Министерство образования и науки Российской Федерации
Московский физико-технический институт (национальный исследовательский университет)
Факультет управления и прикладной математики
«Кафедра математического моделирования сложных процессов и систем»
при Вычислительном центре им. А. А. Дородницына РАН

Итерационные методы балансировки тем в тематическом моделировании

03.03.01 — Прикладные математика и физика

Выпускная квалификационная работа

Студент 578 группы:
Плюснин Павел Андреевич

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

Москва
2019 г.

Аннотация

Тематическое моделирование основано на принципе правдоподобия, который максимально задействует все свои параметры-темы, что приводит к их выравниванию, в результате чего крупные темы дробятся, а мелкие — объединяются в коллекциях с несбалансированными темами.

В данной работе показана проблема несбалансированности тем и предложен итеративный алгоритм для ее решения. Совместно с итеративным алгоритмом использовались подходы регуляризации и кластеризации.

Содержание

1 Введение

Тематическое моделирование — один из самых важных методов анализа тестовых данных и естественного языка. Вероятностная тематическая модель позволяет кластеризовать коллекцию документов, описывая каждую тему дискретным распределением на множестве токенов, а каждый документ — распределением на множестве тем. Таким образом, тематическая модель является эффективным средством систематизации и анализа больших коллекций текстов, и данных, которые можно свести к ним (например, в работе [1] показан анализ банковских транзакций посредством тематического моделирования).

Построение тематической модели сводится к задаче стохастического матричного разложения, которая решается EM-алгоритмом, который в свою очередь максимизирует правдоподобие. Но модели, основанной на принципе максимизации правдоподобия выгодно делать темы примерно равными по мощности, что приводит к дроблению крупных тем и объединению мелких, если в коллекции темы несбалансированы, ведь мощность реальных тем определяется историей формирования коллекции. Эта проблема была названа проблемой несбалансированности тем.

В данной работе было экспериментально показано существование проблемы несбалансированности тем, а также предложен итеративный алгоритм для ее решения. Итеративный алгоритм был исследован совместно с другими техниками, такими как регуляризация и кластеризация.

1.1 Вероятностное тематическое моделирование

Тематическое моделирование — метод выявления скрытой семантической структуры в корпусе текстов. Тематическая модель выявляет латентные темы по наблюдаемым распределениям слов в документах.

Пусть даны D - множество документов (также называется коллекцией, корпусом текстов), каждый документ $d \in D$ представляет собой последовательность слов из словаря W: $d = (w_1, w_2, \ldots, w_n),\ w_i \in W\ \forall i \in \overline{1 \ldots n}$, где n - длина документа d.

Будем предполагать, что существует конечное множество тем T, такое, что каждое слово документа связано с некоторой темой $t \in T$. Другими словами, мы предполагаем, что в процессе написания текста d автор задумывал некоторую тему t, пиша слово w.

Мы считаем, что в процессе написания текста d в нем выражено относительно малое количество тем, т.е каждый текст относится к небольшому числу тем из T.

Определение 1.1. *Задачей тематического моделирования является по заданному корпусу текстов определить число тем, распределения частот слов для каждой темы и степени принадлежности текстов к каждой из тем*

В вероятностной постановке задачи можно считать коллекцию множеством (d, w, t), взятым из дискретного распределения $p(d, w, t)$ на множестве $D \times W \times T$.

Вероятностные тематические модели учитывают следующие предположения:

- **Порядок документов в коллекции не важен**. Время добавления документа в коллекцию и другие его характеристики, за исключением его описания в виде последовательности токенов, не учитываются
- **Гипотеза «мешка слов»**: порядок слов в документе не важен. На самом деле, это очень сильное утверждение, но предполагается, что даже без учета порядка слов, без семантических причинно следственных связей, возможно восстановить тематики
- Каждая тема $t \in T$ описывается неизвестным распределением $p(w|t)$ на множестве слов W
- Каждый документ $d \in D$ описывается неизвестным распределением $p(t|d)$ на множестве тем T

- **Гипотеза условной независимости**: $p(w|t,d) = p(w|t)$, т.е появление слова w с темой t не зависит от документа d

Также, на практике довольно часто используются предположения о том, что

- слова, часто встречающиеся в большинстве документов не важны для выявления тем. Такие слова называются **стоп-словами** и обычно исключаются из словаря W и из документов из D
- слово в разных формах - одно и то же слово. Действительно, приняв гипотезу «мешка слов» нецелесообразно различать различные падежи, рода и прочие формы

Определение 1.2. *Построить тематическую модель - значит восстановить неизвестные распределения $p(w|t), p(t|d)$, т.е найти матрицы $\Phi = \|p(w|t)\|$ и $\Theta = \|p(t|d)\|$ по коллекции D*

По формуле полной вероятности, тематическая модель

$$p(w|d) = \sum_{t\in T} p(w|d,t)\cdot p(t|d) = \sum_{t\in T} p(w|t)p(t|d) = \sum_{t\in T} \varphi_{wt}\theta_{td} \tag{1.1}$$

т.е эта задача - задача стохастического матричного разложения.

1.1.1 PLSA: вероятностный латентный семантический анализ

Эта задача стохастического разложения матриц Φ, Θ обычно решается максимизацией правдоподобия:

$$\mathscr{L}(\Phi,\Theta) = \prod_{d\in D}\prod_{w\in d} p(w|d)^{n_{dw}} \to \max_{\Phi,\Theta} \tag{1.2}$$

где n_{dw} - количество вхождений слова w в документ d.
Переходя к логарифму правдоподобия $\mathscr{L}$ и заменяя $p(w|d)$ на $\sum_{t\in T}\varphi_{wt}\theta_{td}$:

$$L(\Phi,\Theta) = \sum_{d\in D}\sum_{w\in d} n_{dw}\sum_{t\in T}\varphi_{wt}\theta_{td} \to \max_{\Phi,\Theta} \tag{1.3}$$

при условиях, что Φ и Θ — стохастические, т.е

$$\varphi_{wt} \geqslant 0, \quad \sum_{w\in W}\varphi_{wt} = 1 \tag{1.4}$$

$$\theta_{td} \geqslant 0, \quad \sum_{t\in T}\theta_{td} = 1 \tag{1.5}$$

Для решения данной оптимизационной задачи обычно применяется EM-алгоритм [4]:
На E-шаге, по φ_{wt} и θ_{td} вычисляются $p(t|d,w)$, а на M-шаге по найденным вероятностям $p(t|d,w)$ пересчитываются φ_{wt} и θ_{td}. Более подробно EM-алгоритм будет приведен ниже.

Такой подход получил название «Вероятностный латентный семантический анализ» (англ. PLSA, Probabilistic Latent Semantic Analysis) и был впервые предложен в 1999 году[6]

1.1.2 ARTM: аддитивные регуляризаторы в модели PLSA

Стоит заметить что оптимизационная задача 1.3 с ограничениями 1.5 является некорректно поставленной[7], так как существует бесконечное число решений Φ, Θ. Действительно, если Φ и Θ — решения, то и $\Phi' = \Phi S,\ \Theta' = S^{-1}\Theta$ — тоже решения для любой матрицы S ранга $|T|$:

$$\Phi'\Theta' = (\Phi S)(S^{-1}\Theta) = \Phi\Theta$$

Определение 1.3. *Регуляризация - стандартный прием доопределения решения путем добавления дополнительных ограничений.[2]*

Модель ARTM является модификацией PLSA путем добавления регуляризатора $R(\Phi,\Theta) = \sum_i \tau_i R_i(\Phi,\Theta)$ к логарифму правдоподобия [5][10]:

$$L(\Phi,\Theta) = \sum_{d\in D}\sum_{w\in d} n_{dw} \sum_{t\in T} \varphi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta} \tag{1.6}$$

при тех же ограничениях 1.5 на стохастичность матриц Φ, Θ
Тогда EM-алгоритм принимает вид метода простой итерации: 1.7–1.9

E-шаг: $$p_{tdw} \equiv p(t|d,w) = \underset{t\in T}{\operatorname{norm}}(\varphi_{wt}\theta_{td}) \tag{1.7}$$

M-шаг: $$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right), \qquad n_{wt} = \sum_{d\in D} n_{dw}p_{tdw} \tag{1.8}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right), \qquad n_{td} = \sum_{w\in d} n_{dw}p_{tdw}$$

где $$\underset{t\in T}{\operatorname{norm}}(x_t) = \frac{\max(x_t,0)}{\sum\limits_{s\in T}\max(x_s,0)} \tag{1.9}$$

Использование регуляризаторов существенно улучшает спектр возможностей тематического моделирования, предлагает обобщенный способ учитывания как эвристических наблюдений, так и строго вероятностных знаний [8].
Например, модель LDA [3] можно представить моделью ARTM с регуляризатором

$$R(\Phi,\Theta) = \ln \prod_{t\in T} Dir(\varphi_t,\beta) \prod_{d\in D} Dir(\theta_d,\alpha) + const =$$
$$= \sum_{t\in T}\sum_{w\in W} (\beta_w - 1)\ln \varphi_{wt} + \sum_{d\in D}\sum_{t\in T} (\alpha_t - 1)\ln \theta_{td} \tag{1.10}$$

Модель LDA основана на предположении, что столбцы матриц φ_t, θ_d являются случайными величинами, порождаемыми из распределения Дирихле с параметрами $\alpha \in \mathbb{R}^{|T|}$ И $\beta \in \mathbb{R}^{|W|}$

Описанный выше подход ARTM реализован в библиотеке BigArtm[9] в Python.

1.1.3 Разреживающий аддитивный регуляризатор

В качестве примера остановимся на разреживающих регуляризаторах.
Естественно предположить, что каждый документ d и каждое слово w связано лишь с небольшим количество тем t, ведь если токен принадлежит большому числу тем, тогда он является общеупотребительным и не поможет определить тематику документа (является стоп-словом). Тогда получаем, что значительная часть вероятностей φ_{wt} и θ_{td} должны быть равны 0.
При увеличении разреженности распределения, его энтропия увеличивается, а максимальную энтропию имеет равномерное распределение. Тогда будем максимизировать дивергенции Кульбака-Лейблера (перекрестную энтропию) между искомыми распределениями и равномерными:

$$\mathrm{KL}_w\left(\frac{1}{|W|}\|\varphi_{wt}\right) \to \max \tag{1.11}$$
$$\mathrm{KL}_t\left(\frac{1}{|T|}\|\theta_{td}\right) \to \max \tag{1.12}$$

Тогда приходим к регуляризатору

$$R(\Phi,\Theta) = -\beta_0 \sum_{t\in T}\sum_{w\in W} \beta_w \ln \varphi_{wt} - \alpha_0 \sum_{d\in D}\sum_{t\in T} \alpha_t \ln \theta_{td} \to \max \tag{1.13}$$

где $\alpha = (\alpha_t)_{t\in T}$, $\beta = (\beta_w)_{w\in W}$ - равномерные распределения, α_0 и β_0 - задаваемые неотрицательные коэффициенты регуляризации.

