Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»
Физтех-школа Прикладной Математики и Информатики
Кафедра интеллектуальных систем

Направление подготовки / специальность: 03.03.01 Прикладные математика и физика
Направленность (профиль) подготовки: Математическая физика, компьютерные технологии и математическое моделирование в экономике

ВЕРОЯТНОСТНОЕ ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ НЕСБАЛАНСИРОВАННЫХ ТЕКСТОВЫХ КОЛЛЕКЦИЙ

(бакалаврская работа)

Студент:
Панкратов Виктор Владимирович
_____
(подпись студента)

Научный руководитель:
Воронцов Константин Вячеславович,
д-р физ.-мат. наук
_____
(подпись научного руководителя)

Консультант *(при наличии)*:
_____
_____
(подпись консультанта)

Москва 2021

Тематическое моделирование - одно из направлений обработки текстовых коллекций. Задачей тематического моделирования является определение тем, к которым можно отнести каждый документ текстовой коллекции, а также выделение соответствующих темам слов. Каждое слово может принадлежать более чем одной теме, таким образом образуя дискретное распределение вероятностей принадлежности слова темам. Аналогично, существует распределение вероятностей и для принадлежности документов различным темам. Модели, решающие задачу тематического моделирования дают на выходе такие распределения для заданной текстовой коллекции. Коллекции могут обладать различной структурой и свойствами в зависимости от их истории формирования. Желаемое требование к тематическим моделям состоит в том, чтобы качество решения было одинаковым при любых обрабатываемых данных, что случается далеко не всегда. В качестве примера можно рассмотреть коллекцию из 5000 документов по литературе и 50 по математике. Тематические модели, разработанные на текущий момент, не могут корректно обработать коллекции с этим свойством, в данном случае разбивая литературу на более мелкие подтемы и сливая математику с другими получаемыми темами. Это не всегда является желаемым результатом. В работе демонстрируется данная проблема, рассматривается ее решение путем добавления регуляризатора семантической неоднородности и экспериментально показана эффективность данного решения при работе с синтетическими коллекциями с различным балансом тем.

Содержание

1 Введение

Рассматриваются текстовые коллекции, состоящие из документов, каждый из которых в свою очередь состоит из термов. Полагается, что порядок вхождения термов не важен: каждый документ можно представить как неупорядоченное множество термов. Одной из задач обработки текстов является определение по заданным множествам термов в каждом документе доли документов в коллекции, отвечающей различным темам, а также набора слов, соответствующего этим темам. Каждый документ и каждое слово для этого соотносится с некоторой, возможно нулевой, вероятностью к каждой теме. Таким образом, задача состоит в поиске двух данных дискретных распределений вероятностей. Для их нахождения используются тематические модели, которые решают задачу приближенного матричного разложения. В описанном виде это некорректно поставленная оптимизационная задача: у нее существует бесконечное множество решений, поэтому необходимо дополнительно использовать регуляризацию для выделения одного из них.

Решение задачи тематического моделирования основано на максимизации логарифма правдоподобия путем EM-алгоритма. В процессе максимизации темы получаются примерно равными по мощности. В то же время равномощные темы не являются обязательным условием входных данных. Процесс сбора документов некотролируем и в коллекции могут оказаться одновременно группы тем, число документов одних из которых в десятки и более раз превышает число документов других. При решении оптимизационной задачи мелкие темы сольются с более крупными; крупные же разделятся, образуя схожие подтемы, что ухудшит изначальную интерпретируемость тем.

Баланс тем может быть произвольным. Мы будем считать коллекцию сбалансированной, если документы равномерно распределены по темам. Если же распределение существенно отличается от равномерного, такую коллекцию называют несбалансированной; чем больше отличие, тем больше степень несбалансированности.

Сама по себе постановка задачи тематического моделирования не нова [2]. В литературе можно найти много методов и попыток ее решения. Одним из наиболее известных является LDA[1], использующий в качестве регуляризатора логарифм априорного распределения Дирихле. Этот подход был позже расширен до ARTM [5], описывающий подход, использующий произвольные регуляризаторы. На этом исследования в

направлении тематических моделей не ограничились, но в настоящее время одной из наиболее общей идей является ARTM. Вид используемых регуляризаторов зависит от конкретной задачи. В то же время проблема несбалансированности не является свойством конкретной задачи, она связана лишь с данными на входе модели. Это порождает дополнительные трудности в использовании какого-либо регуляризатора для ее решения - мы не имеем возможность проверить его необходимость. Поэтому он должен быть применим для любой возможной коллекции.

Поводом к желанию использовать регуляризацию является отсутствие полного решения проблемы несбалансированности. Попытки ее решения другими способами уже известны[4]. Эксперименты в таких работах обычно проводятся на синтетических данных: для них можно оценить баланс тем в процессе генерации. В данной работе также были сгенерированы синтетические коллекции с различной степенью несбалансированности; кроме этого было проиллюстрировано существование проблемы несбалансированности путем демонстрации качества восстановления тем и предложено ее решение путем добавления в оптимизационную задачу регуляризатора семантической неоднородности[3].

2 Постановка задачи

2.1 Общая постановка задачи тематического моделирования

Пусть D - множество документов, T - конечное множество тем, W - множество термов. Каждый из документов $d \in D$ задается его длиной n_d, $\sum_{d \in D} n_d = n$ и последовательностью термов $\{w_i \in W\}_{i=1}^{n_d}$, элементы которой в дальнейшем будем называть словами. Вероятностная модель порождения коллекции вводится при следующих дополнительных предположениях:

- Гипотеза мешка слов: вышеописанное представление документа d эквивалентно представлению документа d в виде неупорядоченного множества входящих в него слов, в которое каждое слово w входит n_{dw} раз.

- Гипотеза условной независимости: вероятность появления слова w в документе d по теме t не зависит от документа d и описывается

распределением

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

С учетом данных предположений вероятность появления слова w в документе d описывается распределениями $p(w|t) = \phi_{wt},\ p(t|d) = \theta_{td}$. Задача тематического моделирования заключается в нахождении этих распределений. Это эквивалентно задаче получения матричного разложения

$$F = \Phi\Theta \tag{1}$$

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

Ставится задача максимизации функции правдоподобия. Данную задачу решают с помощью EM-алгоритма.

