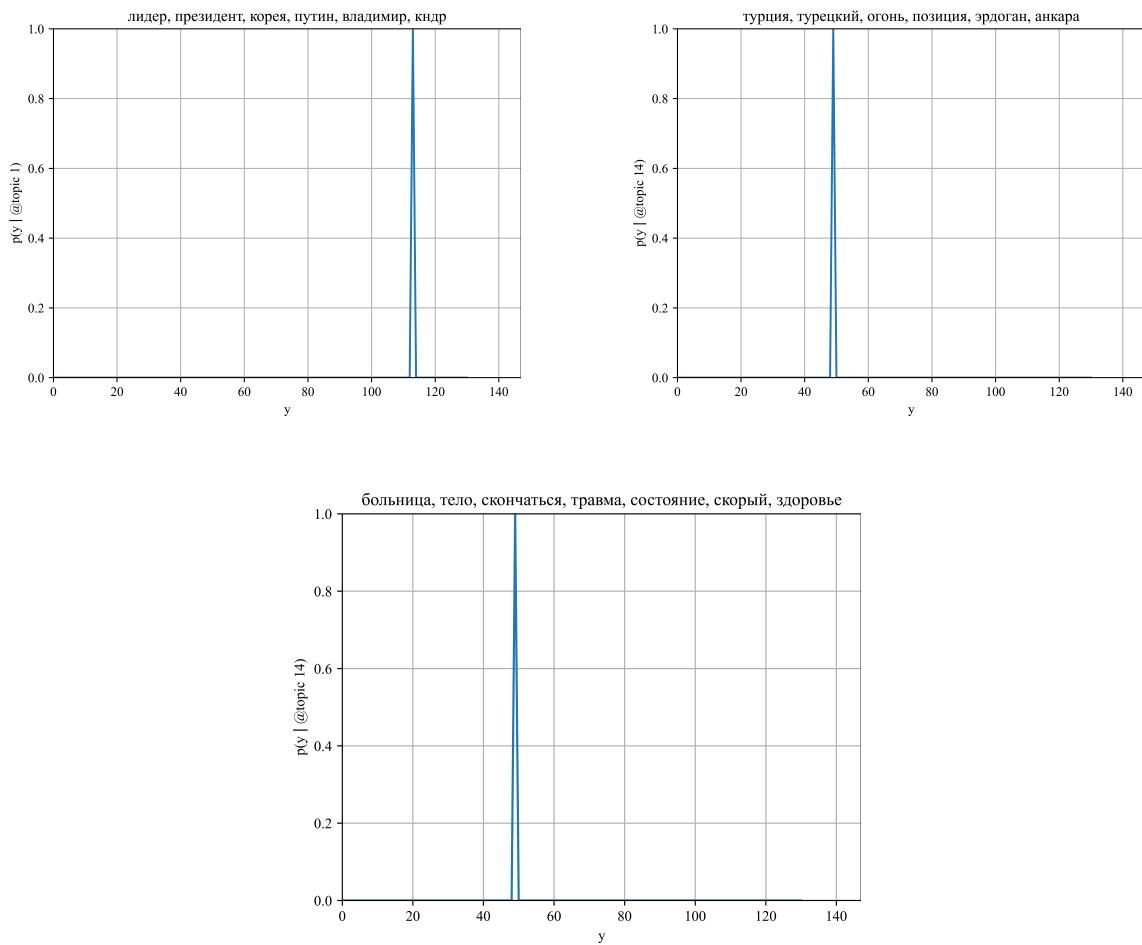


Рис. 8: Распределения некоторых событийных тем Lenta.Ru

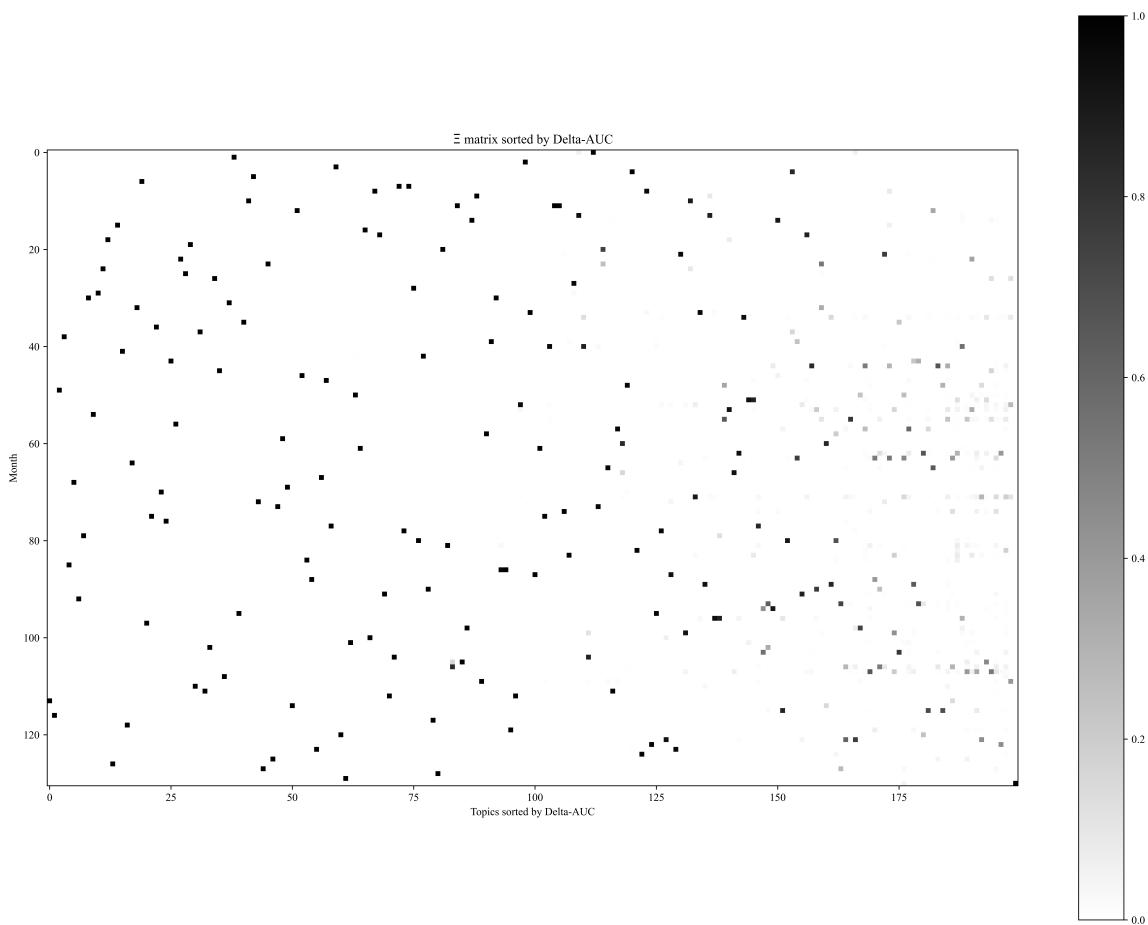


Рис. 9: Матрица  $\Xi$  Lenta.Ru

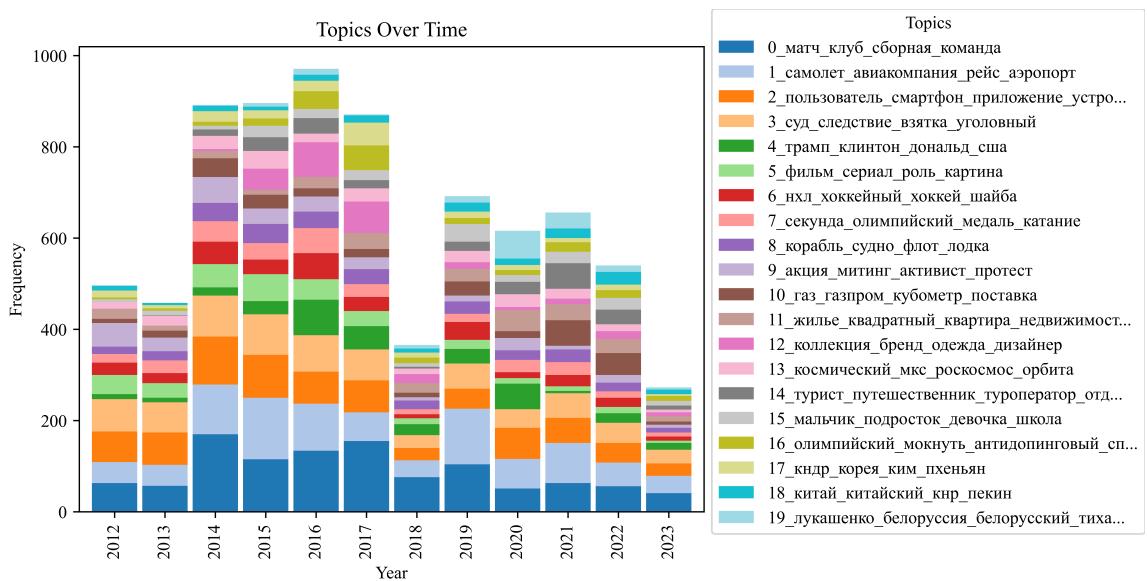


Рис. 10: Значимые темы модели BERTopic Lenta.ru

Таблица 2: Примеры выделенных тем Lenta.ru

@topic_0	исследование, ученый, организм, клетка, болезнь, заболевание, риск, врач, страдать, опасный, мозг, кровь, специалист, боль, доктор
@topic_1	президент, владимир, путин, лидер, глава, государство, зеленский, указ, экс, телефонный, ход, саммит, кремль, встреча, состояться
@topic_2	продукт, магазин, товар, продукция, килограмм, ресторан, питание, посоветовать, вес, хозяйство, еда, мясо, сельский, производитель, правильный
@topic_3	нато, церковь, альянс, общество, вступление, швеция, финляндия, храм, православный, религиозный, кирилл, эстония, ценность, святой, член
@topic_4	миллиард, фонд, бюджет, правительство, увеличивать, средство, налог, выделять, долг, расход, доход, триллион, повышение, условие, финансирование
@topic_5	источник, ссылка, сообщать, база, техника, передавать, технический, вертолет, вестись, тип, выполнять, настоящий, заказ, тасс, за действовать
@topic_6	министр, глава, бывший, пост, премьер, занимать, возглавлять, должность, отставка, политик, обязанность, заявлять, правительство, политика, назначать

#### 5.4. РИА Новости

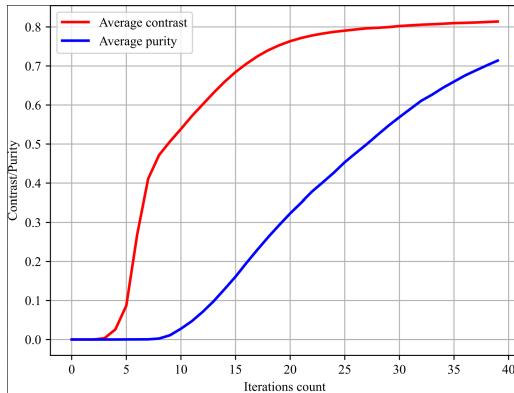