Но обычно этот регуляризатор записывается в виде

$$R(\Phi,\Theta)=\beta_0\sum_{t\in T}\sum_{w\in W}\beta_w\ln\varphi_{wt}+\alpha_0\sum_{d\in D}\sum_{t\in T}\alpha_t\ln\theta_{td}\to\max \tag{1.14}$$

Тогда при отрицательных значениях α_0 и β_0 имеем разреживающий регуляризатор, а при положительных — сглаживающий.

Сглаживающий регуляризатор эквивалентен предположению, что столбцы матриц Φ и Θ порождаются априорными распределениями Дирихле и имеет непосредственное отношение к модели LDA [3].

Легко видеть, что можно занулить один из коэффициентов α_0 или β_0 и тогда получить разреживание только матрицы Φ или Θ соответственно.

1.1.4 Перплексия: внутренний критерий качества

Критерии качества тематических моделей принято делить на внутренние и внешние. Внутренние критерии качества оценивают модель по исходному текстовому корпусу. Внешние критерии качества оценивают полезность модели с точки зрения конечного пользователя. Зачастую, это очень трудно сделать в автоматическом режиме, приходится прибегать к помощи асессоров.

Определение 1.4. *Перплексия — мера несоответствия модели $p(w|d)$ токенам w, наблюдаемым в документах коллекции D. Является внутренним критерием качества и определяется через логарифм правдоподобия как*

$$\mathscr{P}(D,p)=\exp\left(-\frac{1}{n}L(\Phi,\Theta)\right)=\exp\left(-\frac{1}{n}\sum_{d\in D}\sum_{w\in W}n_{dw}\ln p(w|d)\right), \tag{1.15}$$

где n — длина коллекции

1.2 Проблема балансирования тем

Вероятностные тематические модели основаны на принципе максимума правдоподобия, а для этого они должны полностью задействовать все свои параметры для описания коллекции. Модели не выгодно как сокращать количество тем (это означает уменьшение числа доступных параметров), так и сокращать доли отдельных тем в коллекции, так как это привело бы к неполному использованию параметров, а в пределе — к уменьшению числа тем. Получается, что модели выгодно использовать темы примерно в равных долях.

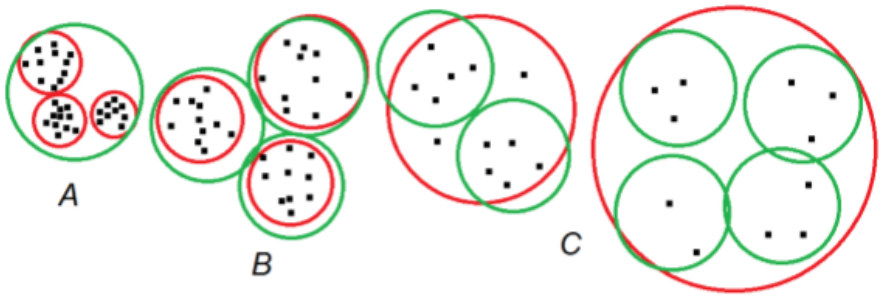

Рис. 1: Вероятностные тематические модели выравнивают темы по их мощности (красные кластеры), что приводит к появлению тем-дубликатов (A) и семантически разнородных тем (C). Лишь часть тем остаются семантически однородными и нераздробленными(B)

Определение 1.5. *Мощностью темы t будем называть число n_t слов данной темы в коллекции. Вероятность темы выражается как $\frac{n_t}{n}$, где n — общее количество слов*

С другой стороны, реальные текстовые коллекции определяются не принципом правдоподобия, а историей их создания, и могут оказаться несбалансированными.
Например, если в коллекции из тысячи документов 980 относятся к теме математики, 10 – биологии, а остальные 10 — психологии, то в тематической модели с тремя темами все три будут о математике, а 20 нематематических текста будут случайно распределены по этим темам.
Так как семантически однородные темы могут отличаться в разы или на порядки, а тематическая модель выравнивает их по мощности, то наиболее мощные темы окажутся разделенными на много мелких, отличающиеся друг от друга деталями, когда наименее мощные объединятся. В качестве иллюстрации смотри Рис.1
Нужен новый критерий, основанный не на максимуме правдоподобия, а на семантической близости.

2 Итеративное балансирование тем

Рассмотрим EM-алгоритм, приведенный выше, как метод простой итерации 1.7–1.9.

Обозначим за n_{tdw} — число раз, когда слово w из документа d относилось к теме t. Тогда попробуем решить проблему несбалансированности, домножив n_{tdw} на величину, обратно пропорциональную мощности темы n_t. Допустим, что n_{tdw} увеличилось в k_t раз по всей коллекции: $n'_{tdw} = k_t n_{tdw}$. Тогда вероятности перепишутся следующим образом:

$$p'_{tdw} = \frac{n'_{tdw}}{\sum_{s\in T} n'_{sdw}} = \frac{k_t n_{tdw}}{\sum_{s\in T} k_s n_{sdw}} = k_t p_{tdw} \frac{\sum_{s\in S} n_{sdw}}{k_s n_{sdw}} = \underset{t\in T}{\operatorname{norm}}(k_t p_{tdw}) \tag{2.1}$$

Положив $k_t = \frac{1}{n_t}$ получим модифицированный EM-алгоритм. В нем мощности n_t оцениваются через немодифицированные вероятности p_{tdw}, а параметры модели — через модифицированные p'_{tdw}

$$\text{E-шаг:} \qquad p_{tdw} \equiv p(t|d,w) = \underset{t\in T}{\operatorname{norm}}(\varphi_{wt}\theta_{td}) \tag{2.2}$$

$$p'_{tdw} = \underset{t\in T}{\operatorname{norm}}\left(\frac{\varphi_{wt}\theta_{td}}{n_t}\right), \qquad n_t = \sum_{d\in D}\sum_{w\in d} n_{dw} p_{tdw}$$

$$\text{M-шаг:} \qquad \varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right), \qquad n_{wt} = \sum_{d\in D} n_{dw} p'_{tdw} \tag{2.3}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right), \qquad n_{td} = \sum_{w\in d} n_{dw} p'_{tdw}$$

$$\text{где} \qquad \underset{t\in T}{\operatorname{norm}}(x_t) = \frac{\max(x_t, 0)}{\sum\limits_{s\in T} \max(x_s, 0)} \tag{2.4}$$

3 Вычислительные эксперименты

Для экспериментов использовался корпус текстов postnauka, состоящий из 3447 текстов

3.1 Генерация коллекций различной степени несбалансированности

Для проведения дальнейших экспериментов необходимо построить несколько наборов текстов с различными степенями несбалансированности тем.
Для этого сначала проведем базовую предобработку текстов:

1. Удалим из текстов все знаки, не являющиеся буквами, цифрами и апострофами. Апострофы оставляем для того, чтобы различать в английском языке слова навроде its и it's которые имеют совершенно разный смысл и роль в тексте.

2. Все года вида 19XX и 20XX заменяем на токен _YEAR_, все остальные числа - на токен _NUMBER_

3. Проводим нормализацию слов

4. Исключаем из коллекции стоп-слова

3.1.1 Отбор монотематических документов

Далее отберем из коллекции монотематические документы. Для этого на предобработанной коллекции postnauka строим модель с различным количеством тем $T+1$. Последнюю тему будем строить как тему общей лексики, остальные - уже специфичны. При обучении используем сглаживающий регуляризатор матрицы Φ на последней теме и разреживающий регуляризатор матрицы Φ на остальных темах. Коэффициенты регуляризации соответственно равны $\tau_s = 10^5,\ \tau_r = -10^5$. Далее для различных коэффициентов k определяем, сколько документов коллекции признается монотематическими, т.е $p(t_1|d) > kp(t_2|d)$, где t_1, t_2 - две наибольшие темы в документе d, не считая тему общей лексики.

Также, для каждой из моделей было определено количество монотематичных документов при $k = \overline{2\ldots9}$. В качестве примера, на Рис. 2 приведены зависи-

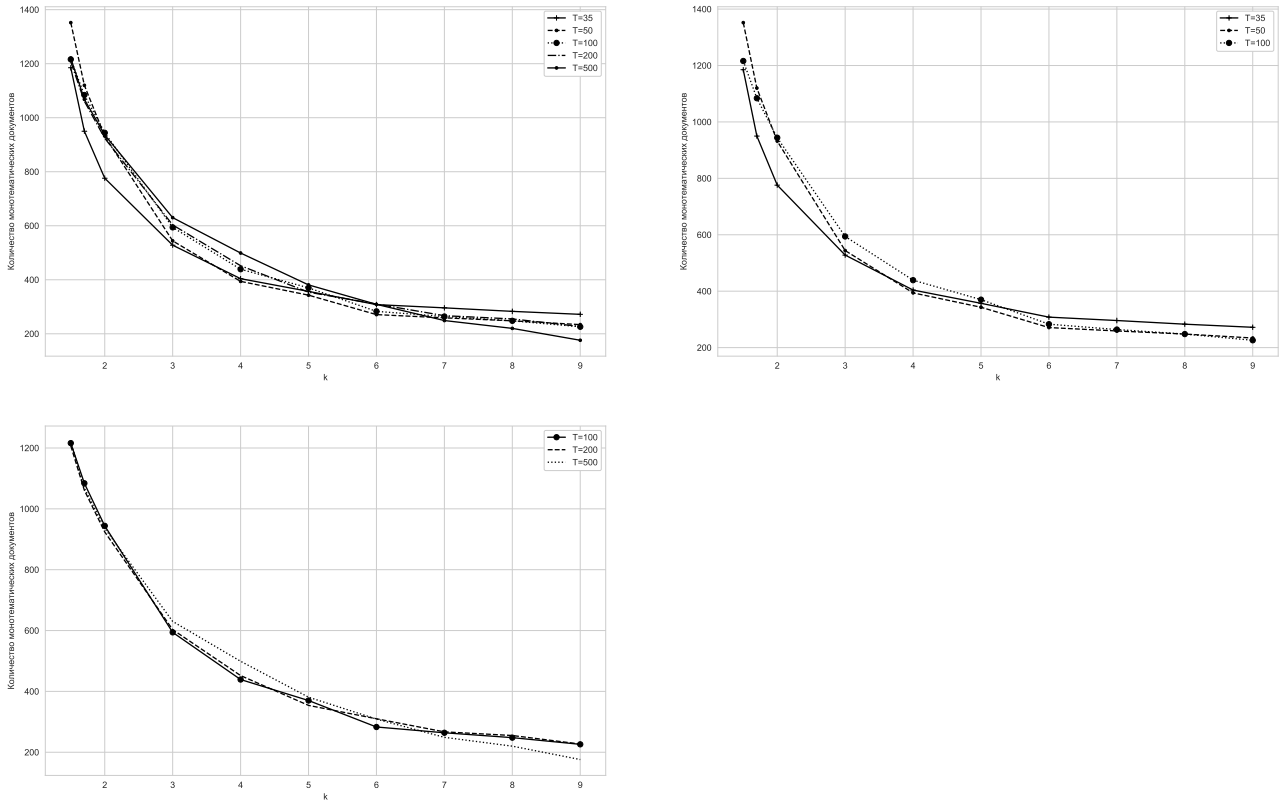

Рис. 2: График зависимости общего количества монотематичных документов от параметра k