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} \to \max_{\Phi,\Theta} \tag{2}$$

Задача (1) поставлена некорректно: в общем случае ее множество решений бесконечно. Для выделения из этого множества решений одного в функцию (2) добавляют один или несколько регуляризаторов, зависящих от матриц Φ, Θ. Вид регуляризаторов определяется тем, какие свойства ожидаются от результата. Функция правдоподобия при этом принимает следующий вид:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} + \sum_i \tau_i R_i(\Phi,\Theta) \to \max_{\Phi,\Theta} \tag{3}$$

2.2 Проблема несбалансированности

Результатом, который выдает вышеописанная модель являются две матрицы Φ, Θ. Они описывают порождение коллекции: матрица Φ показывает вероятность конкретного слова в документе с заданной темой, а матрица Θ показывает распределение тем между документами. Данный алгоритм уже исследовался в литературе, в частности в [6] показывалось, что устойчивость результата зависит от априорного представления о матрицах Φ, Θ. Выданные такой моделью матрицы Θ часто свидетельствуют о равенстве мощностей всех тем, то есть распределение $p(t)$ определенное как $p(t) = \sum\limits_{d\in D} p(t|d)n_d$ не сильно отличается от равномерного, а именно

$\forall\ t_i, t_j \in T \to \frac{p(t_1)}{p(t_2)} \approx 1$ вне зависимости от истинного вида $p(t), t \in T$, что было упомянуто ранее как "проблема несбалансированности". Такой эффект возникает из-за используемого алгоритма: при максимизации правдоподобия модели выгодно использовать все свои параметры.В свою очередь, сокращение долей отдельных тем приводит к неполному использованию, а в пределе - к уменьшению числа параметров. В реальных же коллекциях темы могут оказаться несбалансированными. Чтобы повысить качество решения для несбалансированных коллекций, в модель предлагается добавить регуляризатор.

Вернемся к гипотезе условной независимости. Ее эквивалентная формулировка: $p(w, d|t) = p(w|t)p(d|t)$. Она описывает распределение термов для всех документов заданной темы. Чтобы проверить эту гипотезу для темы t предлагается [3] статистика

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

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

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

Данную статистику также можно записать в виде

$$S_t = \sum_{d \in D} \sum_{w \in d} \hat{p}(w, d|t) \ln \frac{\hat{p}(w, d|t)}{p(w|t)p(d|t)} \tag{4}$$

Предполагая, что гипотеза условной независимости верна, значение данной статистики, получившей название семантической неоднородности темы, должно быть мало, а распределения $p(w|d, t)$ должны быть близки к $p(w|t)$. Можно представить каждую тему как кластер размерности $|W|$, центром которого является $p(w|t)$. Статистика семантической неоднородности показывает удаленность $p(w|d, t)$ от центра кластера.

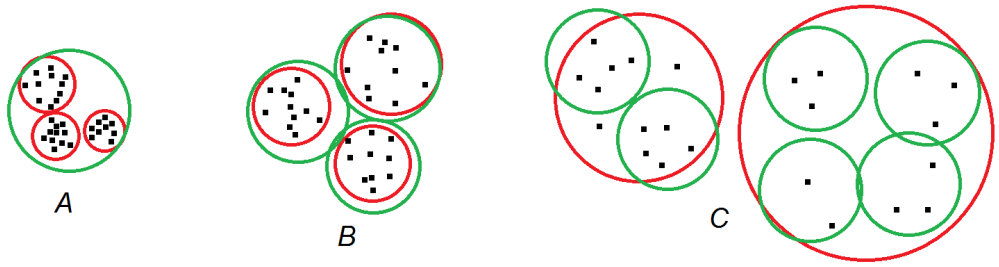

Рис. 1: Выравнивание тем по мощности (красные кластеры) и по равенству статистики семантической неоднородности (зеленые кластеры)

На рисунке 1 представлено два способа разбиения коллекции на темы. Первый - красные кластеры - выравнивание тем по мощности. Именно такой результат обычно дают тематические модели. Второй - зеленые кластеры - выравнивание тем по отклонению от центра кластера - семантической неоднородности. Как видно из части B рисунка, распределение тем может быть одинаковым для данных двух способов, но так происходит не всегда. По сравнению с первым способом, во втором некоторые темы объединились в одну большую, некоторые же наоборот разделились исходя из удаленности точек в них между собой. Ожидается, что учет статистики семантической неоднородности приведет к решению исходно заявленных проблем, а именно дроблению крупных тем и слиянию мелких, что и случилось для красных кластеров на рисунке 1.

Исходя из этого в задачу можно добавить регуляризатор R, учитывающий значение (4). Чтобы это сделать, будем суммировать статистику S_t по всем темам

$$\sum_{t\in T} S_t = \sum_{d\in D}\sum_{w\in d}\left(\sum_{t\in T}\hat{p}(w,d|t)\right)\ln\frac{\hat{p}(w,d|t)}{p(w|t)p(d|t)} \to \min_{\Phi,\Theta} \tag{5}$$

Преобразовав сумму, используя $\hat{p}(w,d|t) = \frac{p(t|d,w)\hat{p}(w|d)p(d)}{p(t)}$ для преобразования логарифма и домножив на постоянное для конкретной задачи n, получим формулу для вычисления. Вставим ее в оптимизационную задачу в качестве регуляризатора R:

$$R = \sum_{d\in D}\sum_{w\in d}\beta_{dw}n_{dw}\ln\sum_{t\in T}\phi_{wt}\theta_{td} \tag{6}$$

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

Заметим, что выражение (6) это в точности выражение (2), домноженное на весовые множители β_{dw}. По смыслу эти множители увеличивают вес маломощных тем $p(t) << 1$

Цель данной работы - исследовать влияние добавления R на получаемое решение задачи тематического моделирования для синтетических коллекций с различным балансом тем.

3 Вычислительный эксперимент

3.1 Генерация коллекции

Для эксперимента будем генерировать синтетическую коллекцию данных. В этом разделе описан процесс генерации, а также указаны числовые параметры дальнейших экспериментов, которые будут таковыми, если в конкретном эксперименте не указано обратное.

Столбцы матриц Φ_0, Θ_0 порождаются симметричными распределениями Дирихле. Параметр распределения определяется из соображений реалистичности коллекции и берется малым для разреженности получаемых матриц. Для матрицы Φ_0 он берется равным 0.02, для матрицы Θ_0 равным 0.2. Чтобы регулировать баланс тем, будем на этом этапе генерации менять наибольшие значения в столбцах Θ_0 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, порожденная несимметричным распределением Ципфа. Матрица Θ_0 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_0. Затем слово генерируется из столбца Φ_0, соответствующего теме t_i. Таким образом, процесс генерации документов описывается как

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

Данным способом для каждого эксперимента будет сгенерирована коллекция из принадлежащих 100 темам 2000 документов, в каждом из которых по 1000 слов. Общее количество различных слов в словаре 10000.

Необходимо определиться, как измерить степень несбалансированности полученной коллекции.Определим мощность темы t: $|t|$ как сумму

вероятностей ее появлений в каждом документе, то есть сумма соответсвующей строки матрицы Θ. Тогда под первым определением степенью несбалансированности коллекции будем подразумевать

$$\frac{\max\limits_{t\in T}|t|}{\min\limits_{t\in T}|t|} \tag{7}$$

и говорить о ней как о степени несбалансированности в первом смысле.

Упоминая степень несбалансированности во втором смысле, мы будем рассматривать сумму элементов матрицы

$$|\Theta - \frac{1}{|T|}| \tag{8}$$

Здесь за $\frac{1}{|T|}$ обозначена матрица одной с Θ размерности, состоящая из элементов $\frac{1}{|T|}$. Степень несбалансированности во втором смысле является отклонением от равномерного распределения по темам всех документов. Необходимость введения двух определений будет объяснена в разделе **3.7** .