В качестве базового эксперимента модель обучается на коллекции новостных потоков агентства РИА Новости. Количество тем 100. Для оценки качества модели были добавлены те же метрики: перплексия, разреженность матриц, контрастность, чистота ядра темы. Сначала модель обучается с 15 проходами по коллекции. После добавления метрики Top Tokens, которая возвращает наиболее вероятные токены к запрашиваемым темам, происходит обучение уже с 25 проходами по коллекции. Получаем следующие результаты (рис.12). Разреженность матрицы  $\Phi$  - 88,1%, раз-

Таблица 3: Примеры постоянных тем согласно Delta-AUC Lenta.Ru

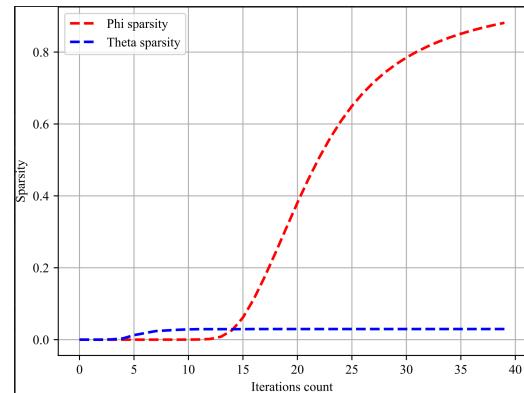
0.477	проект, школа, класс, планироваться, студент, учебный, школьник, обучение, образование, вуз, туризм, научно, образовательный, среда, реализация
0.399	змея, крокодил, укусить, ядовитый, укус, гигантский, синтез, плаэма, прах, кобра, змей, квинсленд, туркменистан, зуб, килограммовый
0.363	фильм, актер, роль, картина, режиссер, снимать, сериал, отель, съемка, выходить, известный, сценарий, лента, сниматься, премьера
0.261	кукла, театральный, мюзикл, чехов, балерина, бриллиант, танцевать, дирижер, филармония, композитор, украшение, универмаг, гусман, худрук, симфонический
0.338	услуга, размер, оплата, получать, мошенник, коммунальный, обеспечение, банка, социальный, работник, долг, жкх, месяц, условие, возможность
0.278	компания, память, менеджер, отчет, компьютер, концерн, доступ, топ, входить, экран, компьютерный, становиться, конкурент, представлять, владелец

реженность матрицы  $\Theta$  - 3,0%, средняя контрастность ядра по темам - 0,814, средняя чистота ядра - 0.714, перплексия - 1611.553.