мости общего количества монотематичных документов от параметра k для $T \in \{35, 50, 100, 200, 500\}$

Экспериментально был выбран коэффициент монотематичности $k = 1.7$ с приемлемым количеством отобранных документов при достаточной степени монотематичности.

При $T = 35$ получается относительно много отобранных документов, но темы, к которым они относятся очень плохо интерпретируются. Уже при небольшом изменении количества тем T количество отобранных документов резко падает и меняется незначительно вплоть до $T = 50$. Учитывая хорошую интерпретируемость тем при $T = 50$ и далее именно для таких T продолжим дальнейший анализ

Название темы, количество соответствующих монотематических документов и топ15 слов для первых 4х тем с наибольшим количеством монотематических документов при $T = 50$ можно увидеть на Рис.3

Название темы	N	Топ15 слов
Квантовая физика (sbj1)	66	теория вселенная квантовый математический волна физика пространство физик уравнение измерение математика закон состояние атом фотон
Генетика (sbj2)	49	организм ген молекула днк клетка белка белок химический геном молекулярный биологический бактерия биология соединение генетический
Право (sbj3)	24	право закон правовой римский суд юрист юридический платон диалог собственность судья тирания договор судебный владение
Лингвистика (sbj4)	24	язык русский языковой лингвист глагол английский лингвистический ’ носитель лингвистика диалект звук грамматика гласный согласный

Рис. 3: Самые богатые на монотематические документы темы и их интерпретация

3.1.2 Генерация новых коллекций

Из отобранных документов приступаем к созданию коллекций из $N = 1000$ документов с различной степенью несбалансированности k.

Для генерации требуемых коллекций для каждого T сперва получаем теоретические распределения

$$p(t) = p(t,k) = \frac{(1^k, 2^k, 3^k, \ldots, T^k)}{\sum_{i=1}^{T} i^k}$$

имеющие степенной тренд в зависимости от степени несбалансированности k. Например, при $k = 0$ имеем сбалансированные темы, при $k = 1$ их мощность линейно возрастает, при $k = 2$ - мощность возрастает квадратично.
Вид векторов $p(t,k)$ можно видеть на Рис.4

Будем считать, что этот вектор $p(t,k)$ задает дискретное распределение.
Теперь, генерируя iый из N текстов сперва выберем n ненулевых компонент вектора

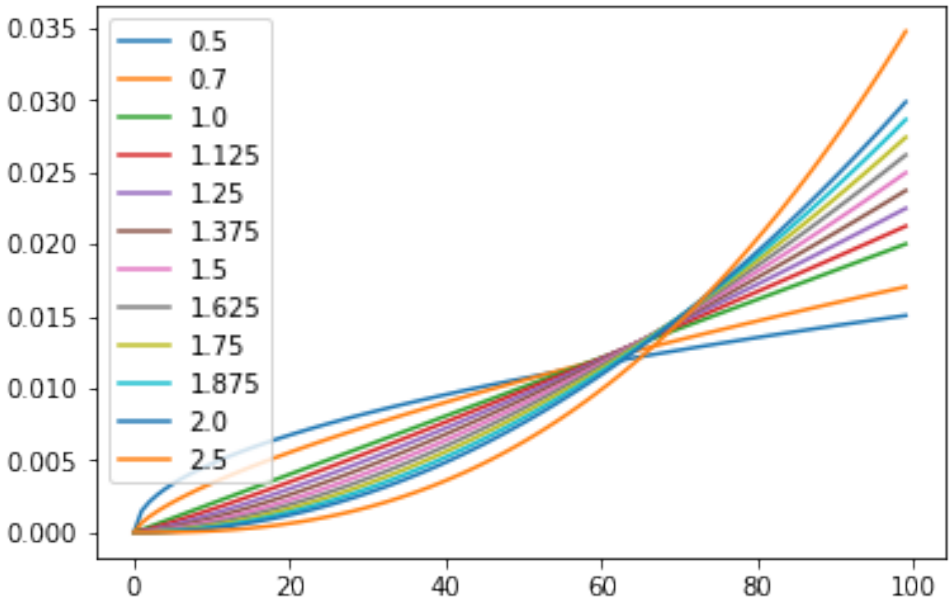

Рис. 4: Вектор $p(t)$ размерности $T = 100$ для различных степеней несбалансированности k

$p(t|d_i)$ из дискретного распределения $p(t, k)$, где

$$n = \begin{cases} 3, & \text{если } T = 50 \\ 4, & \text{если } T = 100 \\ 6, & \text{если } T = 200 \\ 12, & \text{если } T = 500 \end{cases}$$

Выбранные n соответствует эмпирическим наблюдениям количества существенно ненулевых компонент в реальных коллекциях документов (с использованием разреживающих регуляризаторов).

Далее генерируем из распределения Дирихле с $\alpha = 10$ случайный вектор размерности n, который помещаем в соответствующие ненулевые компоненты $p(t|d_i)$. Параметр распределения Дирихле $\alpha = 10$ обеспечивает достаточный разброс значений генерируемых компонент.

При большом количестве документов N реально получаемый $p(\hat{t}, k)$ близок к теоретически выбранному $p(t, k)$: См Рис.5

В генерируемый документ d_i берем соответствующие доли слов из случайных монотематических документов соответствующих тем, согласно $p(t|d_i)$ (предполагаем,

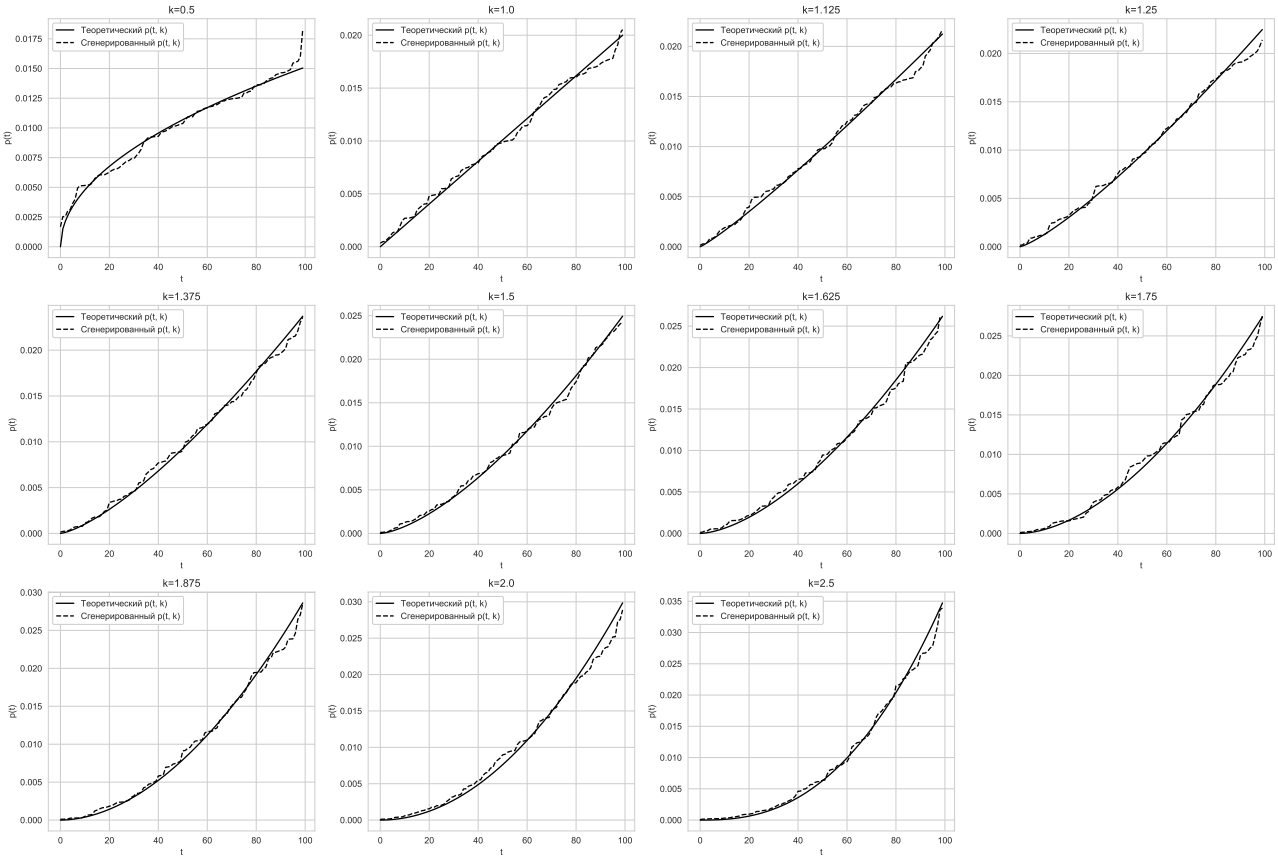

Рис. 5: Теоретический и сгенерированный $p(t,k)$ при $T = 100$, $N = 1000$

что все документы примерно равной длины)

