«Московский физико-технический институт (национальный исследовательский университет)
Физтех-школа прикладной математики и информатики
Кафедра «Интеллектуальные системы»

Панкратов Виктор Владимирович

Вероятностное тематическое моделирование несбалансированных текстовых коллекций

03.04.01 — Прикладные математика и физика

Выпускная квалификационная работа магистра

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

Москва
2023

Аннотация

Рассматриваются текстовые коллекции, состоящие из множества документов, каждый из которых состоит из множества слов. Слова каждого документа порождаются некоторым множеством тем - вероятностных распределений на множестве слов коллекции. Задача тематического моделирования состоит в нахождении тем, которые порождают слова каждого документа коллекции. Если одна тема порождает в десятки раз большую долю слов, чем остальные, коллекция является несбалансированной. Тематические модели, построенные для решения задачи тематического моделирования, не могут корректно восстановить темы для несбалансированных коллекций. В данной работе демонстрируется проблема некорректного нахождения тем для несбалансированных коллекций и предлагается её решение с помощью добавления регуляризатора на основе близости тем как распределений.

Содержание

1 Введение

Тематическое моделирование - одно из направлений обработки текстовых коллекций. В тематическом моделировании рассматриваются коллекции, состоящие из множества документов. Каждому документ состоит из некоторого множества термов, которое обычно предполагается неупорядоченным. Задачей тематического моделирования является описание тем. Каждый документ соответствует некоторому множеству тем, а каждую тему образует некоторое множество слов. Каждый документ и каждое слово соотносятся с каждой темой с некоторой, возможно нулевой, вероятностью. Таким образом, задача тематического моделирования состоит в нахождении двух семейств дискретных условных вероятностных распределений.

Для нахождения распределений и решения задачи тематического моделирования решается задача приближенного матричного разложения. Решение задачи приближенного матричного разложения основано на максимизации правдоподобия. В процессе максимизации темы становятся близкими по мощности - числа документов, соответствующих каждой из тем, получаются близки. Если в коллекции с документами по математике и биологии мощности математической и биологической тем отличаются не более чем в 2-3 раза, решение задачи для нахождения двух тем выделит математическую и биологическую составляющие. В другом случае, если документов по математике несколько сотен, а по биологии десять, обе полученные темы окажутся математическими. Близость выделяемых тем по мощности является следствием максимизации правдоподобия. При решении оптимизационной задачи малые темы объединяются в более крупные или сливаются с уже существующими, а крупные разделяются и образуют множество схожих с исходной мелких тем.

Задача приближенного матричного разложения ставится некорректно: в общем случае она имеет бесконечное множество решений для одной и той же коллекции. Для выделения из множества решений одного, удовлетворяющего требуемым свойствам, используется регуляризация. Одним из свойств решений, которые учитываются при построении тематической модели, является интерпретируемость тем: по словам, с наибольшей вероятностью соот-

ветствующим теме, однозначно определяется смысл темы. В случае, когда коллекция состоит из одной части документов по математике и другой по биологии, интерпретируемые темы будут с наибольшей вероятностью состоять из математических и биологических терминов соответственно.

Процесс порождения документов коллекции неизвестен. По произвольной коллекции невозможно определить наличие двух тем, одна из которых значительно превышает по мощности другую. Если такие темы существуют, то большая по мощности тема будет разделяться, а мелкая сливаться с ней или другими темами. Эффект расщепления крупных тем и слияния мелких получил название проблемы несбалансированности[3]. Проблема несбалансированности приводит к ухудшению интерпретируемости тем, построенных моделью. Значительно отличающихся по мощности пары тем может и не существовать - при построении модели для несбалансированной коллекции необходимо, чтобы модель могла решить задачу и для равных по мощности тем.

Для задачи тематического моделирования существуют различные методы решения[2][5]. Одним из основных подходов является LDA[1], который использует априорное распределение Дирихле в качестве регуляризатора. Этот подход был обобщен для использования произвольных регуляризаторов в ARTM[4]. Использование регуляризаторов в ARTM позволяет адаптировать решение под конкретную задачу. Однако существование различных по мощности тем возможно для каждой задачи - это свойство обрабатываемой коллекции. Для решения проблемы несбалансированности для произвольной коллекции с помощью регуляризаторов необходимо, чтобы модель, снабженная набором регуляризаторов, восстанавливала темы как в случае сбалансированной коллекции с равными по мощности темами, так и в случае несбалансированной, с темами, отличающимися по мощности в десятки и более раз.

В данной работе экспериментально показано, что проблема несбалансированности может наблюдаться на реальных коллекциях и приводит к расщеплению крупных и слиянию мелких тем, рассмотрен регуляризатор на основе статистики семантической неоднородности, добавленный для решения

проблемы несбалансированности и проведены эксперименты, демонстрирующие интепретируемость получаемых при помощи добавленного регуляризатора тем.

2 Постановка задачи

2.1 Общая постановка задачи тематического моделирования

Пусть D - множество документов, W - множество термов. Каждый документ из D задается его длиной $n_d : \sum_{d \in D} n_d = n$ и упорядоченной последовательностью термов $\{w_i \in W\}_{i=1}^{n_d}$.Для описания вероятности появления термов в документе вводится вероятностная модель порождения коллекции. В рамках вероятностной модели документы описываются конечным множеством тем T: вероятность появления слова в документе связывают с условными вероятностями появления тем в документе и вероятностями появления слова для каждой пары темы и документа. При этом предполагается выполнение следующих гипотез:

- Гипотеза мешка слов: для каждого документа d представление термов в виде последовательности $\{w_i \in W\}_{i=1}^{n_d}$ эквивалентно представлению термов в виде неупорядоченного множества $\cup_{i=1}^{n_d} w_i$, в котором каждый терм w встречается n_{dw} раз. Множество тем не зависят от порядка термов в документе и порядка документов в коллекции

- Гипотеза условной независимости: вероятность появления терма w в документе d по теме t описывается распределением

$$p(w|d,t) = p(w|t)$$

Согласно последнему предположению, вероятности появления термов не зависят от документа. Они описываются вероятностями появления слов для каждой темы $p(w|t) = \varphi_{wt}$ и вероятностями порождения слов документов из