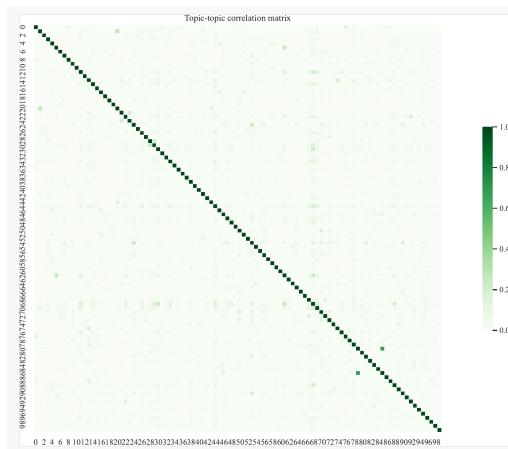
Далее была построена модель ARTM на этой же коллекции новостных потоков. Были добавлены регуляризаторы декорреляции, разреживания и сглаживания матриц  $\Theta$  и  $\Phi$ . Добавили к модели дату публикации новостной статьи. Метрики качества рассматривались таки же, что и у классической модели PLSA. Результаты этой модели показаны на рис.13. Разреженность матрицы  $\Phi$  - 85,4%, разреженность матрицы  $\Theta$  - 30,4%, средняя контрастность ядра по темам - 0.597, средняя чистота ядра - 0.160, перплексия - 2650.423. С точки зрения метрики Delta-AUC были приведены примеры постоянных тем (чем тема постояннее, тем Delta-AUC больше),



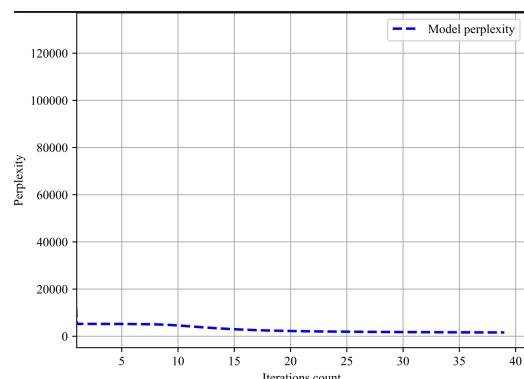
(а) График чистоты и контрастности модели



(б) График разреженности матриц  $\Theta$  и  $\Phi$



(с) Матрица корреляций между темами



(д) График перплексии

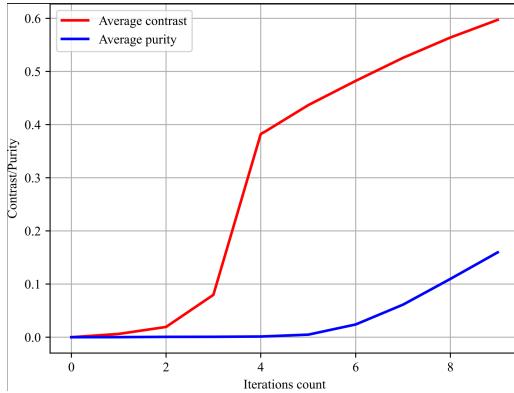
Рис. 11: Графики для модели PLSA коллекции РИА Новости

Таблица 4: Примеры выделенных тем РИА Новости

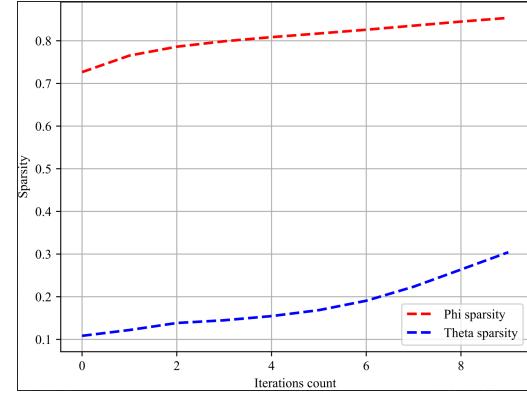
@topic_0	исследование, ученый, продукт, журнал, обнаруживать, клетка, изучать, мозг, эксперимент, университет, статья, использовать, влиять, страдать, организм
@topic_1	президент, владимир, путин, глава, вица, заявлять, секретарь, спикер, песок, государство, кремль, парламентарий, госдума, советник, комитет
@topic_2	партия, германия, греция, немецкий, демократический, парламентский, оппозиционный, фрг, канцлер, политик, править, греческий, политический, депутат, берлин
@topic_3	принимать, международный, участие, процесс, предложение, необходимость, готовый, отмечать, решение, швеция, заявка, подчеркивать, представитель, заявлять, дальнейший
@topic_4	экономический, миллиард, фонд, финансовый, условие, курс, финансирование, долг, бюджет, евро, увеличивать, обязательство, финансы, расход, механизм
@topic_5	рф, крым, состав, федерация, неоднократно, соответствие, возвращение, прежний, крымский, полный, временно, руководство, заявлять, путем, севастополь
@topic_6	пункт, провинция, населенный, прекращение, сутки, гуманитарный, формирование, зона, огонь, примирение, алеппо, эль, вооруженный, зафиксировать, действие

табл. 5. На рис.14 показаны распределения некоторых событийных тем. На рис.15 показана матрица, отсортированная по метрике Delta-AUC.

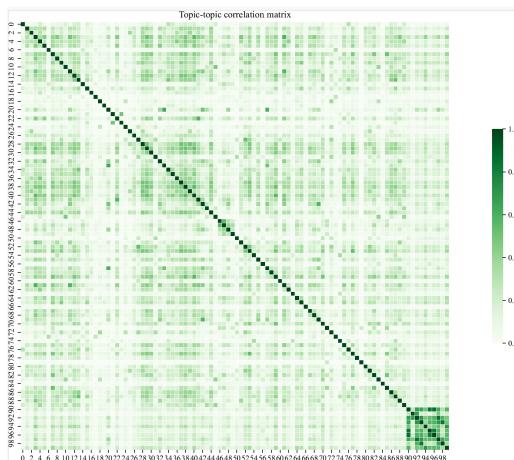
Кроме того, была построена модель BERTopic, в которой реализован подход DTM. На рис.16 изображен график изменения топ-20 самых популярных тем с течением времени