3.2 Тематические модели PLSA

3.2.1 Тематическая модель PLSA без регуляризаторов

Целью этого эксперимента является иллюстрация проблемы несбалансированности тем. Для этого для каждого из построенных корпусов текстов построим тематическую модель PLSA без регуляризаторов как используя, так и не используя балансировку тем. Используя венгерский алгоритм, восстановим соответствие исходных и построенных тем. Сравним полученные распределения тем $\hat{\Theta}$ с истинным Θ

Результаты этого эксперимента видны на Рис.6—Рис.8. На графиках для каждого коэффициента несбалансированности k построены истинная и экспериментальные $p(t)$ в зависимости от темы $sbji,\ i = 1 \dots T$, причем соответствие исходных и построенных тем восстановлено. На Рис.9—Рис.11 венгерским алгоритмом соответствие не

восстанавливалось, а темы сортировались по мощности

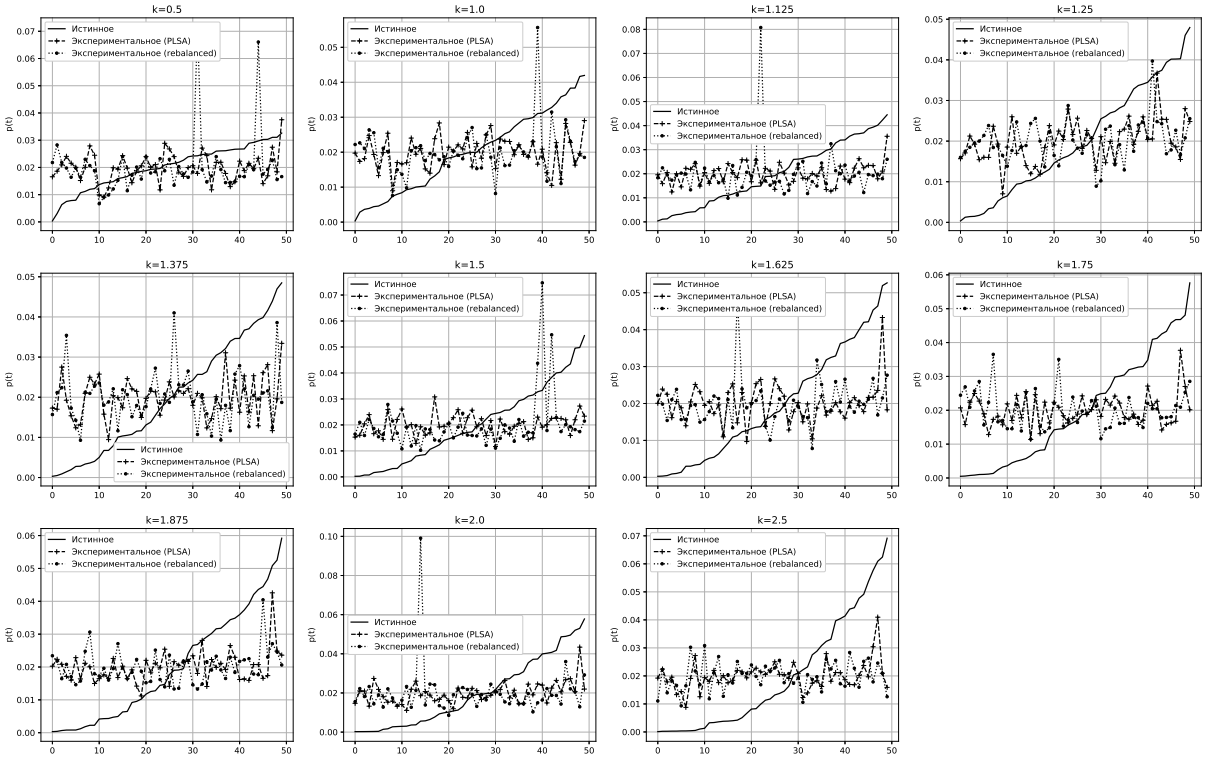

Рис. 6: Распределение p(t) для небалансированной и балансированной моделей без регуляризаторов $T = 50$, построенные темы были соотнесены с истинными темами

Видно, что модели PLSA строят примерно равномощные темы, вне зависимости от распределения мощностей истинных тем. Стоит заметить, что при этом модель с балансировкой тем последнюю тему делает существенно больше остальных, за счет чего в целом немного лучше восстанавливает исходный вид зависимости $p(t)$

Проверим, насколько хорошо модели восстанавливают сами темы, а именно их вероятностные профили.

Для этого для каждой истинной темы t_i (из истинной матрицы Θ строка $\Theta[t_i]$, описывающая принадлежность каждого документа к этой теме) найдем ближайшую к ней построенную тему $\hat{t}_j$ ($\hat{\Theta}[j]$). Пример такого соответствия приведен на Рис.12. Все такие графы приведены в Приложении.

Количество верно восстановленных тем (т.е таких, которые были признаны взаимно ближайшими, причем венгерский алгоритм их тоже соотнес вместе) в зависимости от коэффициентов несбалансированности k для различных T можно видеть

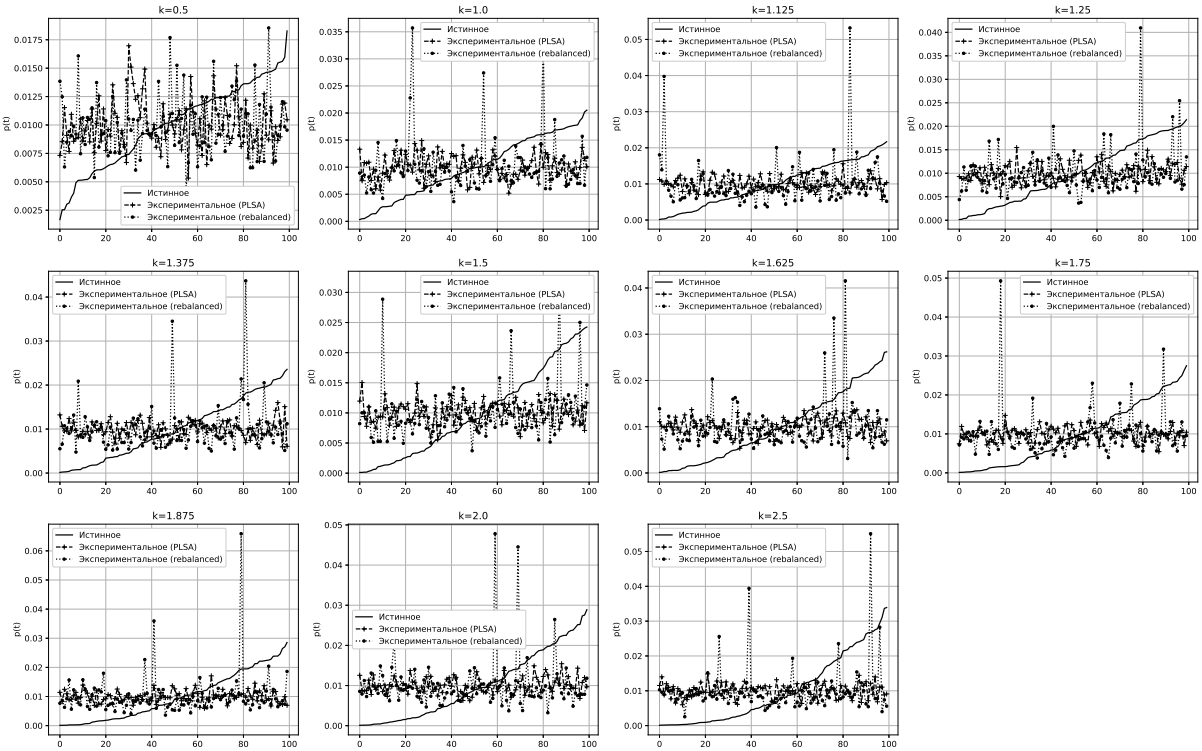

Рис. 7: Распределение p(t) для небалансированной и балансированной моделей без регуляризаторов $T = 100$, построенные темы были соотнесены с истинными темами

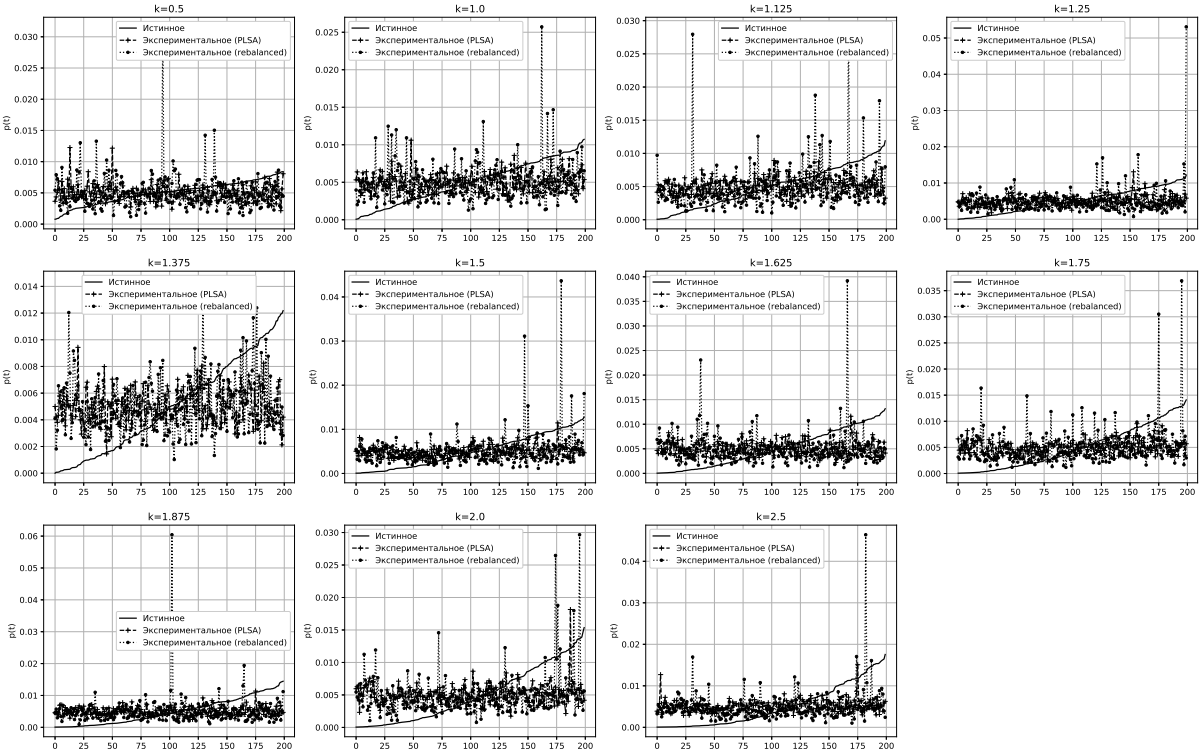

Рис. 8: Распределение p(t) для небалансированной и балансированной моделей без регуляризаторов $T = 200$, построенные темы были соотнесены с истинными темами