каждой темы $p(t|d) = \theta_{td}$. Задача тематического моделирования состоит в нахождении распределений $\varphi_{wt}, \theta_{td}$. Для нахождения $\varphi_{wt}, \theta_{td}$ ставится задача приближенного матричного разложения:

$$F = \Phi\Theta \tag{2.1}$$

$$F = \left(\frac{n_{wd}}{n_d}\right)_{W\times D} \quad \Phi = (\varphi_{wt})_{W\times T} \qquad \Theta = (\theta_{td})_{T\times D}$$

Для решения (2.1) максимизируется следующая функции правдоподобия:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} \to max_{\Phi,\Theta} \tag{2.2}$$

Задача приближенного матричного разложения поставлена некорректно: множество ее решений в общем случае бесконечно. Чтобы наложить дополнительное ограничение на множество решений задачи, в оптимизируемую функцию (2.2) добавляют несколько слагаемых - регуляризаторов, зависящих от матриц Φ, Θ. Набор регуляризаторов зависит от задачи - различные регуляризаторы наделяют решение различными свойствами. Функция правдоподобия при этом принимает следующий вид:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + \sum_i \tau_i R_i(\Phi, \Theta) \to max_{\Phi,\Theta} \tag{2.3}$$

2.2 Проблема несбалансированности

Результатом работы тематической модели являются две матрицы: матрица Φ, показывающая вероятность появления конкретного слова при заданной теме, и матрица Θ, показывающая распределение тем между документами. Из матрицы Θ часто следует, что темы близки по мощности: при вероятностях тем, определенных как $p(t) = \sum_{d\in D} p(t|d)n_d$ выполнено $\forall t_1, t_2 \to \frac{p(t_1)}{p(t_2)} < C$ для C не превосходящих 10. При этом в обработанной коллекции мощности тем произвольные и могут существовать такие t_1, t_2, что отношение $\frac{p(t_1)}{p(t_2)}$ больше любой наперед заданной константы. Такой эффект является свойством максимизации правдоподобия: модели выгодно исполь-

зовать все свои параметры. Сокращение доли отдельной темы приводит к неполному использованию или, в пределе, уменьшению числа параметров. Чтобы выделять не равные по мощности темы в случае несбалансированной коллекции, в модель предлагается добавить регуляризатор.

Для построения тематической модели была использована гипотеза условной независимости, которая формулируется следующим образом: $p(w, d|t) = p(w|t)p(d|t)$. Гипотеза условной независимости проверяется для темы t статистикой семантической неоднородности темы:

$$S_t = KL(\hat{p}(w, d|t)||p(w|t)p(d|t))$$

Здесь и далее $\hat{p}$ обозначает частотные оценки вероятностей

$$\hat{p}(w, d|t) = \frac{n_{dw} p(t|d, w)}{n \cdot p(t)},\ \hat{p}(w|d) = \frac{n_{dw}}{n_d}$$

Статистика семантической неоднородности также записывается в виде

$$S_t = \sum_{d \in D} \sum_{w \in d} \hat{p}(w, d|t) \ln \frac{\hat{p}(w, d|t)}{p(w|t)p(d|t)} \tag{2.4}$$

При выполнении гипотезы условной независимости значение статистики семантической неоднородности мало, а распределение $p(w|d, t)$ близко к $p(w|t)$. Каждая тема представляется в виде кластера в пространстве размерности $|W|$, центром которого является $p(w|t)$. Статистика семантической неоднородности показывает удаленность $p(w|d, t)$ от центра кластера. Чтобы учитывать значение статистики семантической неоднородности, в модель добавляется регуляризатор. Статистика семантической неоднородности темы суммируется по всем темам:

$$\sum_{t \in T} S_t = \sum_{d \in D} \sum_{w \in d} \left(\sum_{t \in T} \hat{p}(w, d|t) \ln \frac{\hat{p}(w, d|t)}{p(w|t)p(d|t)} \right) \to \min_{\Phi, \Theta} \tag{2.5}$$

Используется формула $\hat{p}(w, d|t) = \frac{p(t|d,w)\hat{p}(w|d)p(d)}{p(t)}$ для преобразования логарифма. Таким образом, при домножении на постоянное для конкретной за-

дачи n формула (2.5) преобразуется и вставляется в исходную постановку задачи в качестве регуляризатора R:

$$\sum_{d\in D}\sum_{w\in d}\left(\sum_{t\in T}\frac{p(t|d,w)\hat{p}(w|d)p(d)}{p(t)}\ln\frac{p(t|d,w)\hat{p}(w|d)p(d)}{p(w|t)p(d|t)p(t)}\right)\to\min_{\Phi,\Theta} \tag{2.6}$$

$$\beta_{dw}=\sum_{t\in T}\frac{p(t|d,w)}{p(t)},\ n_{dw}\sim\hat{p}(w|d)p(d),\ p(t|d)=\frac{p(d|t)p(t)}{p(d)}$$

$$\frac{p(t|d,w)\hat{p}(w|d)p(d)}{p(w|t)p(d|t)p(t)}=\frac{p(t|d,w)\hat{p}(w|d)}{p(w|t)p(t|d)}=\frac{\hat{p}(w|d)}{p(w|d)}$$

$$\sum_{d\in D}\sum_{w\in d}\left(\sum_{t\in T}\frac{p(t|d,w)n_{dw}}{p(t)}\ln\frac{\hat{p}(w|d)}{p(w|d)}\right)\to\min_{\Phi,\Theta} \tag{2.7}$$

$$\sum_{d\in D}\sum_{w\in d}\beta_{dw}n_{dw}\ln\frac{p(w|d)}{\hat{p}(w|d)}\to\max_{\Phi,\Theta} \tag{2.8}$$

$$R=\sum_{d\in D}\sum_{w\in d}\beta_{dw}n_{dw}\ln\sum_{t\in T}\varphi_{wt}\theta_{td} \tag{2.9}$$

Выражение (2.9) является в точности выражением (2.2), домноженным на весовые множители β_{dw}. По смыслу эти множители увеличивают веса маломощных тем $p(t) << 1$

Цель данной работы - предложить решение проблемы несбалансированности, состоящее в добавлении регуляризатора семантической неоднородности, и экспериментально проверить предложенное решение для коллекций с различным балансом тем и для моделей с разными параметрами.