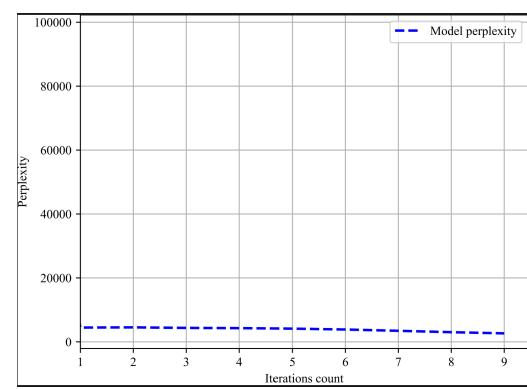
(а) График чистоты и контрастности модели



(б) График разреженности матриц  $\Theta$  и  $\Phi$



(с) Матрица корреляций между темами



(д) График перплексии

Рис. 12: Графики для модели ARTM коллекции РИА Новости

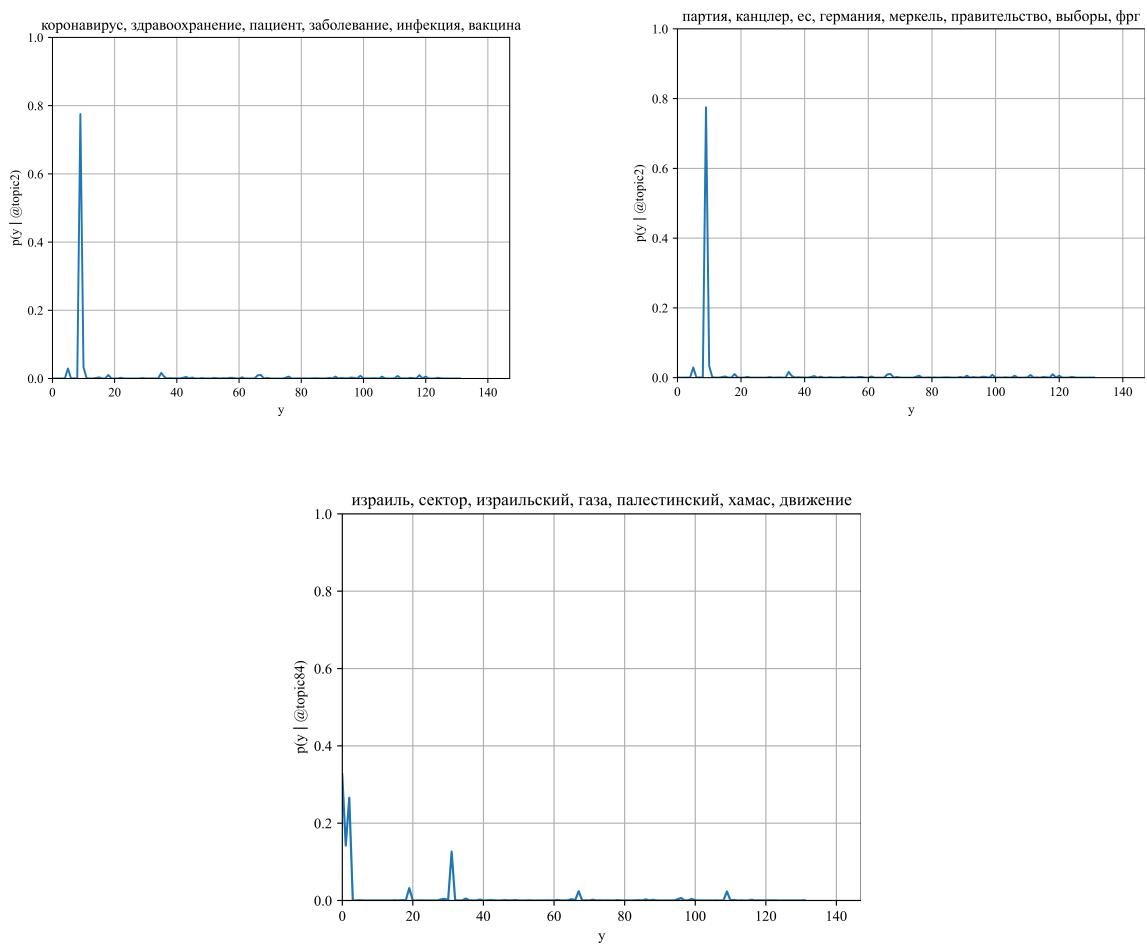


Рис. 13: Распределения некоторых событийных тем

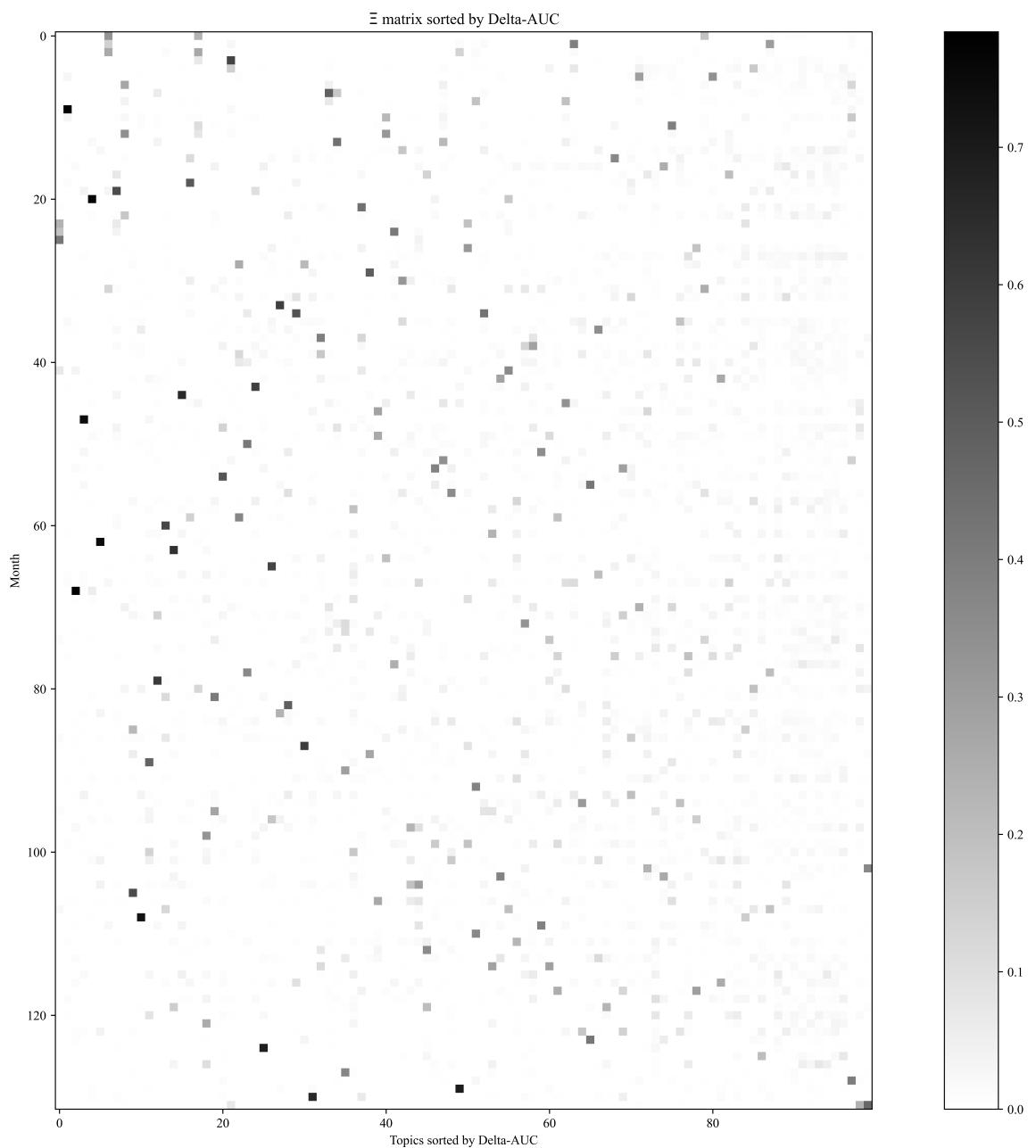


Рис. 14: Матрица  $\Xi$  РИА Новости

Таблица 5: Примеры постоянных тем согласно Delta-AUC РИА Новости

0.578	клиントон, франция, президентский, глава, лидер, ес, правительство, хиллари, отмечать, становиться
0.579	пожар, площадь, метр, здание, квадратный, мчс, область, жилой, пожарный, возгорание
0.650	самолет, аэропорт, рейс, воздушный, посадка, авиакомпания, маршрут, находится, полет, информация
0.945	турция, вооруженный, турецкий, мирный, эрдоган, военный, огонь, анкара, действие, боевой
0.679	премьер, посольство, ответ, решение, действие, иерусалим, министр, еврейский, мид, называть
0.774	президент, сша, пресс, сообщать, результат, агентство, данные, представитель, заявлять, российский

### 5.5. Газета.Ru

В качестве базового эксперимента модель обучается на коллекции новостных потоков агентства Газета.Ru. Количество тем 200. Для оценки качества модели были добавлены те же метрики: перплексия, разреженность матриц, контрастность, чистота ядра