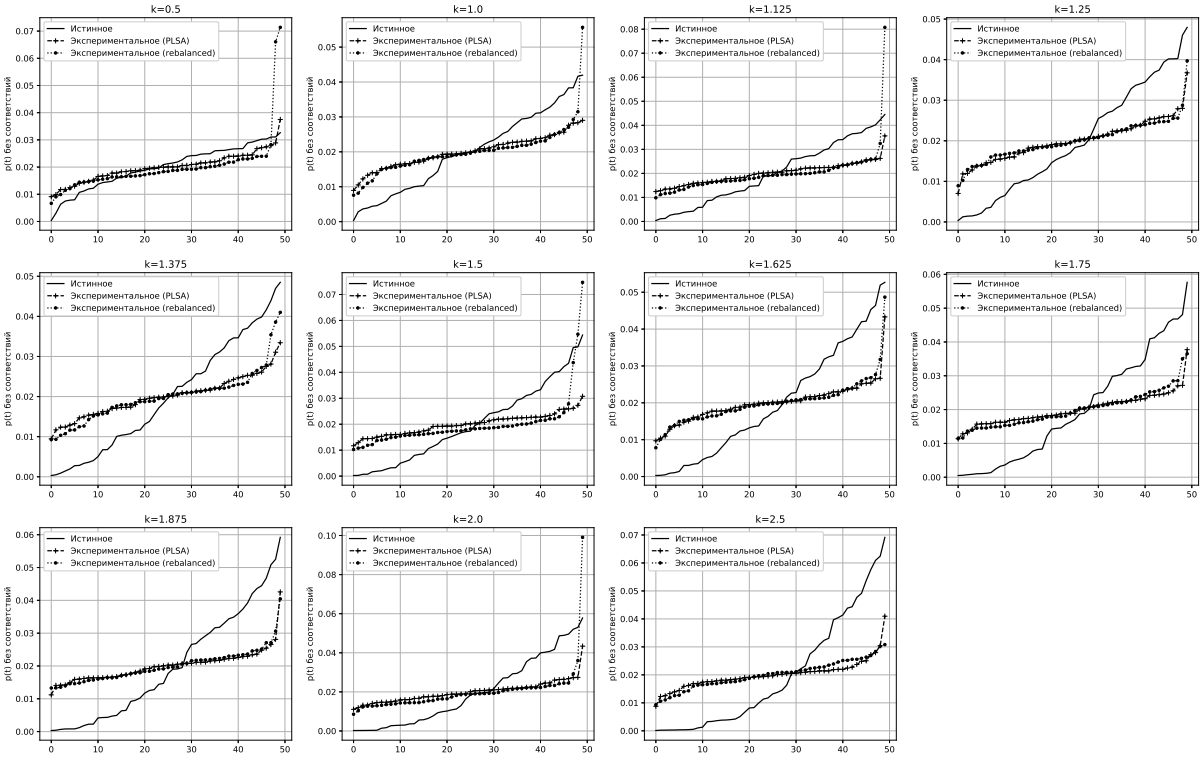

Рис. 9: Распределение p(t) для небалансированной и балансированной моделей без регуляризаторов $T = 50$, построенные темы отсортированы в порядке неубывания мощности

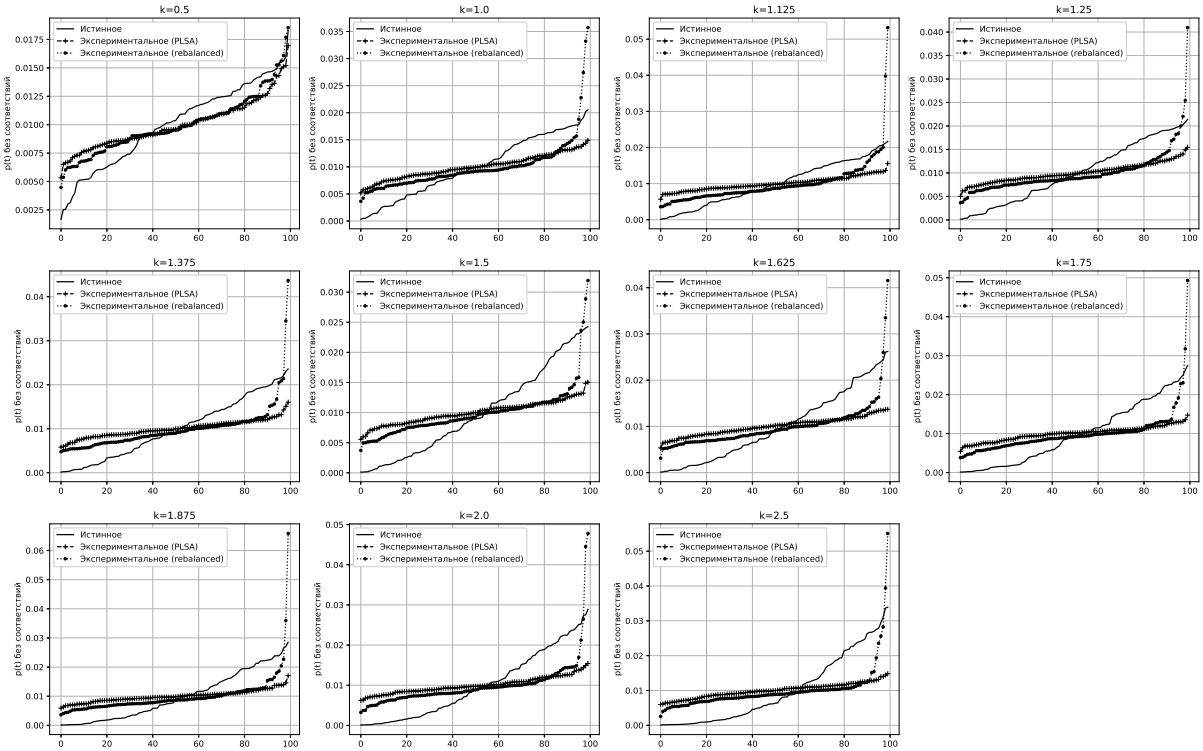

Рис. 10: Распределение p(t) для небалансированной и балансированной моделей без регуляризаторов $T = 100$, построенные темы отсортированы в порядке неубывания мощности

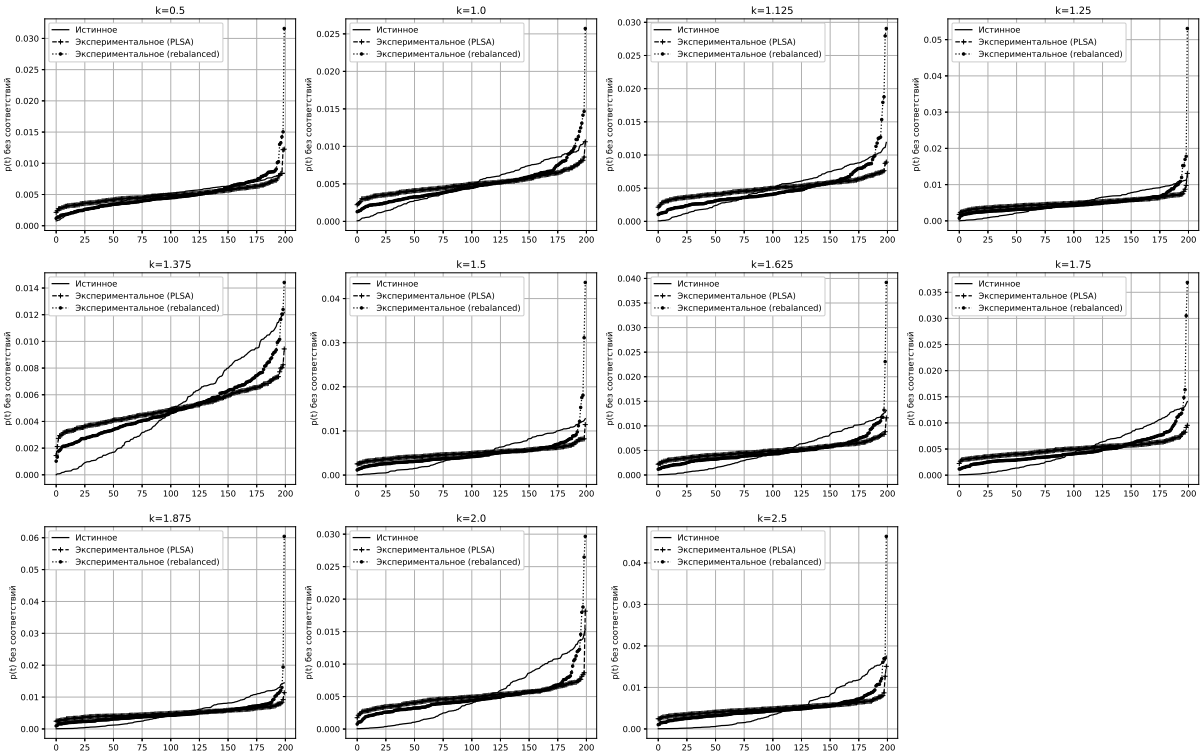

Рис. 11: Распределение p(t) для небалансированной и балансированной моделей без регуляризаторов $T = 200$, построенные темы отсортированы в порядке неубывания мощности