3 Вычислительный эксперимент

3.1 Генерация коллекций

Эксперименты проводились на синтетических и реальных коллекциях. Для получения синтетических коллекций генерируются матрицы Φ, Θ - искомые матрицы Столбцы матриц Φ, Θ порождаются симметричными распре-

делениями Дирихле. Параметр распределения определяется из соображений реалистичности коллекции и берется малым, чтобы получаемые матрицы были разреженными. Для матрицы Φ он берется равным 0.02, для матрицы Θ равным 0.2. Чтобы регулировать баланс тем, на этом этапе генерации наибольшие значения в столбцах Θ меняются со значениями в строках, которые соответствуют необходимым темам. После этого в обе матрицы добавляется еще одна фоновая тема, доля которой во всех документах равна 0.5, если не указано иного. Фоновая тема порождается несимметричным распределением Ципфа. Матрица Θ перед этим перенормируется в зависимости от желаемой доли фоновой темы в документах.

Для генерации очередного слова w_i сначала генерируется тема t_i документа из соответствующего этому документу столбцу матрицы Θ. Затем слово генерируется из столбца Φ, соответствующего теме t_i. Таким образом, процесс генерации документов описывается как

$$t_i \sim Dir(t|d) \quad w_i \sim Dir(w|t_i), i \in 1 \dots n_d$$

Реальные коллекции были получены предобработкой следующих датасетов: 20newsgroups (в дальнейшем обозначается 20news), записей TED Talks(TED), описаний книг(Books) и новостей BBC(BBC). Основные параметры используемых датасетов приведены в таблице 1:

	20news	TED	Books	BBC
Число документов	18846	3865	3000	1926
Число слов, среднее	252.89	405.88	187.26	141.98
Число слов, медиана	101	407	188.0	130
Число уникальных слов	42616	26902	19608	10924

Таблица 1: параметры реальных коллекций, используемых в экспериментах

Датасеты были предобработаны: проведена лемматизация и стемминг, удалены стоп-слова и слова, встречающиеся менее чем в двух документах. На основе полученных датасетов генерировались используемые в экспериментах коллекции. Для генерации несбалансированной коллекции использо-

валась комбинация одного из следующих подходов:

1. Выбирается подвыборка документов исходной коллекции. Для исходной коллекции отношения чисел документов, соответствующих двум темам не превосходило 10. Для таких коллекций тематическая модель без регуляризаторов способна корректно соотнести большинство документов с темами так, чтобы малые темы не объединялись, а крупные не разбивались на несколько более мелких. Поэтому для получения несбалансированной коллекции из исходной удаляются документы, соответствующие более мелким темам. Число удаленных документов зависит от требуемого баланса тем в коллекции.

2. Отбираются монотематические документы. Для эта исходной коллекции строится тематическая модель без регуляризаторов. Для каждого документа рассматриваются две темы, порождающие его с наибольшими вероятностями. Если отношение вероятностей рассматриваемых тем больше 2, документ считается монотематическим.

3. Генерируются новые документы. Для этого выбирается число слов генерируемого документа и его тема. Пока в документе не сгенерировано достаточное количество слов, выбираются случайные n_{group} подряд идущих слов из случайного документа выбранной темы. Здесь n_{group} - параметр алгоритма. Эксперименты проводились при $n_{group} = 10$, если не указано иное

Определим понятие степени несбалансированности. Для синтетической коллекции известна вероятность появления темы в документе для каждой пары темы и документа. Это следует из процесса генерации коллекции на основе матрицы Θ - искомой вероятностью будет элемент соответствующей строки и соответствующего столбца матрицы. Будем генерировать документы с равным числом слов. Тогда мощностью темы $p(t)$ назовем сумму вероятностей её появления по документам - сумму соответствующего столбца Θ.

$$p(t) = \sum_d p(t|d)p(d) \sim \sum_d p(t|d)$$

Степень несбалансированности в таком случае определяется как

$$R_{imb} = \frac{\max\limits_{t \in T}|t|}{\min\limits_{t \in T}|t|} \tag{3.1}$$

Для получения матрицы Θ для реальной коллекции строится тематическая модель и отбираются монотематические документы. Тогда матрица Θ составляется как матрица размера $T \times D$ в столбце i и строке j которой стоит 1, если для документа d_j с наиболее вероятна тема i и 0 в противном случае. Тогда определение степени несбалансированности через равенство (3.1) применимо и к реальным или сгенерированным на их основе коллекциям.

3.2 Сравнение моделей

Темы документов не заданы изначально. В случае, если документы коллекции изначально разделены на категории, эти категории могут не совпадать с реальными темами. Поэтому необходимо определить способ оценки качества восстановления тем моделью. Для этого матрицы Φ, Θ, полученные в ходе эксперимента, сравниваются с аналогичными матрицами, полученными другой тематической моделью. Для корректности сравнения другая тематическая модель строится на сбалансированной коллекции, содержащей все документы исследуемой несбалансированной коллекции как подмножество. Строки и столбцы матриц, найденных для сбалансированной коллекции, соответствующие документам несбалансированной коллекции, сравниваются со строками и столбцами матриц Φ, Θ, полученными в ходе эксперимента. Темы, найденные моделью, построенной на сбалансированной коллекции, будем называть исходными темами.