темы. Сначала модель обучается с 15 проходами по коллекции. После добавления метрики Top Tokens, которая возвращает наиболее вероятные токены к запрашиваемым темам, происходит обучение уже с 40 проходами по коллекции. Получаем следующие результаты (рис.12). Разреженность матрицы  $\Phi$  - 93.0%, разреженность матрицы  $\Theta$  - 0.0%, средняя контрастность ядра по темам - 0.798, средняя чистота ядра - 0.640, перплексия - 2470.247.

Далее была построена модель ARTM на этой же коллекции новостных потоков. Были добавлены регуляризаторы декорреляции, разреживания и сглаживания матриц  $\Theta$  и  $\Phi$ . Добавили к модели дату публикации новостной статьи. Метрики качества рассматривались таки же, что и у классической модели PLSA. Результаты этой модели показаны на рис.13. Разреженность матрицы  $\Phi$  - 91.6%, разреженность матрицы  $\Theta$  - 14.1%, средняя контрастность ядра по темам - 0.681, средняя чистота

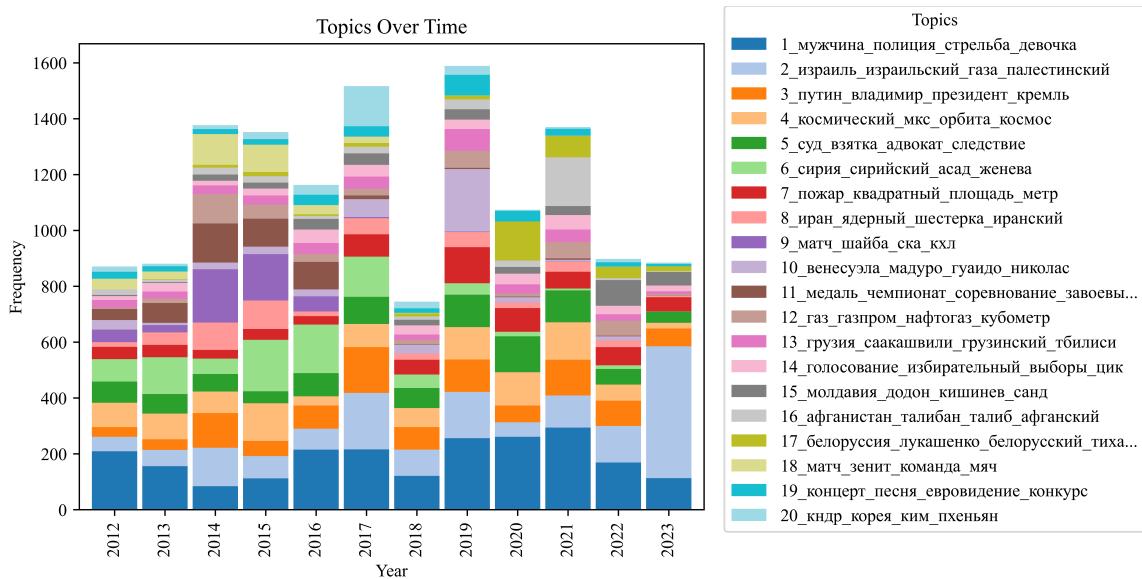
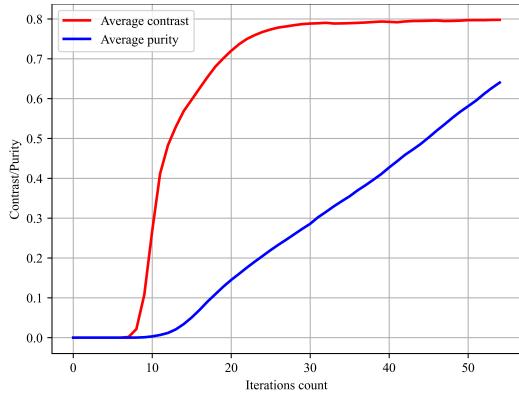


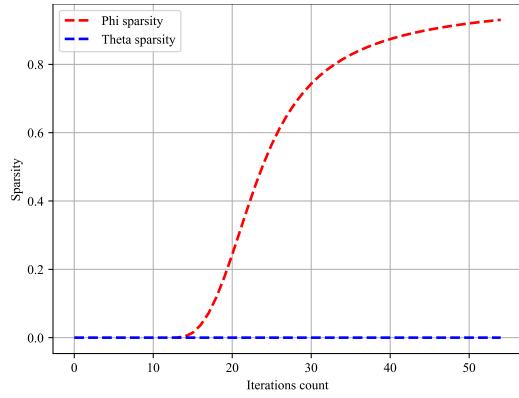
Рис. 15: Значимые темы модели BERTopic РИА Новости

ядра - 0.277. С точки зрения метрики Delta-AUC были приведены примеры постоянных тем (чем тема постояннее, тем Delta-AUC больше), табл. 5. На рис.14 показаны распределения некоторых событийных тем. На рис.15 показана матрица, отсортированная по метрике Delta-AUC.