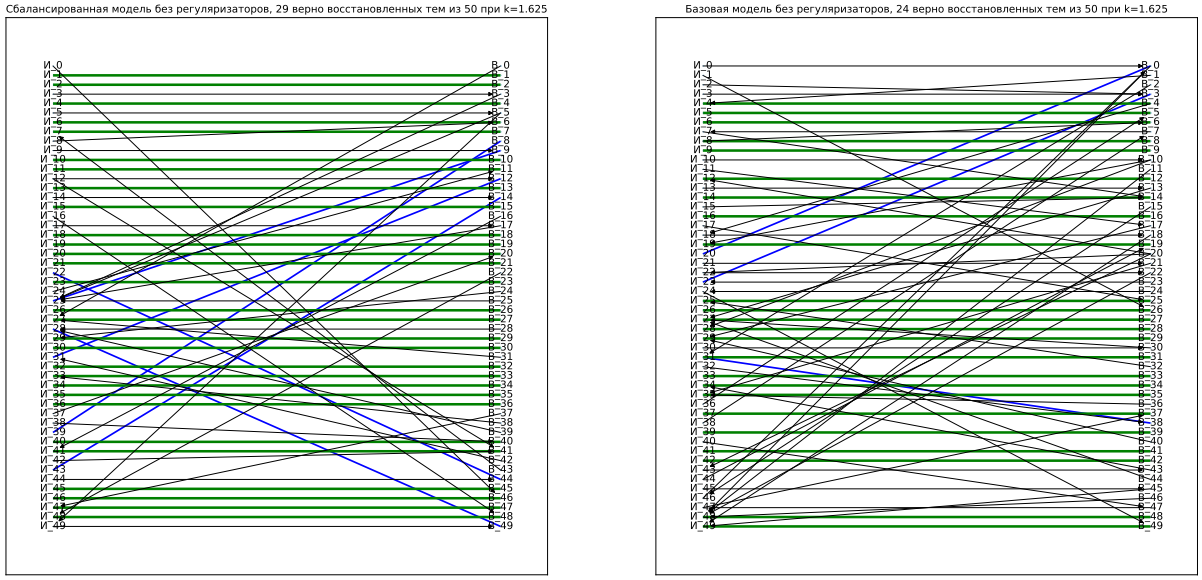

Рис. 12: Граф соответствий тем при $T = 50, k = 1.625$

на Рис.13—Рис.15

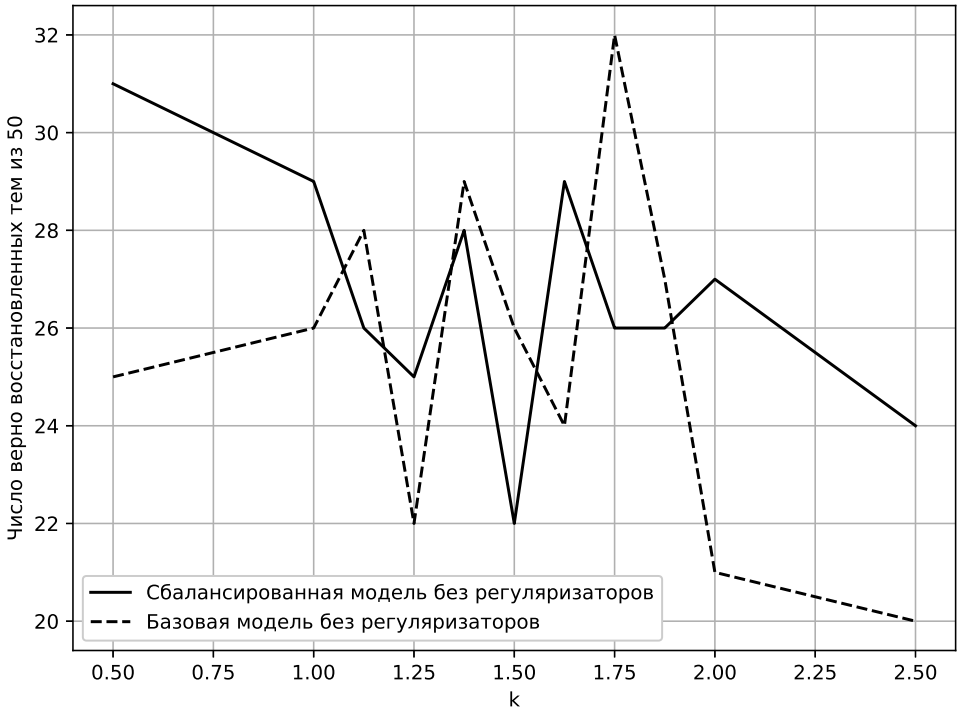

Рис. 13: Зависимость количества верно восстановленных тем от k при $T = 50$

Как видно из этих графиков, сбалансированная модель дает лучшие результаты, чем модель без модификаций, при больших значениях параметра разбалансированности $k \geqslant 2$. Также, сбалансированная модель является более устойчивой: дисперсия результата у нее меньше, чем у базовой модели.

3.2.2 Разреживающий регуляризатор

Теперь добавим в базовую и несбалансированную модель PLSA разреживающий матрицу Φ регуляризатор. Это достигается выбором отрицательного коэффициента регуляризации $\tau = -5 \cdot 10^4$.

Обычно, добавление такого регуляризатора улучшает интерпретируемость получаемых тем.

Сначала обучаем модели 2 итерации без регуляризатора, затем добавляем регуляризатор и дообучаем модель еще на 8 итерациях до сходимости перплексии. Строим графики зависимостей, аналогичные графикам из предыдущего пункта.