Вопрос оценки качества восстановления тем сведен к сравнению матриц Φ_1, Θ_1, полученных одной моделью и Φ_2, Θ_2, полученных второй моделью. Необходимо установить соответствие между строками матриц Θ_1, Θ_2, так как полученные моделью темы неупорядоченные. Строки, соответствующие одной теме в общем случае имеют различные номера в Θ_1, Θ_2. Строки $\Theta_1[:, i]$,$\Theta_2[:, j]$ являются взаимно близкими, если выполнен следующий крите-

рий:

$$\arg\min_k \ (dist(\Theta_1[:,i], \Theta_2[:,k])) = j \tag{3.2}$$

$$\arg\min_k \ (dist(\Theta_1[:,k], \Theta_2[:,j])) = i \tag{3.3}$$

Здесь $dist$ - расстояние по заданной метрике. В экспериментах используется косинусное расстояние. Для столбцов матриц Φ_1, Φ_2 свойство взаимной близости и последующие определения вводятся аналогично.

$$\arg\min_k \ (dist(\Phi_1[i], \Phi_2[k])) = j \tag{3.4}$$

$$\arg\min_k \ (dist(\Phi_1[k], \Phi_2[j])) = i \tag{3.5}$$

Рассмотрим определённую в (3.2), (3.3) взаимную близость тем. Идеальный результат достигается, когда для каждой темы $\Theta_1[:,i]$ существует взаимно близкая ей тема $\Theta_2[:,j]$). Существует другой случай: темы $\Theta_1[:,j]$, которые не являются ближайшими ни для какой темы из Θ_2, то есть для любого i для пары $\Theta_1[:,i], \Theta_2[:,j]$ не выполнено (3.2). Таким образом определяются невосстановленные темы. Аналогично определяются ложные темы: такие темы $\Theta_1[:,i]$, что для любого j (3.3) не выполнено для пары $\Theta_1[:,i], \Theta_2[:,j]$.

В экспериментах взаимная близость, а также взаимно близкие, невосстановленные и ложные темы для реальных коллекций определяются по сравнению строк матриц Θ, полученных описанной в эксперименте моделью и моделью, построенной на сбалансированной коллекции, если не указано иное. Для синтетических коллекций матрица Θ формируется в процессе генерации, поэтому полученная в эксперименте матрица сравнивается с известной.

3.3 Демонстрация проблемы несбалансированности

На примере синтетических коллекций проверим, что при возрастании степени несбалансированности число взаимно близких тем уменьшается. Сгенерированы коллекции с одной большой по мощности темой и множеством малых равномощных. Общее число документов в коллекции равно 2000 для каждой сгенерированной коллекции, число тем равно 100. Результат обработ-

ки сгенерированных коллекций тематической моделью без регуляризаторов представлен на графике 1

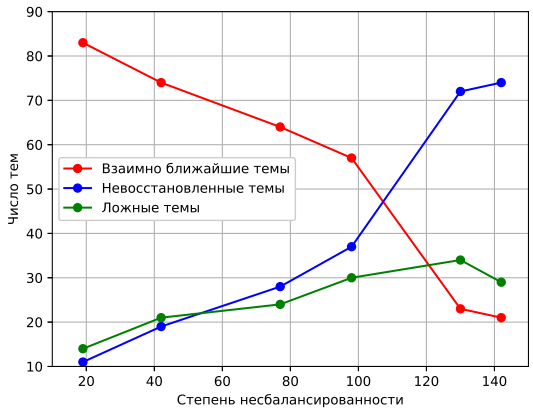

График 1: Зависимость числа восстановленных взаимно близких тем от степени несбалансированности коллекции, синтетические коллекции

При малых степенях несбалансированности число взаимно близких тем, найденных моделью близко к 90. Оно уменьшается с возрастанием степени несбалансированности. Для степеней несбалансированности больше 90 большинство тем оказались не взаимно близки с исходными, так как увеличилось число невосстановленных и ложных тем.

Проверим, что проблема несбалансированности наблюдается и для реальных коллекций. Проведен эксперимент для коллекций, сгенерированной на основе 20newsgroups. Генерировались коллекции с одной темой, состоящей из 3000 документов и девятнадцати темами, состоящими из $\frac{3000}{x}$ документов, где x - зависящая от эксперимента степень несбалансированности. На графике 2 представлены результаты для модели без регуляризатора семантической неоднородности и для модели с добавлением регуляризатора с коэффициентом 0.6. Результаты усреднялись для 50 итераций генерации коллекции и построения модели.

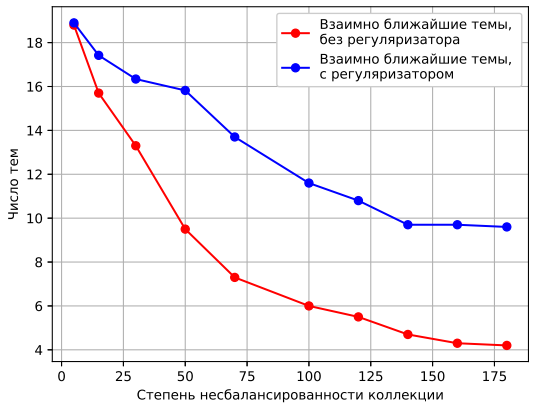

График 2: Зависимость числа восстановленных взаимно близких тем от степени несбалансированности коллекции

Число взаимно близких тем убывает с ростом степени несбалансированности для обеих моделей, однако для модели с добавлением регуляризатора больше половины найденных тем оказались взаимно близки с исходными при значении степени несбалансированности 100. Для той же степени несбалансированности модель без регуляризатора смогла восстановить 6 из 20 тем. Таким образом, модель с добавлением регуляризатора лучше восстановила исходные темы для коллекции с одной большой по мощности темой. Рассмотрим коллекции иного вида баланса тем.

Пусть темы делятся на равномощные крупные и равномощные мелкие и число крупных тем больше одной. На основе 20newsgroups сгенерирована коллекция из 4 тем по 500 монотематических документов и 16 тем по 10 монотематических документов. Таким образом, степень несбалансированности полученной коллекции равна 50. Мотивация для выбора описанного баланса тем в коллекции представлена в секции 3.4.

В модель добавлен регуляризатор семантической неоднородности. Для различных значений коэффициента регуляризации был проведен эксперимент и подсчитано полученное число взаимно близких тем. Генерация коллекции и проведение эксперимента были повторены 50 раз, а значения эксперимента усреднены. На графике 3 представлена зависимость числа найденных моделью взаимно близких тем от коэффициента регуляризатора.