3.2 Оценивание качества восстановления тем.

Для эксперимента мы генерируем коллекции с различной степенью несбалансированности и для каждой из них используем стандартную модель для нахождения матриц Φ, Θ. Чтобы оценить сходство полученных матриц Φ с используемыми при генерации в данной работе считается количество взаимно ближайших по некоторой метрике столбцов матриц Φ, Φ_0, то есть пар столбцов $\Phi[i]$,$\Phi_0[j]$:

$$\arg\min_k \ (dist(\Phi[i], \Phi_0[k])) = j \tag{9}$$

$$\arg\min_k \ (dist(\Phi[k], \Phi_0[j])) = i \tag{10}$$

Здесь *dist* - расстояние по заданной метрике. В экспериментах, как правило, будет представлено косинусное расстояние(сниуз справа на графике), евклидово(сверху справа), Хеллингера(сверху слева) и Йенсена-Шеннона(снизу слева), однако за исключением раздела **3.8** рассуждения

в ней опираются лишь на результаты при *dist* - расстоянии Йенсена-Шеннона. Некоторые обоснования возможности исключения метрик из рассмотрения приведены в **3.8**

Будем полагать, что если вышеописанная оценка сходства низкая, исходная матрица Φ плохо восстанавливается и, напротив, если она высока, то и матрица Φ восстановлена хорошо. Тогда идеальный результат достигается, когда каждой исходной теме сопоставляется ровно одна из тем на выходе модели. На практике такое достигается крайне редко и существуют темы $\Phi_0[j]$, которые не являются ближайшими ни для какой темы из Φ, то есть для любого i для пары $\Phi[i], \Phi_0[j]$ не выполнено (9). Такие темы будем называть невосстановленными. Аналогично определим ложные темы: такие темы $\Phi[i]$, что для любого j (10) не выполнено для пары $\Phi[i], \Phi_0[j]$.

В дальнейшем количество ложных и невосстановленных тем, а также число взаимно ближайших пар будет оцениваться в зависимости от параметров эксперимента.

3.3 Основной вычислительный эксперимент: коллекции с большой степенью несбалансированности.

Для демонстрации одновременно влияния на качество восстановления тем добавления регуляризатора R и существования проблемы несбалансированности эксперимент был проведен на коллекциях с большой, в данном случае в обоих описанных смыслах, степенью несбалансированностью. Были сгенерированы коллекции, в которых мощности всех тем, кроме одной были примерно равны, а мощность оставшейся значительно больше.

Сгенерировав коллекцию по такому принципу мы рассмотрим зависимость указанных в **3.2** метрик от степени несбалансированности, для определенности, в первом смысле. Параметры эксперимента были указаны ранее в **3.1** и отличаются только долей фоновой темы: 0.5 для первого графика и 0.7 для второго. Помимо этого представим на графике зависимость между степенями несбалансированности в двух смыслах для данного эксперимента. Для сравнения: первая коллекция, описанная крайними левыми точками на графиках ниже, сбалансированная:

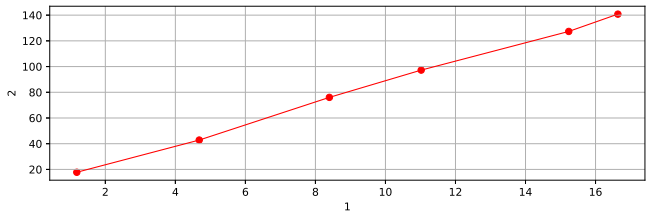

Рис. 2: Эксперимент на коллекциях с большой степенью несбалансированности. Зависимость между степенями несбалансированности в первом и втором смысле для используемых коллекций.

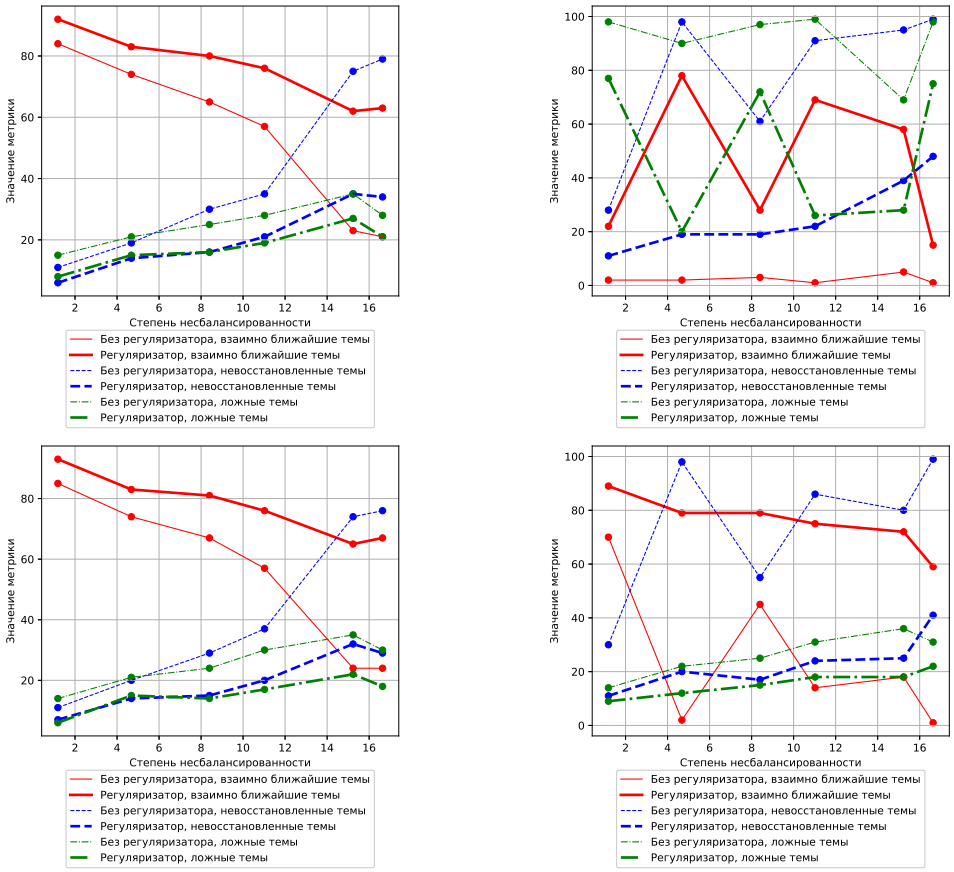

Рис. 3: Эксперимент на коллекциях с большой степенью несбалансированности. Доля фоновой темы 0.5. Значительное улучшение восстановления тем при добавлении регуляризатора (красные кривые)