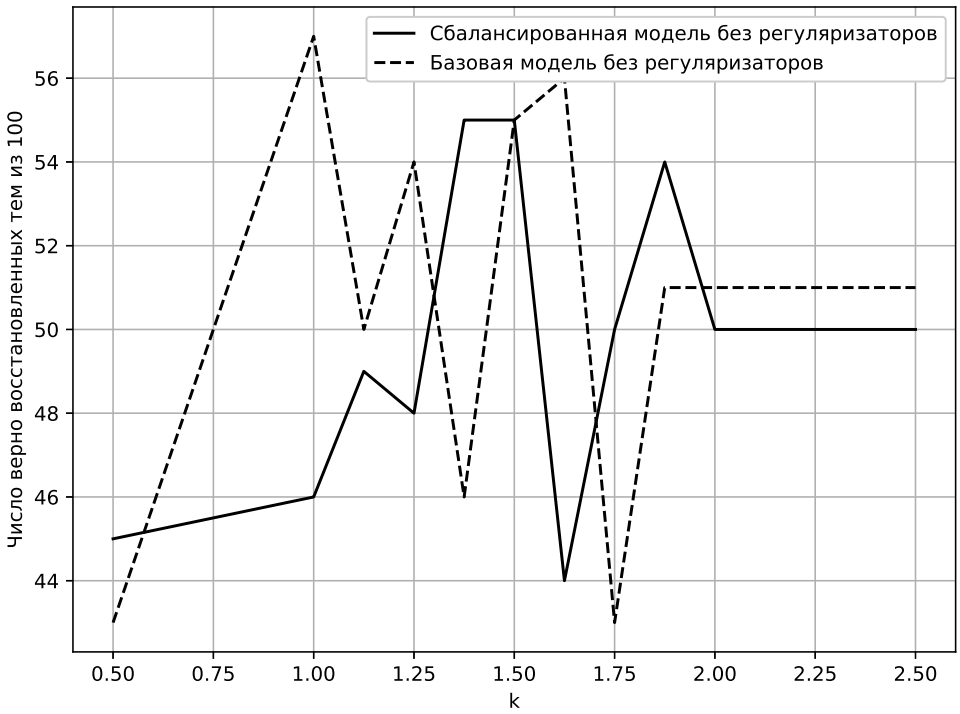

Рис. 14: Зависимость количества верно восстановленных тем от k при $T = 100$

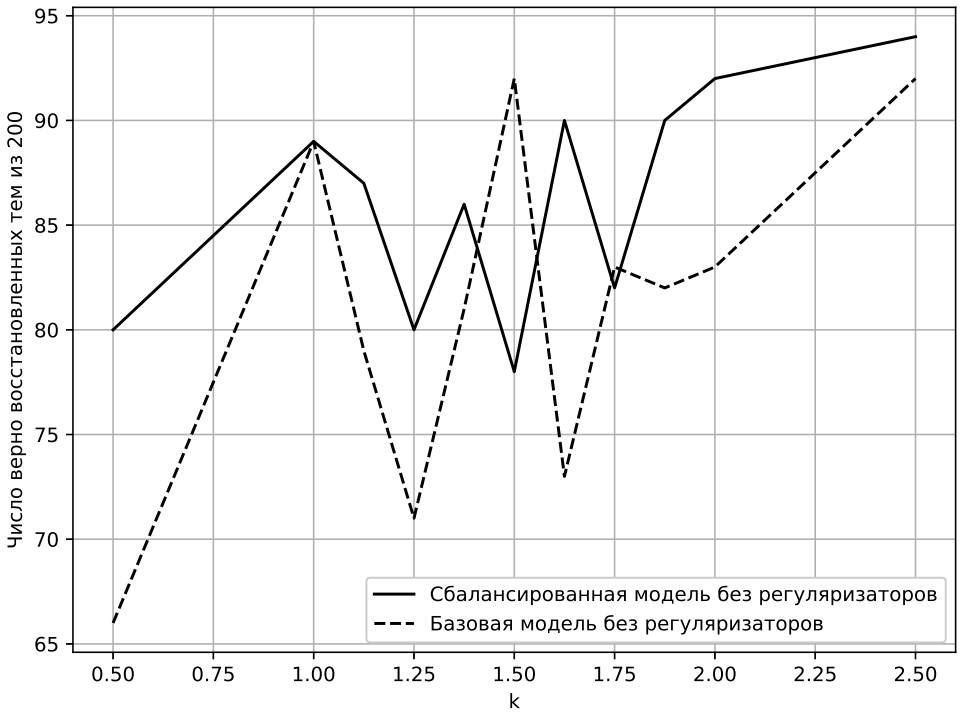

Рис. 15: Зависимость количества верно восстановленных тем от k при $T = 200$

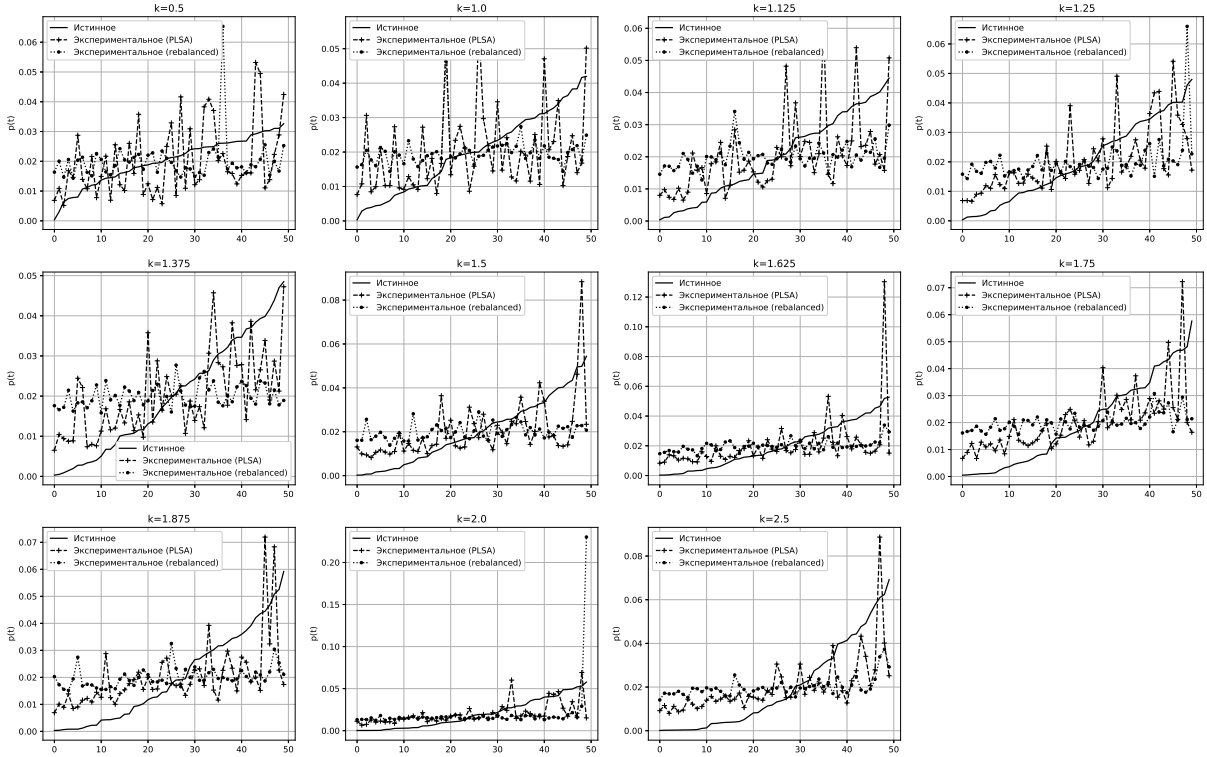

Рис. 16: Распределение p(t) для небалансированной и балансированной моделей с разреживающим регуляризатором $T = 50$, построенные темы были соотнесены с истинными темами

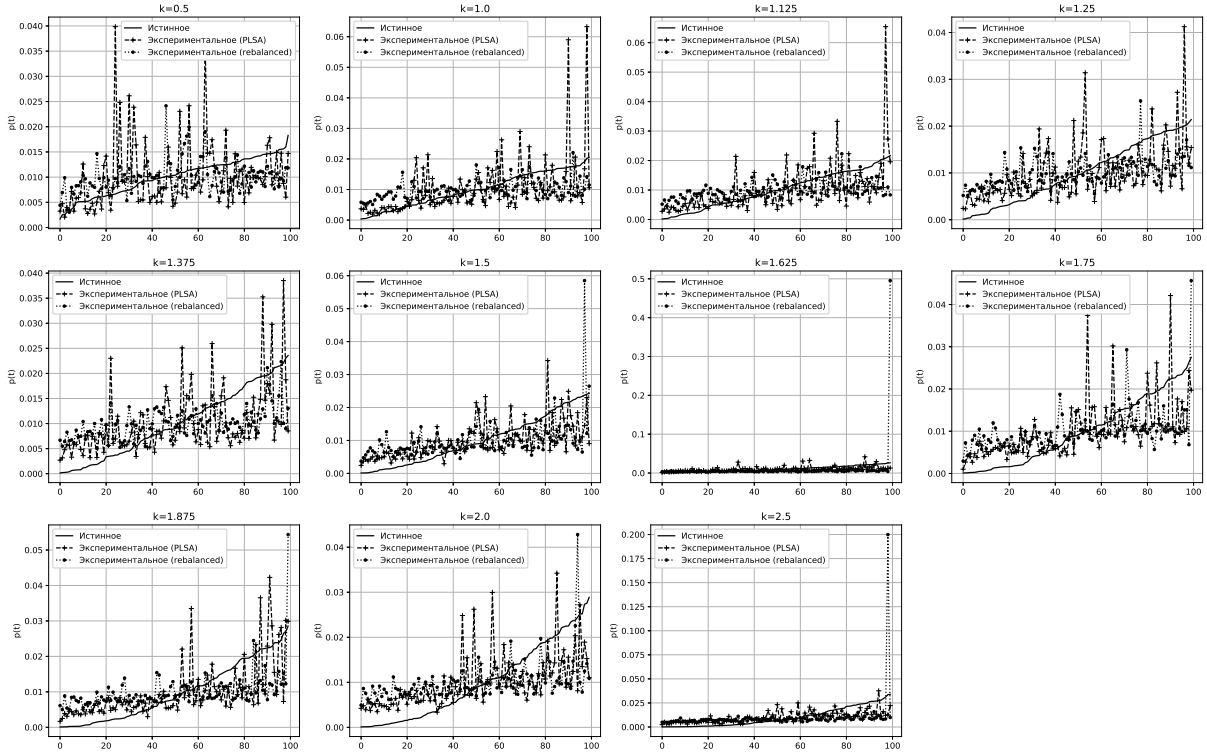

Рис. 17: Распределение p(t) для небалансированной и балансированной моделей с разреживающим регуляризатором $T = 100$, построенные темы были соотнесены с истинными темами

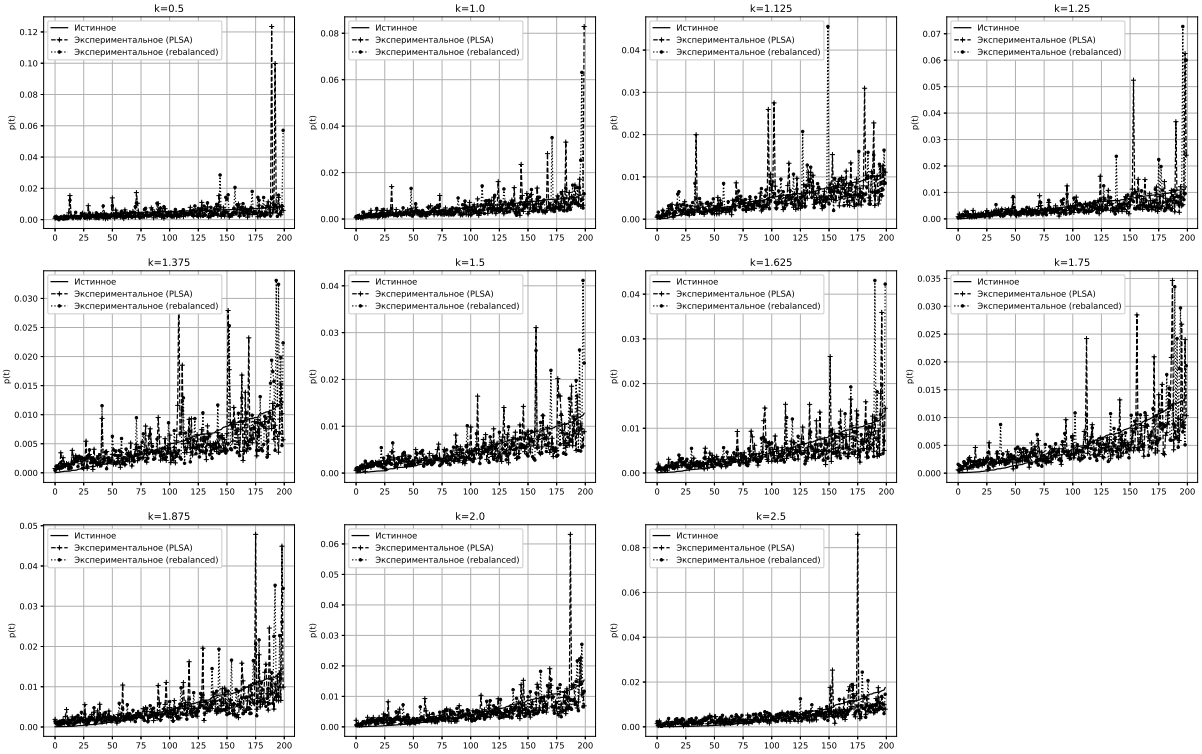

Рис. 18: Распределение p(t) для небалансированной и балансированной моделей с разреживающим регуляризатором $T = 200$, построенные темы были соотнесены с истинными темами

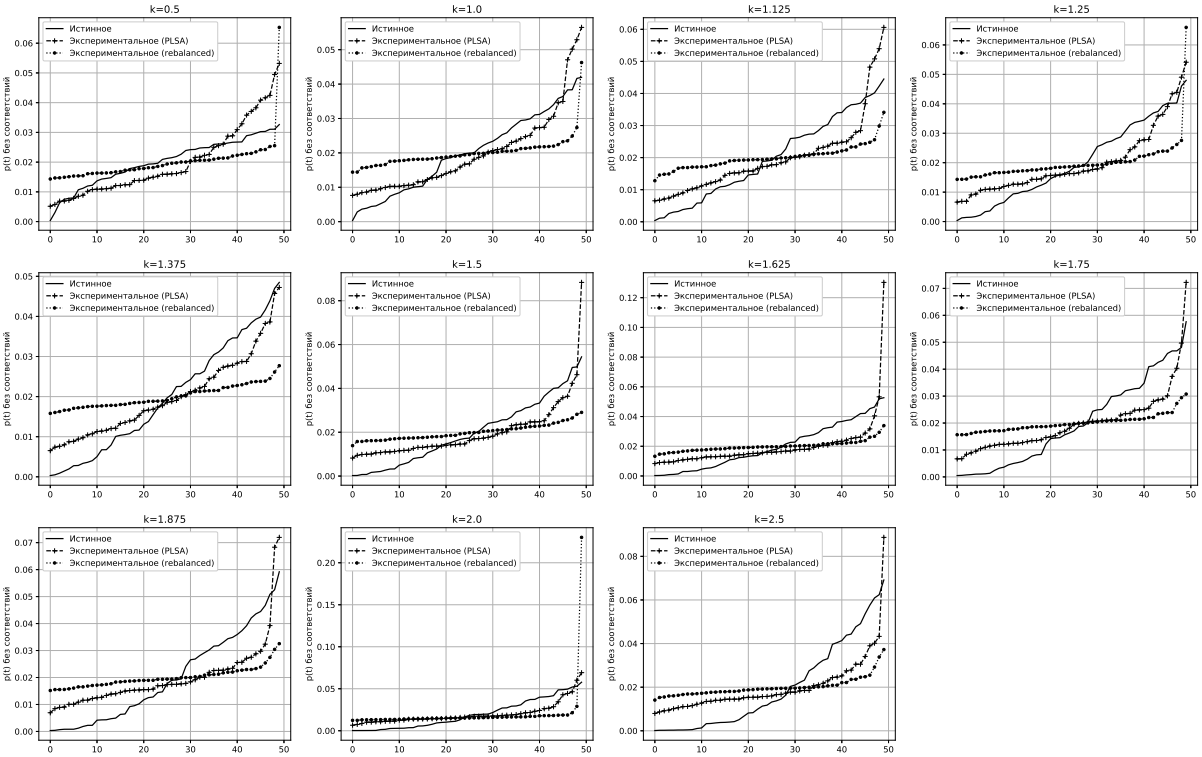

Рис. 19: Распределение p(t) для небалансированной и балансированной моделей с разреживающим регуляризатором $T = 50$, построенные темы отсортированы в порядке неубывания мощности