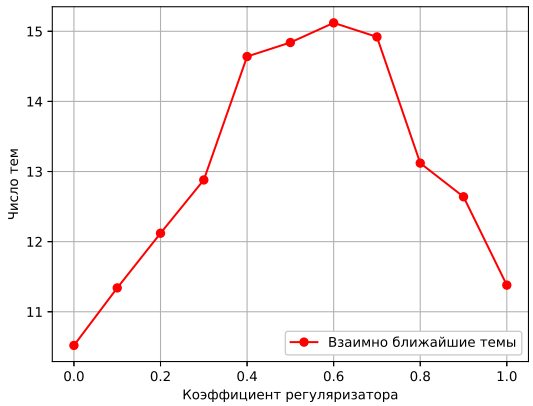

График 3: Эксперимент на коллекции с 4 равными по мощности крупными темами. Зависимость между числом взаимно близких тем и коэффициентом регуляризации

При отсутствии регуляризации модель выделила 10 взаимно близких с исходными тем. В зависимости от коэффициента регуляризации модель с регуляризатором выделяла больше взаимно ближайших тем при максимуме для коэффициента 0.6, что соответствует использованному в предыдущем эксперименте значению.

Пусть 20 тем в коллекции распределены равномерно: сгенерированы коллекции, для которых мощность темы i равна $10 \cdot (ki + 1)$, $k \in \{1, 2, 5\}$. При заданных таким образом мощностях тем степени несбалансированности коллекции принимают значения от 20 до 100. Проведен эксперимент для тематической модели с добавлением регуляризатора семантической неоднородности. Зависимость числа найденных взаимно близких тем от коэффициента регуляризации для таких коллекций представлена на графике 4.

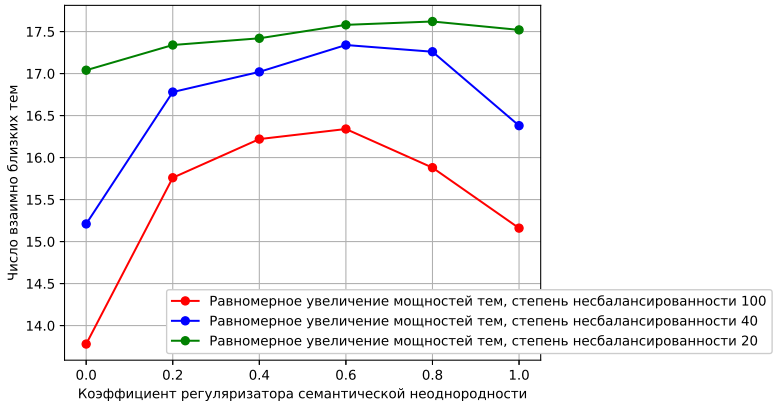

График 4: Эксперимент на коллекциях с равномерно возрастающими мощностями тем. Зависимость между числом взаимно близких тем и коэффициентом регуляризации

Для коллекции со степенью несбалансированности 20 и тем с мощностью большей 20 отношение мощностей тем не превосходит $\frac{200}{30} < 7$. Таким образом, большая часть коллекции сбалансированная и как модель без регуляризатора, так и модель с добавлением регуляризатора восстановили большую часть исходных тем. В то же время для коллекции со степенями несбалансированности 40 и 100 при добавлении регуляризатора число взаимно полученных близких к исходным тем увеличилось.

Таким образом, проблема несбалансированности наблюдается как для реальных так и для синтетических коллекций и добавление регуляризатора приводит к выявлению большего числа взаимно близких с исходными тем.

3.4 Подбор параметров генерации коллекции

Проверим, как меняется качество восстановления тем в зависимости от состава коллекции. Для этого сгенерированы синтетические коллекции с различным числом крупных тем и различными степенями несбалансированности R_{imb}. Сгенерированы коллекции с несколькими крупными темами по $R_{imb} \cdot 5$ документов и оставшимися мелкими темами по 5 документов. Выбор числа документов по каждой теме обусловлен результатами, показанными на графике 1: необходимо проверить, что причиной уменьшения числа найденных моделью взаимно близких тем не являлась слишком малая мощность для каж-

дой мелкой темы. Степеням несбалансированности больше 100 соответствуют мощности мелких тем меньше 10.

Эксперимент был проведен при $R_{imb} \in [30, 50, 70]$ Для каждого эксперимента и каждого числа крупных тем было сгенерировано по 50 коллекций. Общее число тем для каждой коллекции равно 100. Для каждого набора параметров было взято среднее число взаимно близких тем по всем запускам на соответствующих сгенерированным коллекциям. Зависимость среднего числа взаимно близких тем от числа крупных тем при генерации представлена на графике 5

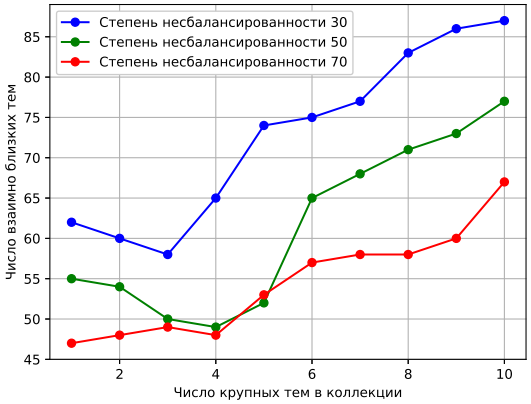

График 5: Зависимость числа взаимно близких тем от числа крупных тем в коллекции

При увеличении числа крупных тем число найденных взаимно близких тем сначала падает, но потом возрастает и становится больше, чем для одной крупной темы. В дальнейшем будут генерироваться коллекции с 4 крупными темами. Из графика 3 следует, что для такого значения число восстановленных моделью без регуляризатора семантической неоднородности низкое. В свою очередь, из графика 5 следует, что увеличение числа крупных тем приводит к лучшему восстановлению тем коллекции. Исходя из этого, в дальнейшем на основе 20newsgroups будут генерироваться коллекции с 4 крупными и 16 мелкими темами, если не указано иное. При этом мощность крупных тем будет равна 500, а мощность мелких 10. Степень несбалансированности для таких коллекций будет равна 50.

Проверим, что для выбранного способа генерации реальной коллекции

для любого рассматриваемого параметра n_{group} число взаимно близких тем не меньше, чем при обработке подмножества документов реальной коллекции с тем же балансом тем.