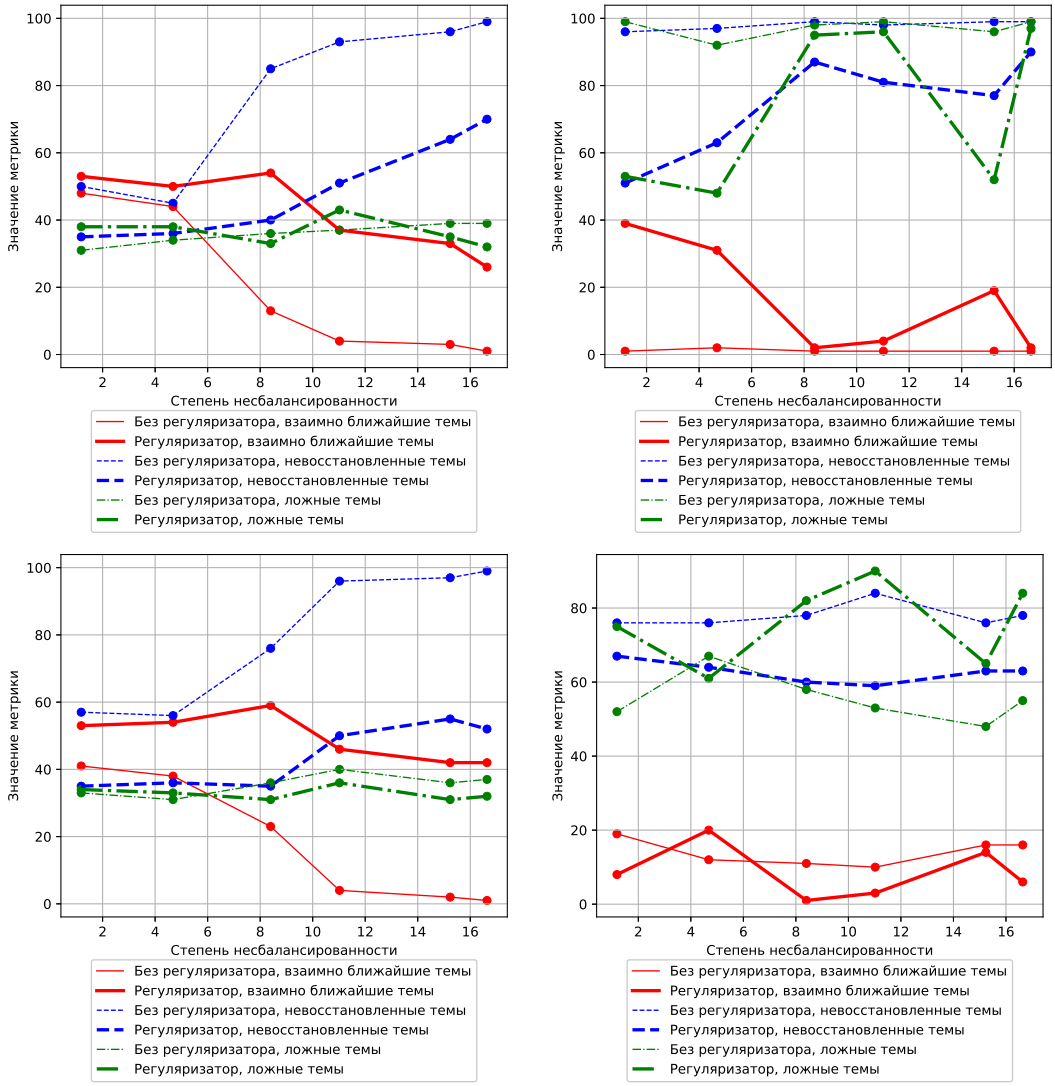

Рис. 4: Эксперимент на коллекциях с большой степенью несбалансированности. Доля фоновой темы 0.7. Значительное улучшение восстановления тем при добавлении регуляризатора (красные кривые)

Из графиков видно, что при использовании регуляризатора число взаимно ближайших пар тем выросло, а число ложных и невосстановленных тем упало. Также можно заметить, что количество невосстановленных тем для модели без регуляризатора растете по мере увеличения

степени несбалансированности. Из этого можно сделать вывод, что так как темы на выходе модели без регуляризатора слабо сопоставимы с исходными, такая модель не справилась с задачей восстановления тем для данной коллекции.

3.4 Эксперимент на сбалансированных коллекциях.

Как уже отмечалось ранее, изначально никакого представления о балансе тем в коллекции у модели не имеется. На вход модели могут подать как сбалансированную, так и несбалансированную коллекцию. Для каждой возможной коллекции модель будет одинаковой и поэтому важно проверить, что введенный нами регуляризатор не ухудшит качество работы модели на сбалансированных коллекциях. Исходя из данных соображений проведен следующий эксперимент.

Сгенерируем несколько сбалансированных коллекций с различной долей фоновой темы в документах и рассмотрим результат работы модели на этих коллекциях с использованием регуляризации и без нее. Когда мы говорим о сбалансированной коллекции мы также имеем это в виду в обоих смыслах. Для корректности эксперимента зависимость степеней несбалансированности в двух смыслах также приведена на графике.

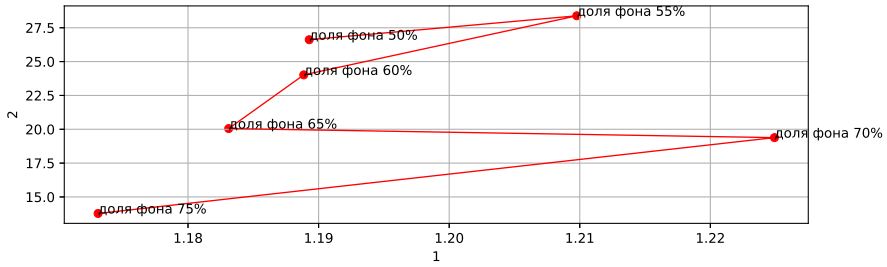

Рис. 5: Эксперимент на сбалансированных коллекциях. Зависимость между степенями несбалансированности в первом и втором смысле для используемых коллекций.

В этот раз связь какого-либо рода между степенями не видна из графика, но так произошло из-за близости точек на нем между собой, поэтому какой-то вывод о несоответствии двух определений степени несбалансированности сделать невозможно. Также можно убедиться, что зна-

чения, представленные на этом графике примерно равны значениям на прошлом графике для сбалансированной коллекции.

Результаты самого эксперимента представлены на графике ниже.