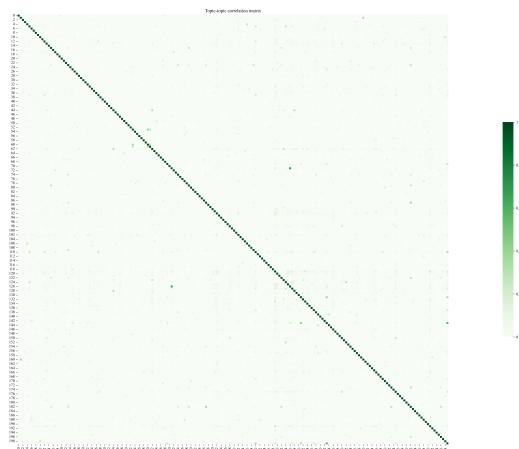
Кроме того, была построена модель BERTopic, в которой реализован подход DTM. На рис.16 изображен график изменения топ-20 самых популярных тем с течением времени



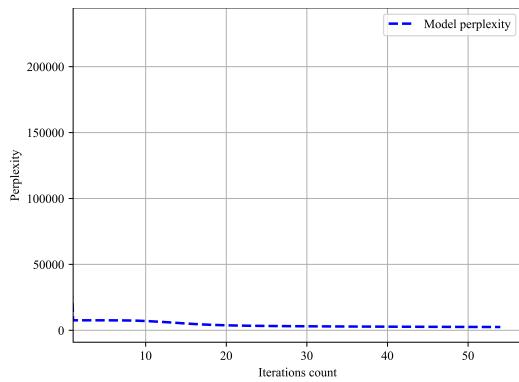
(а) График чистоты и контрастности модели



(б) График разреженности матриц  $\Theta$  и  $\Phi$

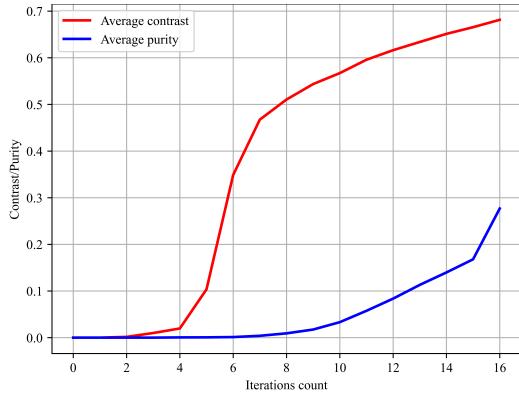


(с) Матрица корреляций между темами

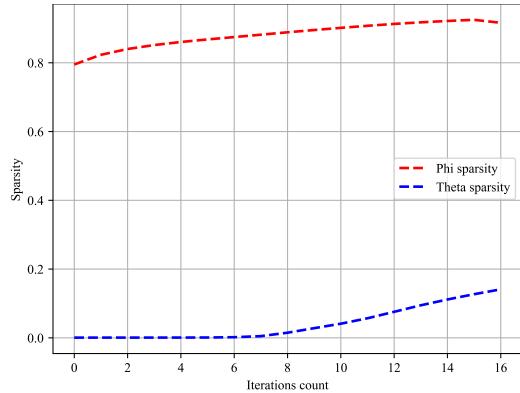


(д) График перплексии

Рис. 16: Графики для модели PLSA коллекции Газета.Ru



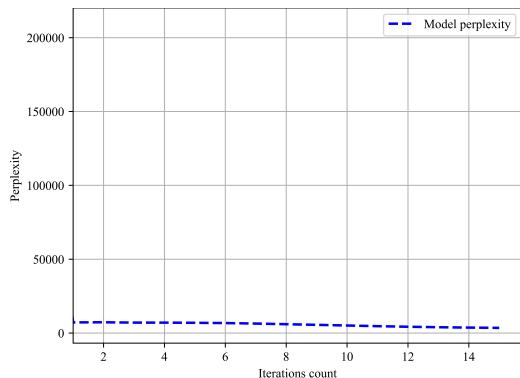
(a) График чистоты и контрастности модели