На основе коллекции 20newsgroups сгенерировано десять коллекций со степенью несбалансированности $\tau = 50$. Каждая из этих коллекций сгенерирована на основе монотематических документов 20newsgroups, при этом для генерации коллекции с номером i использовался параметр $n_{group} = i$. Для всех построенных коллекций проведен эксперимент с добавлением регуляризатора семантической неоднородности и без него. Коэффициент регуляризации был выбран как оптимальный из предыдущей секции. На графике 5 результат обработки таких коллекций сравнивается с результатом обработки коллекций с тем же балансом тем, полученных выбором подмножества монотематических документов 20newsgroups.

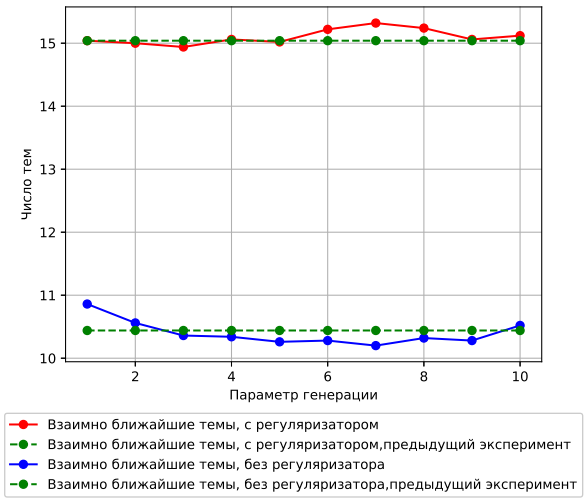

График 6: Эксперимент по подбору параметра n_{group} генерации коллекции

Число восстановленных тем для сгенерированных коллекций незначительно отличается от числа восстановленных тем для реальных. Исходя из этого, такой способ генерации используется в дальнейшем при построении коллекций. При этом, поскольку на графике отсутствует явная зависимость числа взаимно близких тем от параметра n_{group}, его можно выбирать любым от 1 до 10. Значения $n_{group} > 10$ в экспериментах не рассматривались.

3.5 Добавление регуляризатора разреживания

Подход ARTM позволяет использовать несколько регуляризаторов при построении модели. В предыдущих экспериментах эта возможность не была использована: эксперименты проводились для модели с добавлением регуляризатора семантической неоднородности и только его и для модели без регуляризаторов. Рассмотрим, как можно улучшить предыдущие результаты при добавлении других регуляризаторов. В модель был добавлен регуляризатор разреживания: каждые 10 итераций обнуляются 0.1 от всех элементов каждого столбца Φ и каждой строки Θ. Не обнуляются элементы больше 1/10000.

На графике 7 представлена зависимость числа взаимно близких тем от коэффициента регуляризатора семантической неоднородности для 50 итераций генерации коллекции и построения тем. Для сравнения на графике также представлены результаты аналогичного эксперимента из секции 3.3 без добавления регуляризатора разреживания.

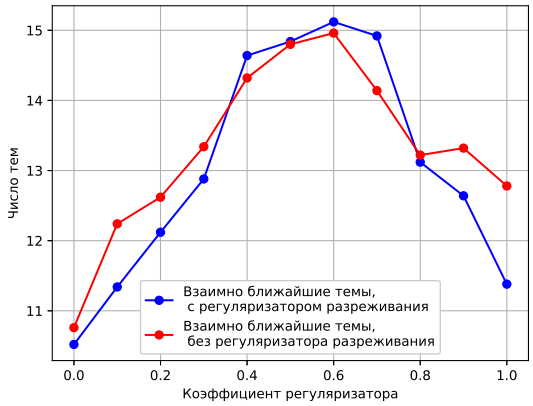

График 7: Зависимость числа взаимно близких тем от степени несбалансированности коллекции при добавлении регуляризатора разреживания

Добавление регуляризатора разреживания не ухудшило, но и не улучшило результаты модели. Для этого же эксперимента на графике 8 приведено стандартное отклонение результатов

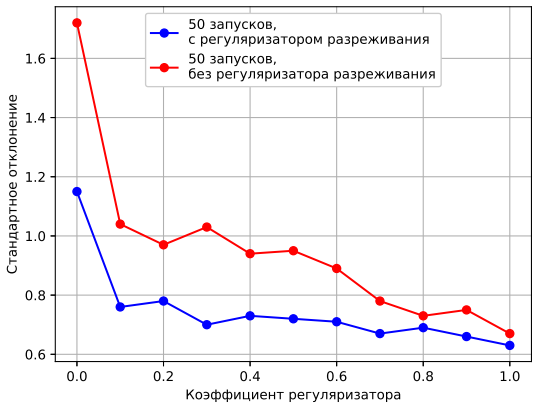

График 8: Стандартное отклонение числа взаимно близких тем для 50 различных запусков модели в зависимости от коэффициента регуляризации

Стандартное отклонение уменьшается с ростом коэффициента регуляризатора семантической неоднородности. Также оно становится меньше при добавлении регуляризатора разреживания. Этот эффект проверен на другой коллекции. Эксперимент проводился на коллекции новостей BBC. Строились тематические модели с добавлением регуляризатора разреживания и без него. Регуляризатор семантической неоднородности добавлен в обе модели с коэффициентом 0.6. Модели запускались 50 раз для различной начальной инициализации. Стандартное отклонение числа взаимно близких тем в зависимости от количества тем модели приведено на графике 9.