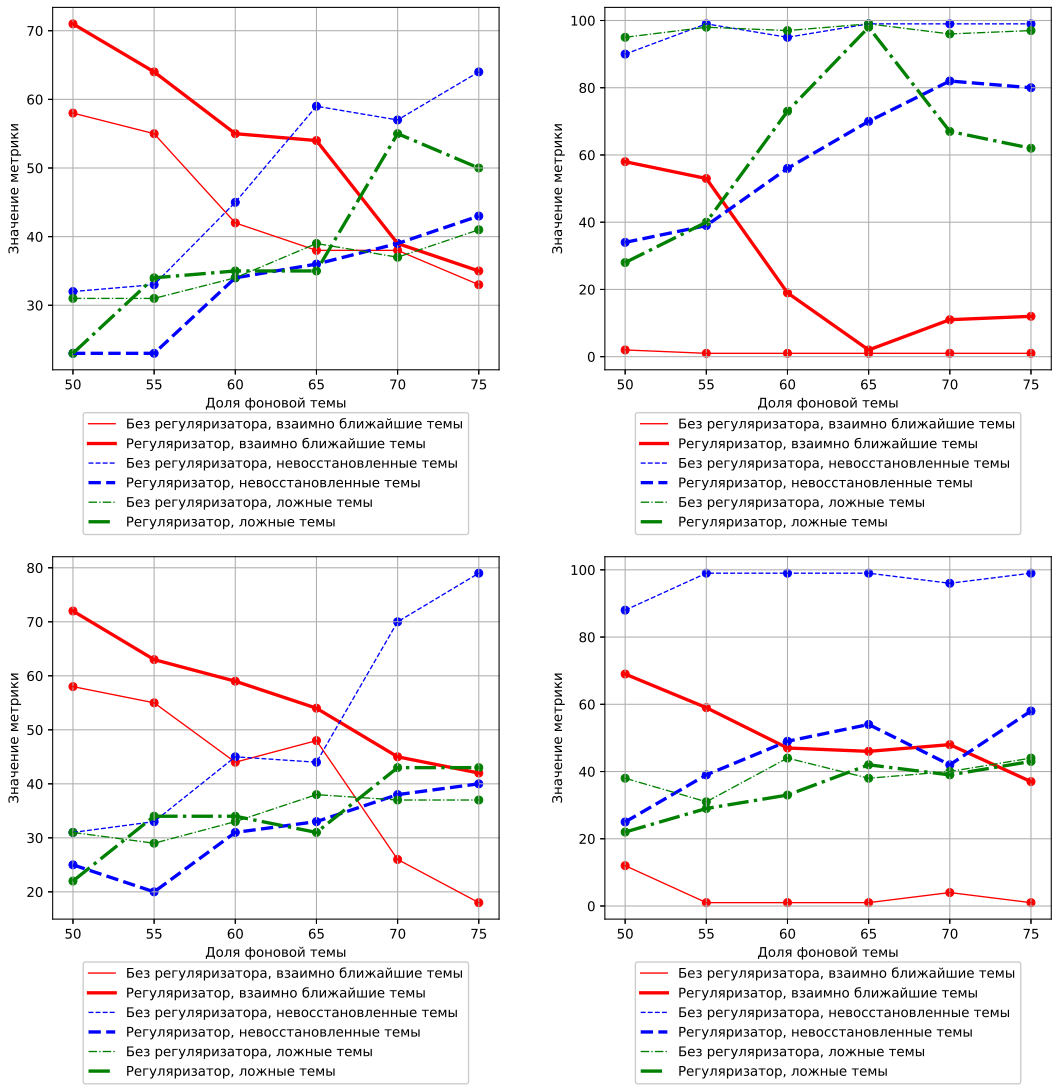

Рис. 6: Эксперимент на сбалансированных коллекциях. Значения метрик для модели без регуляризатора не лучше, чем значения метрик для модели с его добавлением(сравниваются кривые одного цвета).

Из графика видно, что при добавлении регуляризатора число взаим-

но ближайших тем стало выше, чем было изначально, а число невосстановленных тем снизилось. Для корректности результатов поставленного эксперимента было бы достаточно, если бы данные метрики остались на том же уровне, как это и произошло с числом ложных тем, линии которых на графике пересекаются и остаются на уровне 30-40. Таким образом, результаты данного эксперимента не противоречат предположению о том, что регуляризатор пригоден и для сбалансированных коллекций и не ухудшает в смысле описанных ранее метрик качество решения для них.

3.5 Эксперимент на коллекциях с малым числом крупных тем

В первом эксперименте в каждой из коллекций была ровно одна крупная тема. Рассмотрим случай, когда их число k невелико: $k \in \{1 \dots 6\}$. Сами крупные темы при этом сделаем примерно равными по модулю между собой и равными по модулю для каждого эксперимента, то есть вместе с увеличением числа больших тем при генерации очередной коллекции будут меняться вероятности малых тем.

Как и в предыдущих случаях, опишем зависимости между степенями несбалансированности в двух смыслах для коллекций в виде графика.

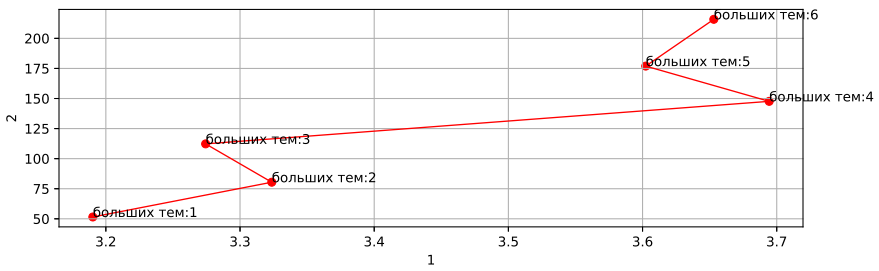

Рис. 7: Эксперимент на коллекциях с малым числом крупных тем. Зависимость между степенями несбалансированности в первом и втором смысле для используемых коллекций.

Чем больше крупных тем, тем меньше вероятности остальных тем и тем больше степень несбалансированности в первом смысле. Аналогично будет расти и отклонение всех вероятностей от равных, поэтому вид графика соответствует ожиданиям.

Ниже представлены результаты данного эксперимента. Мы видим, что с добавлением регуляризатора, используя метрики, для которых без регуляризации темы не восстанавливались, можно взаимно однозначно сопоставить больше половины тем.