(b) График разреженности матриц  $\Theta$  и  $\Phi$



(c) Матрица корреляций между темами



(d) График перплексии

Рис. 17: Графики для модели ARTM коллекции Газета.Ru

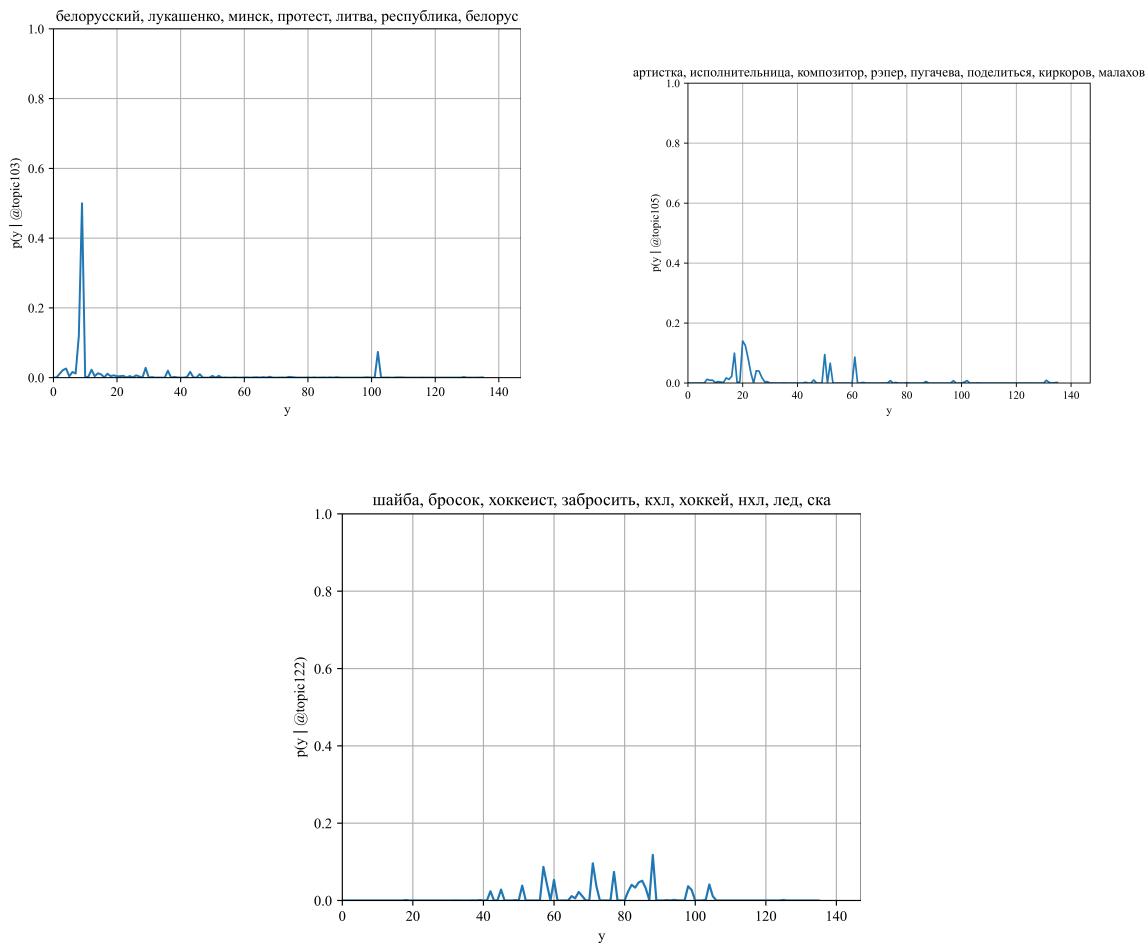


Рис. 18: Распределения некоторых событийных тем Газета.Ru

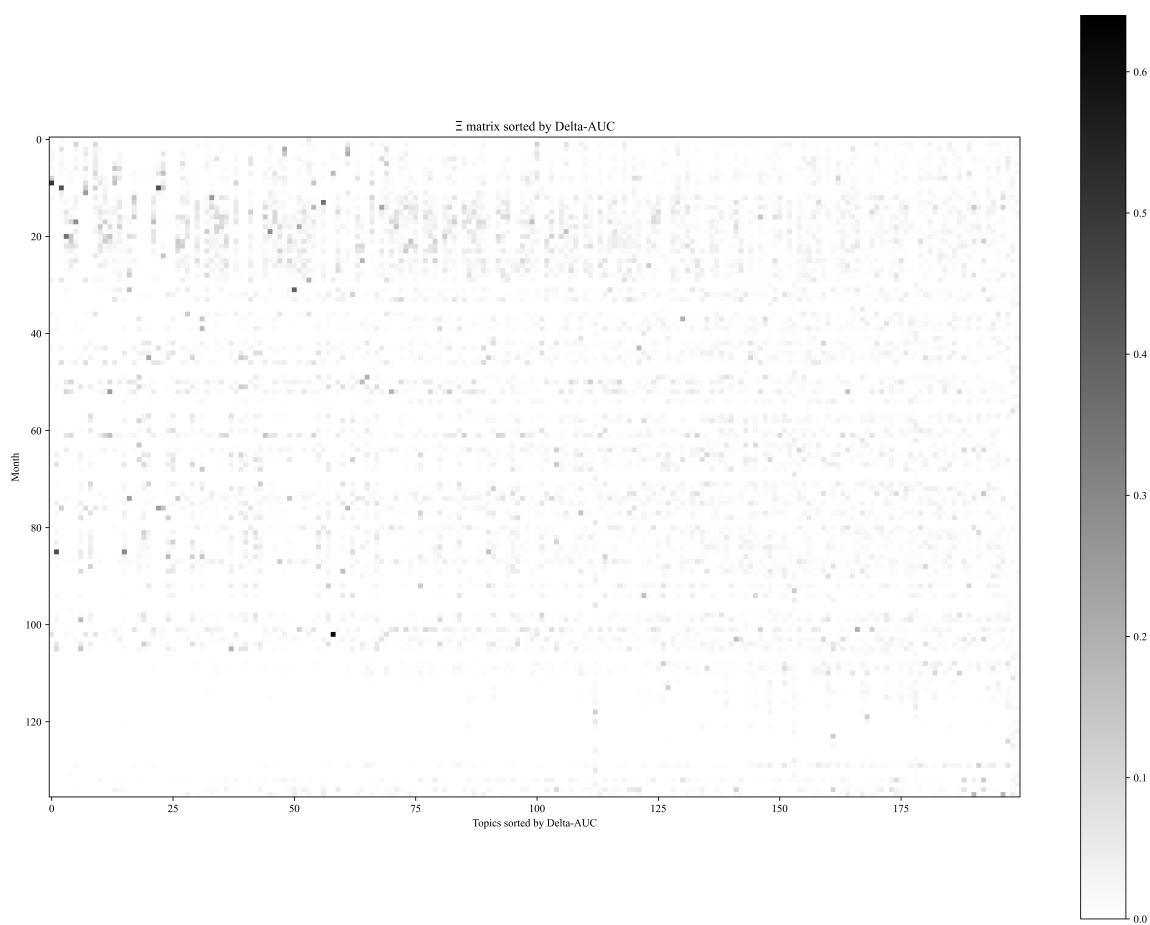


Рис. 19: Матрица  $\Xi$  Газета.Ru

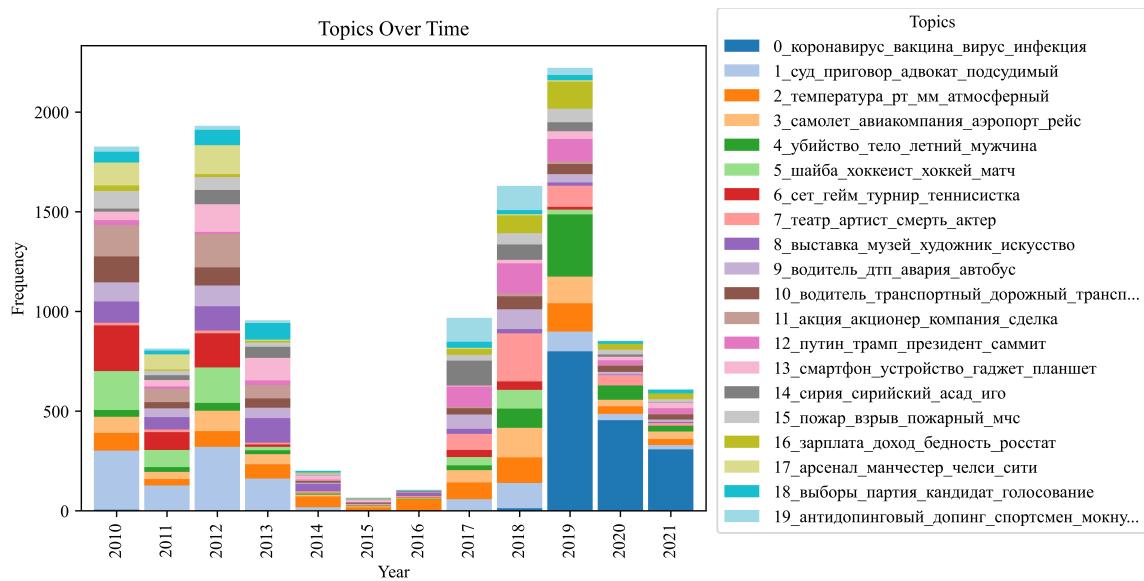


Рис. 20: Значимые темы модели BERTopic Газета.Ru

Таблица 6: Примеры выделенных тем Газета.Ru

@topic_0	врач, исследователь, лечение, болезнь, исследование, пациент, мозг, заболевание, организм, ученый, доктор, медицинский, клиника, рак, боль
@topic_1	президент, владимир, путин, глава, секретарь, кремль, песок, государство, пресс, саммит, обсуждать, послание, песков, визит, заявлять
@topic_2	ведущий, поведение, критика, скандал, повод, высказывание, отреагировать, высказываться, задавать, реакция, вызывать, скандальный, раскритиковывать, возмущаться, извиняться
@topic_3	нато, альянс, запад, присутствие, угроза, союзник, совместный, учение, усиливать, усиление, действие, агрессия, база, оборона, агрессивный
@topic_4	рынок, торговый, товар, объем, продукция, торговля, производство, производитель, экспорт, пошлина, промышленность, предприятие, отрасль, потребитель, таможенный
@topic_5	судно, море, находиться, порт, сообщать, вертолет, берег, моряк, экипаж, морской, капитан, задействовать, направляться, прибывать, член
@topic_6	глава, пост, должность, возглавлять, отставка, назначать, занимать, назначение, бывший, обязанность, кабинет, министр, исполнять, уход, заместитель

## 6. Результаты

Так как была построена составная диаграмма с помощью динамической модели BERTopic, то получились следующие результаты. Для Газета.Ru: Темы, связанные с судебной и криминальной журналистикой, темы про погодные условия, авиаперелеты, спортивные темы (а именно хоккей и теннис), темы про дорожно-транспортные происшествия, темы про культуру присутствуют на всем исследуемом временном промежутке. Эти темы относятся к классу постоянных, а есть тема про пандемию, которая

Таблица 7: Примеры постоянных тем согласно Delta-AUC Газета.Ru

0.665	ядерный, договор, кндр, пекин, кнр, испытание, ракета, дрсмд, корейский, чен, атомный, дальность, меньший, северокорейский, пхеньян
0.672	вуз, студент, учебный, обучение, образовательный, экзамен, университет, лайнер, егэ, выпускник, сдача, комиссия, минобрнаука, задание, наука
0.676	законопроект, поправка, единоросс, собрание, законодательство, думский, заседание, медведев, миронов, справедливый, палата, лдпр, эсер, комиссия, законодательный
0.727	выборы, избирательный, демократ, демократический, предвыборный, оппозиция, оппозиционный, президентский, парламентский, меркель, дебаты, фракция, коалиция, партийный, цик
0.734	боевик, аль, террористический, иго, мадуро, коалиция, оппозиция, куба, николас, колумбия, хуан, сирийский, отряд, захватить, теракт
0.756	зарплата, квартал, росстат, минэкономразвития, пенсия, минфин, бедность, прожиточный, пенсионер, индексация, темп, взнос, минимальный, плата, годовой