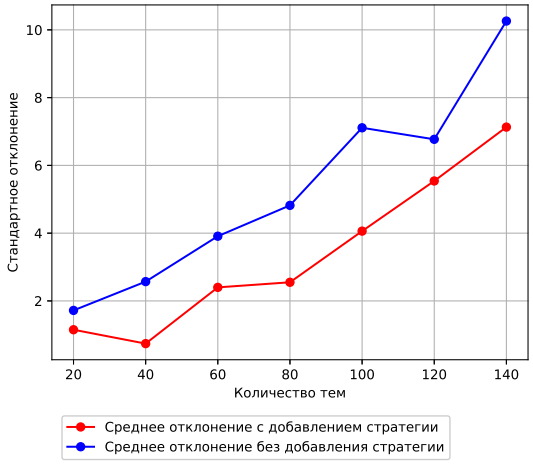

График 9: Стандартное отклонение числа взаимно близких тем для 50 различных запусков модели в зависимости от числа тем коллекции новостей BBC

Стандартное отклонение для модели с добавлением регуляризатора меньше, чем для модели без регуляризатора, но оно растет в зависимости от числа тем. Следующий эксперимент демонстрирует, что это связано со свойствами используемой коллекции. Для всех используемых датасетов построена тематическую модель для каждого числа тем из $[20, 40, 60, 80, 100]$. Регуляризатор семантической неоднородности в модели не добавлялся. Модели запускались 50 раз при различной начальной инициализации. Среднее число взаимно близких тем в зависимости от числа тем приведен на графике 10

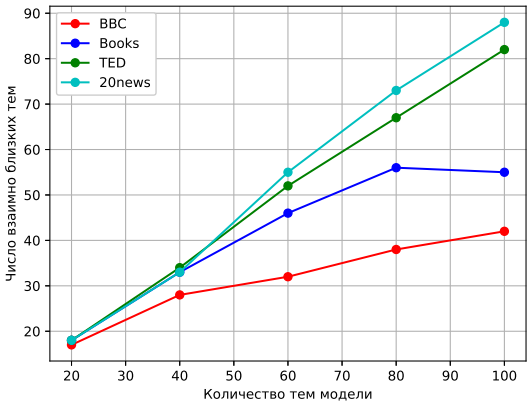

График 10: Среднее число взаимно близких тем между повторными запусками одной модели на одном датасете

В случае, когда число тем превышает реальное, модель не может корректно восстановить темы. Для коллекций описаний книг и новостей BBC модель с регуляризатором не восстанавливает исходные темы при числе тем больше 40. Таким образом, в дальнейших экспериментах для демонстрации обработки моделью различных коллекций используется 40 тем. Для числа тем, не превосходящего это значение, регуляризатор разреживания уменьшает стандартное отклонение получаемых результатов и добавлен во все последующие эксперименты.

3.6 Добавление регуляризатора декоррелирования

Проверим, как изменяется качество восстановления тем при добавлении регуляризатора декоррелирования. Сгенерирована коллекция со степенью несбалансированности 50 на основе 20newsgroups описанным выше способом. Исследовалась зависимость числа найденных моделью взаимно ближайших тем от коэффициента регуляризатора декоррелирования при постоянном коэффициенте регуляризатора семантической неоднородности равном 0.6. Результаты приведены на графике 11.

График 11: Зависимость числа взаимно близких тем от десятичного логарифма коэффициента регуляризатора декоррелирования

Как видно из графика, добавление регуляризатора декоррелирования при подборе коэффициента регуляризации повышает качество восстановления тем.

Так как регуляризатор декоррелирования улучшил результат для коэффициента регуляризатора семантической неоднородности 0.6, проверим, останется ли значение 0.6 оптимальным при его добавлении. Для этого из предыдущего графика подобрано наилучшее значение коэффициента регуляризатора декоррелирования. Повторим эксперимент по подбору коэффициента регуляризатора семантической неоднородности с добавлением регуляризатора декоррелирования в модель. На графике 12 результаты сравниваются с результатами модели с добавлением регуляризатора разреживания без регу-

ляризатора декоррелирования.

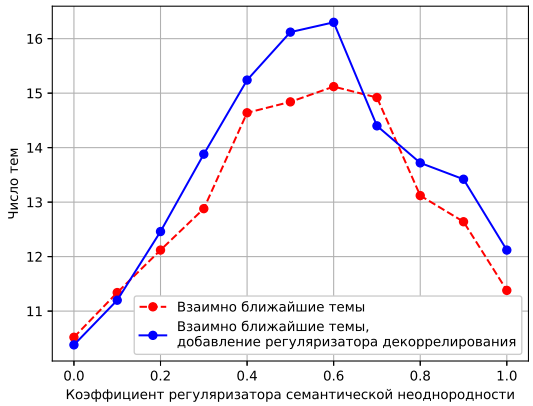

График 12: Зависимость числа взаимно близких тем от коэффициента регуляризатора семантической неоднородности для модели с добавлением декоррелирования

Для обеих моделей максимум достигается при значении 0.6 коэффициента регуляризатора семантической неоднородности. Число взаимно близких тем в максимуме увеличилось и стало равно 16, а не 15.

3.7 Сравнение результатов на 20newsgroups

Покажем, как эффекты слияния и расщепления тем мешают интерпретации результатов. Для коллекции 20newsgroups построена тематическая модель на 40 темах. Полученные темы упорядочивались по числу соответствующих им документов. Для 20 тем, которым соответствуют меньше всего документов в коллекции были оставлены первые 10 документов соответствующих тем. Документы, соответствующие остальным темам, из коллекции не удалялись. Степень несбалансированности полученной коллекции равна 80.6

На полученной коллекции строятся две модели. Первая - с добавлением регуляризатора декоррелирования. Вторая - с добавлением регуляризатора семантической неоднородности, регуляризатора разреживания и регуляризатора декоррелирования. Коэффициенты регуляризаторов выбирались как указанные в предыдущем эксперименте. Для регуляризатора семантиче-

ской неоднородности коэффициент равен 0.6. В таблице 2 приведены наиболее вероятные слова для близких тем между исходной моделью и моделью без регуляризатора.