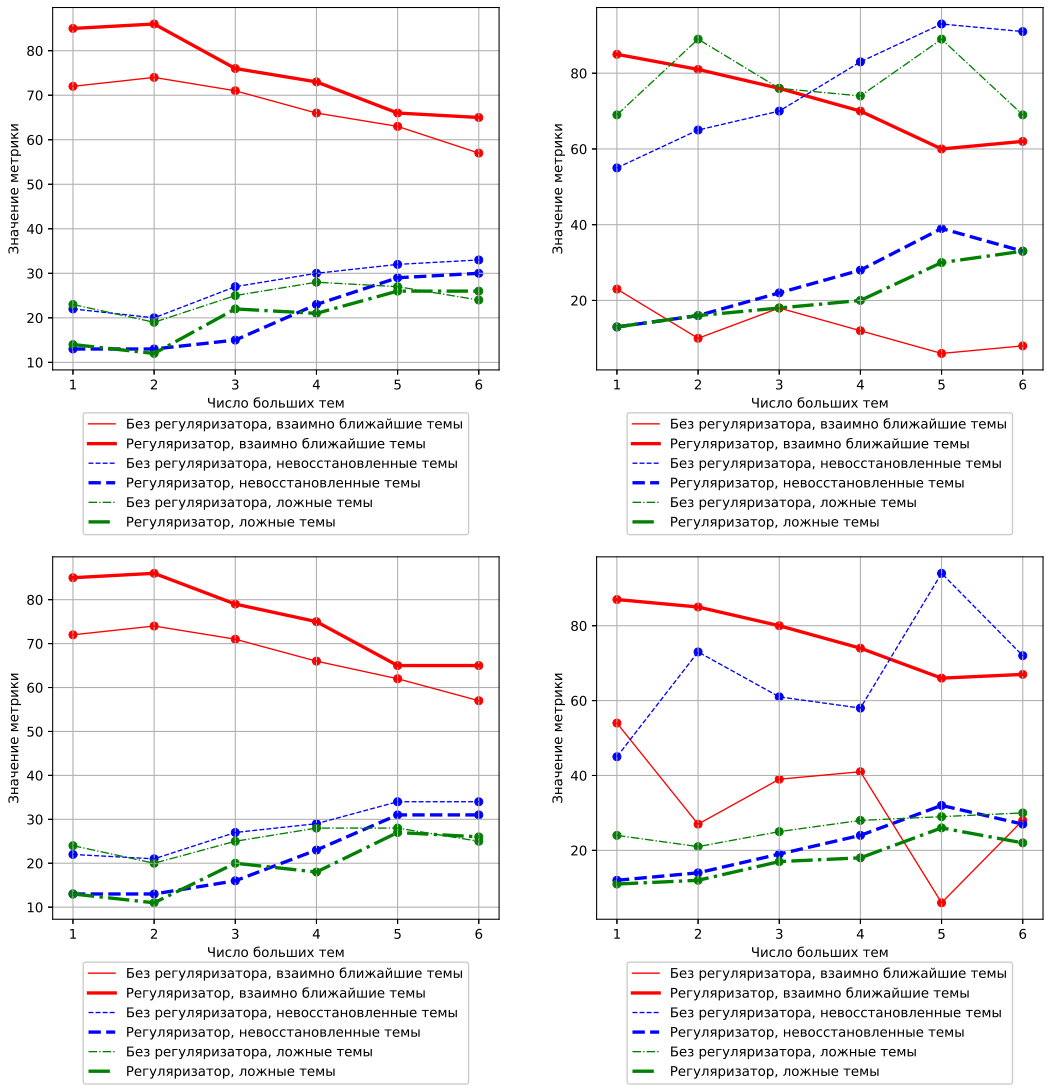

Рис. 8: Эксперимент на коллекциях с малым числом крупных тем. При введении регуляризатора темы стали лучше восстанавливаться во всех исследуемых метриках

3.6 Эксперимент на коллекции с большим числом крупных тем

В предыдущем эксперименте был рассмотрен случай более чем одной большой по мощности тем. Однако их было достаточно мало: от 1 до 6. Эксперимент был построен так для получения большой степени несбалансированности. Однако большее количество крупных тем так и не было рассмотрено.

Исходя из этого был проведен еще один эксперимент. Сгенерируем коллекцию с половиной больших тем и половиной малых, но с меньшей степенью несбалансированности. Оказалось, что в получившейся коллекции в первом смысле коллекция со степенью 1.3 близка к сбалансированной, но во втором значение равно 107, что соответсвует 2-3 крупным темам в прошлом эксперименте. Результат в зависимости от коэффициента регуляризации представлен на графике ниже.

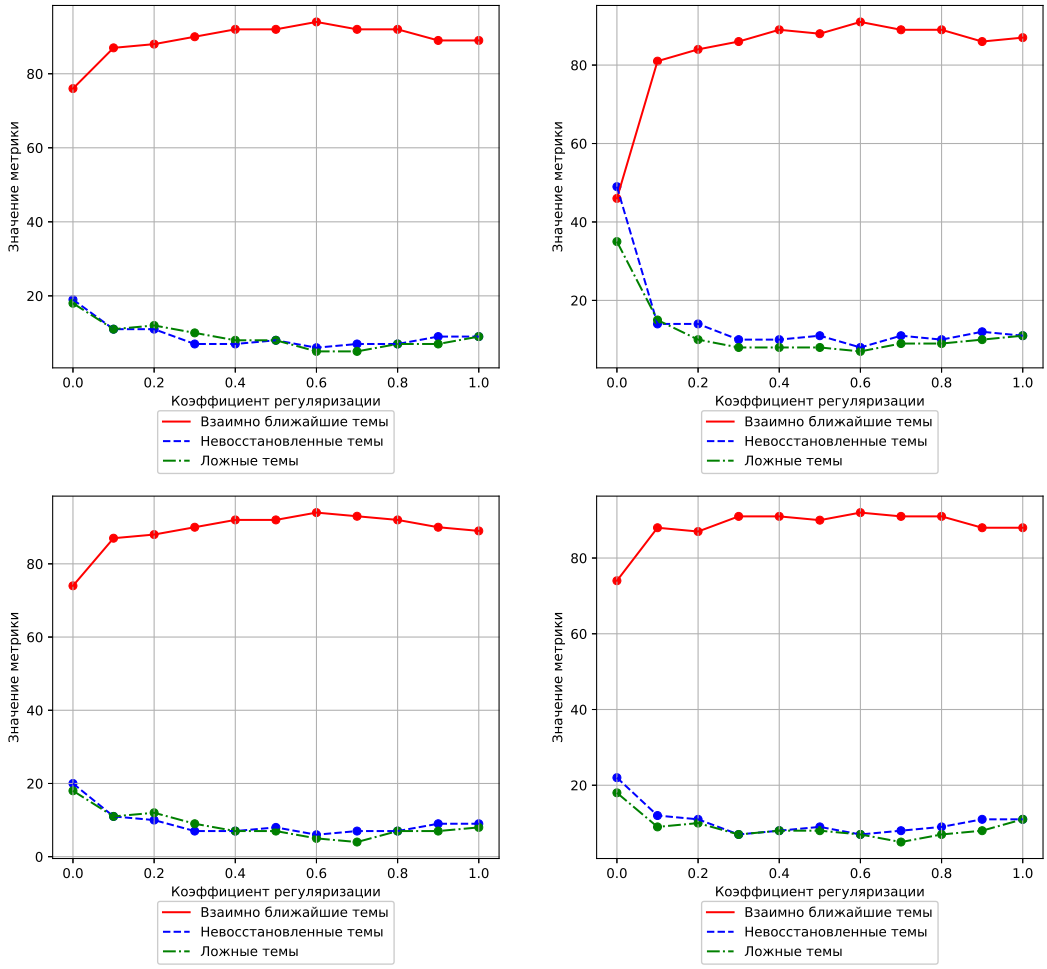

Рис. 9: Эксперимент на коллекции с большим числом крупных тем. Слабая зависимость исследуемых метрик от коэффициента регуляризатора

Здесь зависимость метрик от коэффициента регуляризации почти незаметна. Мы отложим обсуждение этого до следующего раздела.