вспыхивает в 2019г. и уже полностью исчезает в 2021г. Такая тема называется событийная. Анализируя еще более внимательно диаграмму, можно наблюдать, как у тем изменяется частота, т.е. в какие-то года они более или менее популярны.

Для РИА Новости: политические темы, темы про космос, темы о добычи нефти, тема про взяточничество, спортивные темы про футбол и другие есть на всем временном промежутке. Хоккей и прочие спортивные соревнования скорее относятся к событийным темам, т.к. сосредоточены только в 2012-2016гг

Для Лента.ру: Темы про спорт(хоккей, фигурное катание), добычи газа, авиаперелеты, новые технологии и взяточничество, космос постоянные.

После построения темпоральной модели в ARTM появилась возможность узнать меру событийности у каждой темы. Благодаря регуляризаторам, модели ARTM те-

мы более интерпретируемы, чем в классической модели PLSA, в которой собралось много слов из фоновой лексики.

Код скачивания статей с сайта новостного агентства РИА Новости доступен на [GitHub](#). Код предварительной обработки текста, конвертирования данных в формат UCi Bag of Words, построения моделей и графиков можно найти в этом [репозитории](#).

## 7. Заключение

В ходе выполнения данной работы было проведено исследование о динамике новостных тематик с течением времени. Была поставлена цель выяснить, какие темы всегда есть в новостных сводках, а какие темы вовсе постепенно исчезают из внимания СМИ. Были построены модели PLSA, ARTM и динамическая модель BERTopic, все три обучены на коллекциях новостных статей за последние 10 лет. Для моделей были построены графики их матриц, развитие тем во времени и их метрики качества. Темы были разделены на событийные и несобытийные. Проанализировав новостные коллекции, получилось, что темы про политику, экономику, большой спорт, общественно-социальные вопросы и культуру всегда являются актуальными. Кроме того, в результате настоящей работы были рассмотрены возможности инструментов BigARTM и BERTopic.

## 8. Литература

### Список литературы

- [1] Воронцов К.В. Вероятностное тематическое моделирование: теория регуляризации ARTM и библиотека с открытым кодом BigARTM 2023. URL: <http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf>
- [2] D.M. Blei, J.D. Lafferty, Dynamic topic models, Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, 2006. URL: <http://www.cs.columbia.edu/~blei/papers/BleiLafferty2006a.pdf>

- [3] Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet Allocation // Journal of Machine Learning Research. Vol. 3 (4–5). January 2003. URL: <http://jmlr.csail.mit.edu/papers/volume3/blei03a/blei03a.pdf>
- [4] Vorontsov K.V., Voronov S.O. Automatic Filtering of Russian Scientific Content using Machine Learning and Topic Modeling // International Conference on Computational Linguistics and Intellectual Technologies «Dialogue–2015». Moscow, 2015. URL: <https://www.dialog-21.ru/media/2135/vorontsov.pdf>
- [5] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure // arXiv. 2022. URL: <https://arxiv.org/pdf/2203.05794.pdf>