	Исходная модель	Модель без регуляризатора
1	size frame pixel quality image	user graphic support software animation
2	size frame pixel quality image	display image library animation screen
3	spirit soul eternal life heaven	death sin church doctrine god
4	spirit soul eternal life heaven	death faith heaven sin believe
5	void include null int code	application window use get problem code
6	menu mouse button application window	application user window server program
7	firearm legal citizen law crime	wrong society person life crime
8	wrong society orientation partner relationship	wrong society person life crime

Таблица 2: Сравнение ближайших тем исходной модели и модели без регуляризатора. Коллекция 20news

Для первой модели в строках 1-2 представлены слова одной и той же темы. Наиболее близкими для неё являются 2 похожих по смыслу темы модели без регуляризатора. Такой эффект соответствует расщеплению темы - когда одна крупная тема разделяется на 2 более маленьких по мощности. Аналогичный пример приведен в строках 3-4. В то же время мелкая тема в строке 7, у которой было оставлено только 10 документов, слилась с темой из строки 8: они наиболее близки к одной теме модели без регуляризатора. В эксперименте также представлен другой случай слияния, когда 2 различные темы сливались сами с собой и с другими, как в случае исходных тем из строк 5-6. Для них существуют соответствующие им наиболее, но не взаимно, близкие темы. Соответствующие им темы не интерпретируемые.

Темы, построенные моделью без регуляризатора, не соответствуют исходным. Проверим, что модель с добавлением регуляризатора способна восстановить исходные темы. Для этого аналогичное предыдущему сравнение представлено в таблице 3

Исходная модель	Модель с регуляризатором
size frame pixel quality image	animation interactive graphic draw image
spirit soul eternal life heaven	soul holy god divine revelation
void include null int code	function include void return null
menu mouse button application window	button menu user command server
firearm legal citizen law crime	murder person crime moral society
wrong society orientation partner relationship	orientation wife partner relationship marriage

Таблица 3: Сравнение тем исходной модели и ближайших для них тем модели с регуляризатором. Коллекция 20news

Темы, которые восстановила модель с добавлением регуляризатора близки по смыслу к исходным. Поэтому в приведенном эксперименте указанное несоответствие тем не является следствием алгоритма построения коллекции. Таким образом, модель с добавлением регуляризатора помогла устранить эффекты слияния и расщепления тем.

Предыдущий эксперимент был проведен также для коллекций TED, BBC, Books. Результаты представлены в таблицах 4-9

Исходная модель	Модель без регуляризатора
business company buy customer consumer	business money economy market government
economy wealth growth rich tax	business money economy market government
girl sister family father mother	child violence life man family
border crime jail violence prison	child violence life man family
social identity feel self relationship	way culture time person life
social identity feel self relationship	kind come ask friend feel
information data machine pattern computer	product new create build technology
information data machine pattern computer	neuron example different make brain

Таблица 4: Сравнение ближайших тем исходной модели и модели без регуляризатора. Коллекция TED.

Исходная модель	Модель с регуляризатором
business company buy customer consumer	investor customer fund business dollar
economy wealth growth rich tax	economy growth trade sector wealth
girl sister family father mother	grandfather wife door question mother
border crime jail violence prison	prison officer justice crime civil
social identity feel self relationship	friend desire compassion relationship
information data machine pattern compute	pattern process computer memory neuroscience

Таблица 5: Сравнение тем исходной модели и ближайших для них тем модели с регуляризатором. Коллекция Books.

Исходная модель	Модель без регуляризатора
application development technology model analysis	include content method use material
application development technology model analysis	application level development system include
kitchen meal cook delicious recipe	fruit vegetable meal healthy diet
kitchen meal cook delicious recipe	food cookbook kitchen recipe cook
topic criticism forum philosopher introduction	university work history book world
culture author study professor university	university work history book world
fuel engine speed vehicle table	system operation air vehicle table
interface deploy hardware computer software	system operation air vehicle table

Таблица 6: Сравнение ближайших тем исходной модели и модели без регуляризатора. Коллекция Books.

Исходная модель	Модель с регуляризатором
application development technology model analysis	structure analysis chapter model application
kitchen meal cook delicious recipe	food delicious cook meal recipe
topic criticism forum philosopher introduction	opus intellectual early philosopher criticism
culture author study professor university	society modern publish century history
fuel engine speed vehicle table	signal speed vehicle component sensor
interface deploy hardware computer software	engineer software practical model calculation

Таблица 7: Сравнение тем исходной модели и ближайших для них тем модели с регуляризатором. Коллекция BBC.

Исходная модель	Модель без регуляризатора
price rate economy rise growth	expect firm growth market sale
price rate economy rise growth	price last economy month year
comedy star direct producer film	win award star show year
comedy star direct producer film	movie comedy number star film
international black south rugby budget	government meeting budget economy minister
nation minister meeting view leadership	government meeting budget economy minister
concentration remark reporter comment apologise	week spokesman come find police
-	week spokesman come find police

Таблица 8: Сравнение ближайших тем исходной модели и модели без регуляризатора. Коллекция BBC.

Исходная модель	Модель с регуляризатором
price rate economy rise growth	price fall sale analyst market
comedy star direct producer film	movie drama comedy star film
international black south rugby budget	rugby coach squad injury captain
nation minister meeting view leadership	nation government problem area consider
concentration remark reporter comment apologise	explain comment exciting long someone
man society back allege police	person society police community officer

Таблица 9: Сравнение тем исходной модели и ближайших для них тем модели с регуляризатором. Коллекция TED.

4 Заключение

- Показано, что тематическая несбалансированность коллекции приводит к дроблению крупных и слиянию мелких тем

- Предложен алгоритм устранения проблемы несбалансированности, заключающийся в добавлении регуляризации на основе семантической однородности тем

- Проведены эксперименты, демонстрирующие возможные модификации модели и показывающие, что модель улучшает интерпретируемость результатов обработки несбалансированных коллекций

Список литературы

[1] Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res 3* (2003), 993–1022.

[2] Mimno, D. M., and Blei, D. M. Bayesian checking for topic models. In *EMNLP* (2011), ACL, pp. 227–237.

[3] Veselova, E., and Vorontsov, K. Topic balancing with additive regularization of topic models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, ACL 2020, Online, July 5-10, 2020* (2020), S. Rijhwani, J. Liu, Y. Wang, and R. Dror, Eds., Association for Computational Linguistics, pp. 59–65.

[4] Vorontsov, K., and Potapenko, A. Additive regularization of topic models. *Mach. Learn 101*, 1-3 (2015), 303–323.

[5] Wallach, H. M., Mimno, D. M., and McCallum, A. Rethinking lda: Why priors matter. In *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada* (2009), Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds., Curran Associates, Inc, pp. 1973–1981.