3.7 Объяснение необходимости двух определений степени несбалансированности

Проведем также эксперимент на коллекции, где большинство тем равномощные, но существуют по $\frac{1}{20}$ от общего числа тем большие и меньшие

по мощности. При этом отношение мощностей между самыми большими и малыми невелико, как и в предыдущем эксперименте: оно равно 1.65. Во втором смысле значение степени несбалансированности равно 48, что также достаточно мало. Зависимость значений исследуемых метрик от коэффициента регуляризации представлена на графике ниже.

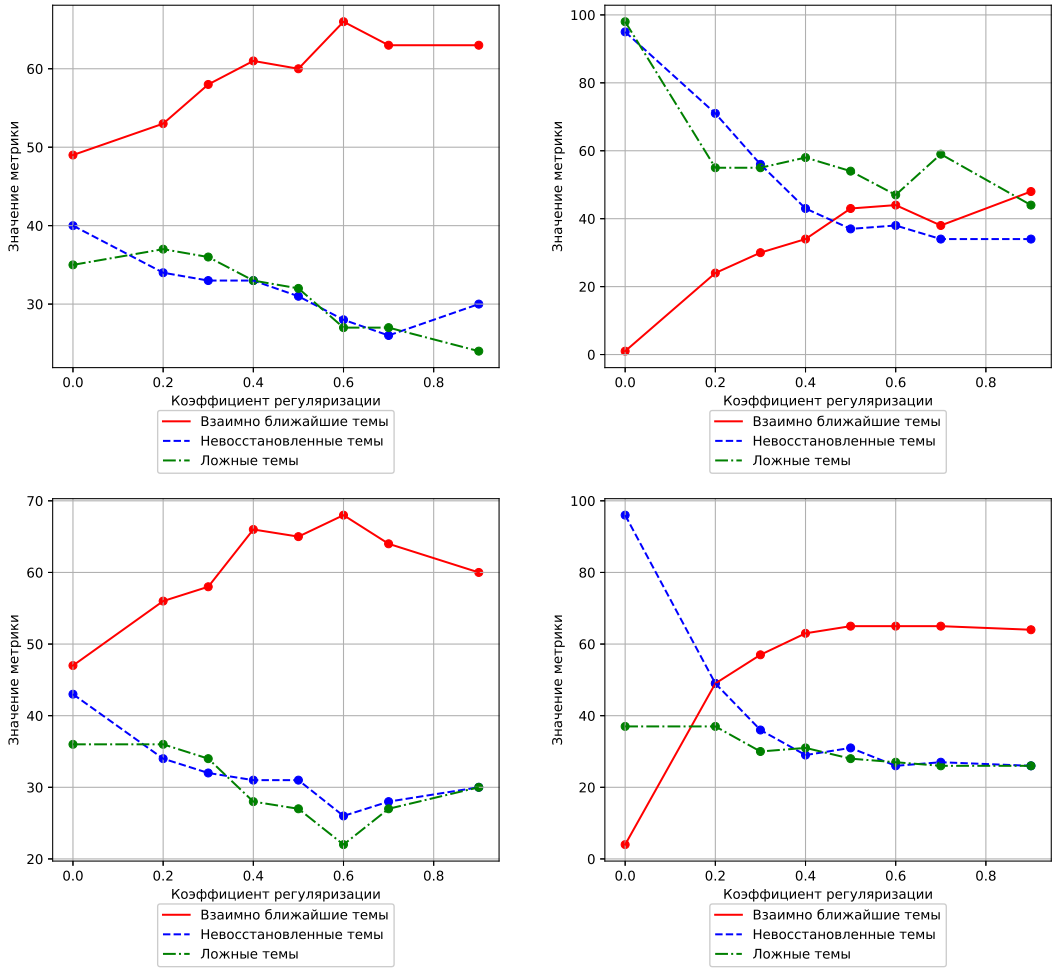

Рис. 10: Эксперимент на коллекции с увеличенной степенью несбалансированности в первом смысле. Сильная зависимость исследуемых метрик от коэффициента регуляризатора

Здесь, в отличие от предыдущего эксперимента, влияние коэффициента регуляции заметно и в отсутствии регуляризатора модель не мо-

жет найти половину пар. Но при этом степень несбалансированности во втором смысле упала по сравнению с предыдущим экспериментом. Более резкую зависимость можно объянить наличием четко выраженных больших и малых тем и повышением, пусть и небольшим, степени несбалансированности в первом смысле. Более того, из первого эксперимента мы понимаем, что одной крупной темы достаточно, чтобы модель без регуляризатора не восстановила почти все темы в наборе. Таким образом, первое определение степени несбалансированности имеет смысл. Однако оно никак не отражает баланс всех, кроме двух, тем в коллекции. Это может быть важно - в качестве примера расссмотрим следующий эксперимент.

3.8 Объяснение возможности выбора метрик

После рассмотрения коллекций с небольшой долей крупных и маломощных тем рассмотрим другую коллекцию, в которой число больших и малых тем будет примерно равно и равно числу остальных. Во втором смысле степень несбалансированности примерно совпала с одним из предыдущих экспериментов (107), что превосходит степень для коллекции в прошлом эксперименте. В первом же смысле степень несбалансированность равна 1.63. Результаты эксперимента для составленной таким образом коллекции представлены на графике ниже

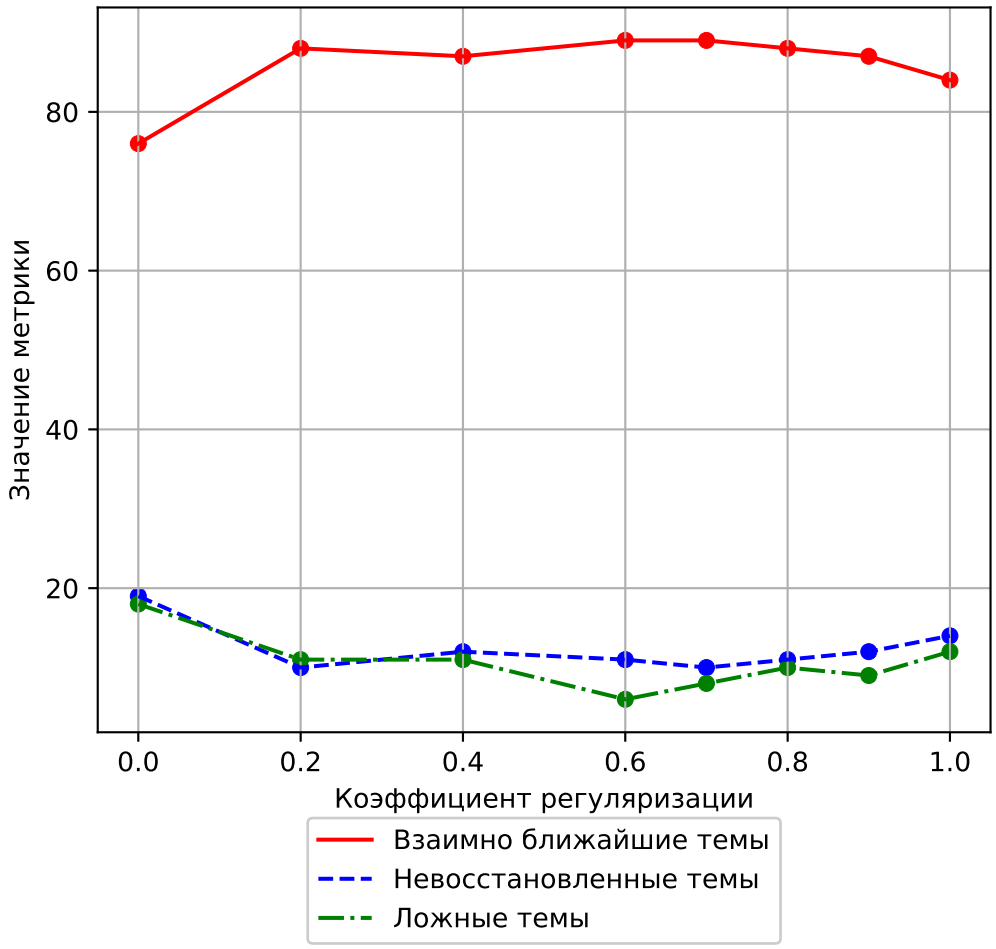

Рис. 11: Эксперимент на коллекции с тремя различными по мощности группами тем. Расстояние Йенсена-Шеннона. Высокие и почти постоянные значения метрик.

Приведу некоторые соображения, по которым выбиралась метрика для представления в данной части работы. Рассмотрим последний график. Можно заметить, что при обработке такой коллекции даже исходная модель восстановила большинство тем: многие пары полученных и исходных тем можно взаимно сопоставить. Также практически отсутствует зависимость исследуемых метрик от коэффициента регуляризации. Теперь рассмотрим тот же график в другой, косинусной, метрике:

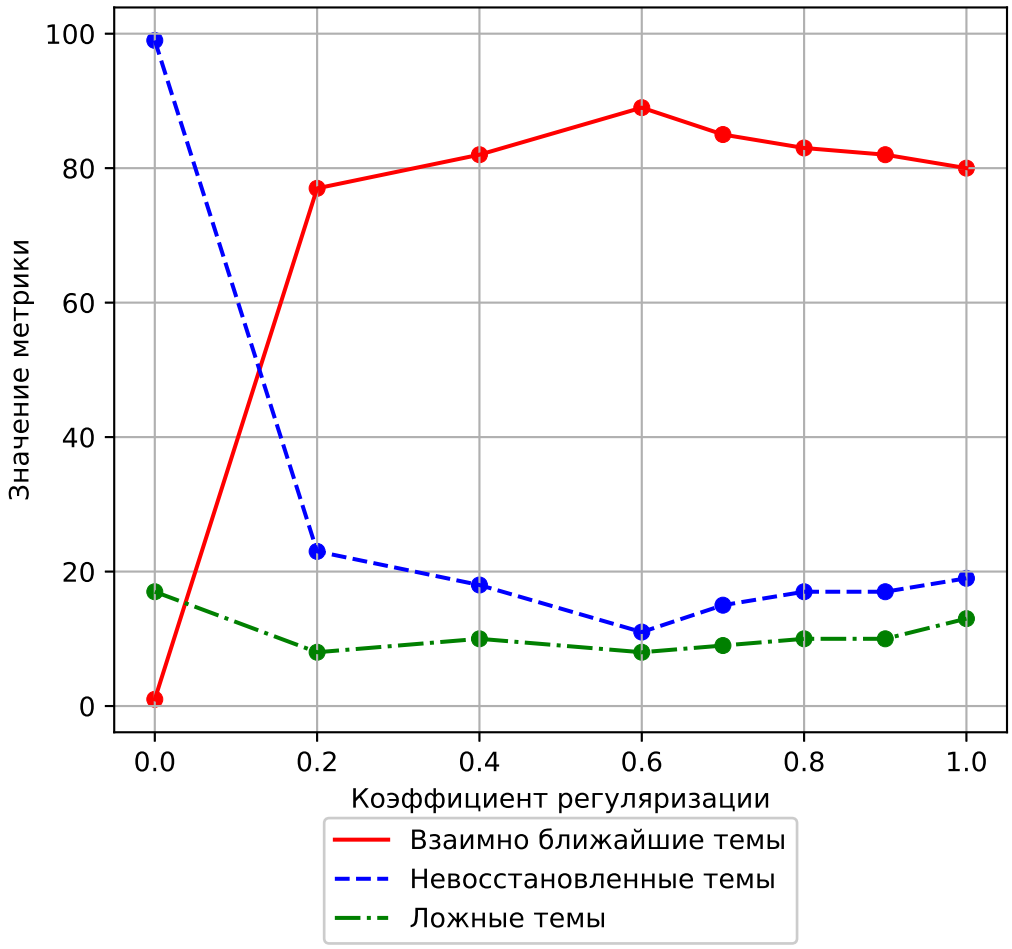

Рис. 12: Эксперимент на коллекции с тремя различными по мощности группами тем. Косинусное расстояние. Значения метрик зависят от наличия регуляризатора в модели и коэффициента регуляризации.

По этому графику число восстановленных моделью без регуляризатора тем почти нулевое. Тем не менее, из первого графика мы видим, что темы все же можно соотнести друг с другом. Сравнения по косинусному расстоянию не показывают свойства получаемого решения. Аналогично, на последнем графике имеется пик для значения регуляризатора 0.6, что неверно для первого графика. Если подобное происходит систематически, что и наблюдается для данных двух расстояний, можно предпочесть его другому. Подчеркнем, что это утверждение не несет характер

критерия и исходя из подобных рассуждений не всегда удается предпочесть одну метрику другой. Поэтому в работе приведены результаты для четырех различных метрик. Но именно из подобных соображений описание результатов эксперимента опирается на результаты, использующие *dist* - расстояние Йенсена-Шеннона.

В качестве дополнительного следствия из этого и предыдущего эксперимента рассмотрим значения метрик для косинусного расстояния в отсутствии регуляризатора. Мы видим, что на последней коллекции мы не смогли сопоставить почти все. В то же время для предыдущего эксперимента была сопоставлена половина. Таким образом, распределение тем между собой и второе определение степени несбалансированности также важны.

4 Заключение

- Было показано, что семантическая несбалансированность приводит к дроблению крупных тем и слиянию мелких.
- Был предложен регуляризатор, использующий статистику семантической неоднородности тем.
- Проведены эксперименты на синттических данных; показано, что регуляризатор решает проблему несбалансированности; показана зависимость характера проблемы от баланса тем в коллекции

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] Hofmann, T. Probabilistic latent semantic analysis. In *UAI* (1999).

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

[4] 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.

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

[6] 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.