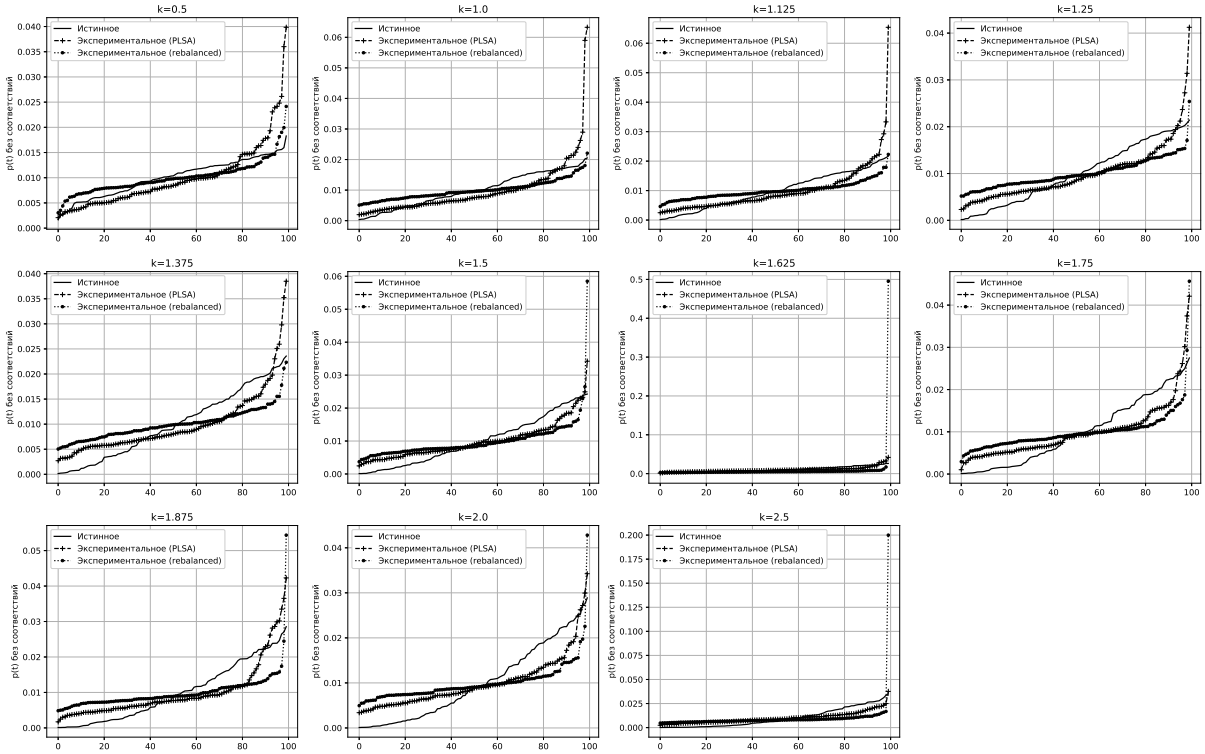

Рис. 20: Распределение p(t) для небалансированной и балансированной моделей с разреживающим регуляризатором $T = 100$, построенные темы отсортированы в порядке неубывания мощности

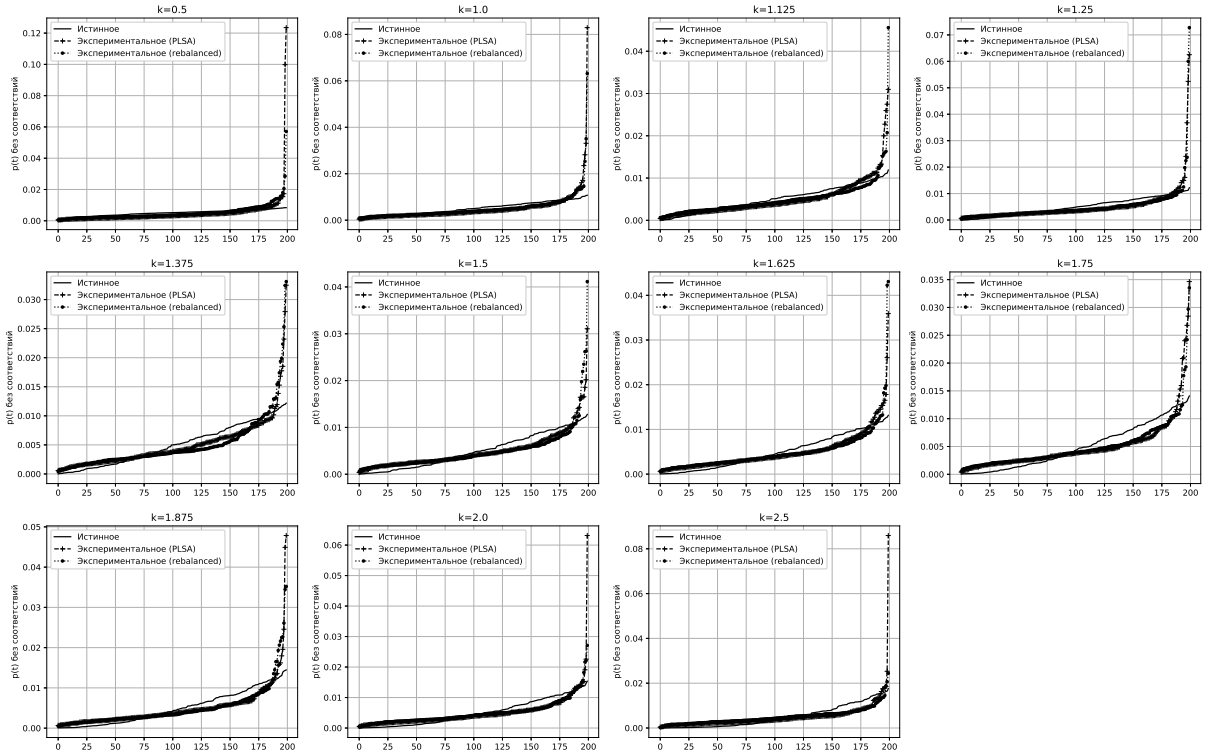

Рис. 21: Распределение p(t) для небалансированной и балансированной моделей с разреживающим регуляризатором $T = 200$, построенные темы отсортированы в порядке неубывания мощности

На Рис.16—Рис.21 построено распределение $p(t)$ для несбалансированной и балансированной моделей с разреживающим регуляризатором. Как видно, восстанавливаемая $p(t)$ качественно не изменилась относительно предыдущего пункта

Количество верно восстановленных тем (т.е таких, которые были признаны взаимно ближайшими, причем венгерский алгоритм их тоже соотнес вместе) в зависимости от коэффициентов несбалансированности k для различных T можно видеть на Рис.22—Рис.24

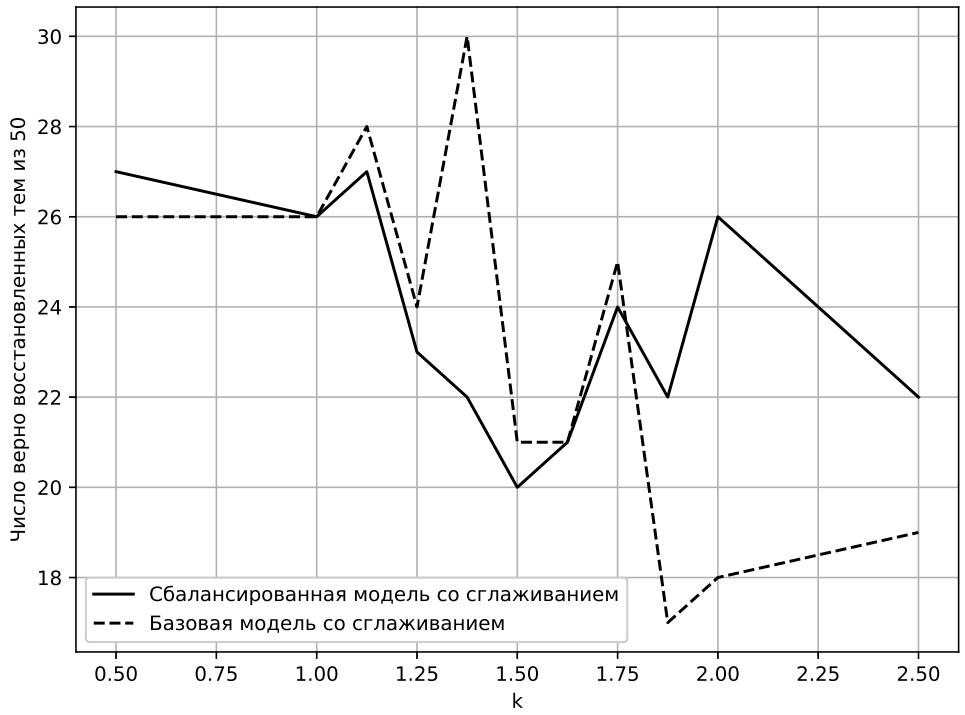

Рис. 22: Зависимость количества верно восстановленных тем от k при $T = 50$ (использовано разреживание)

3.3 Использование алгоритмов кластеризации

Теперь проверим эффективность использования алгоритмов кластеризации для решения проблемы несбалансированности классов.

В этом эксперименте для каждого коэффициента несбалансированности k сперва построим соответствующую тематическую модель на $T' = \left\lceil \sum_{i=0}^{T} k^i \right\rceil$ темах. Далее, для построенных распределений тем $\hat{\Theta}$ применим алгоритм кластеризации $DBSCAN$

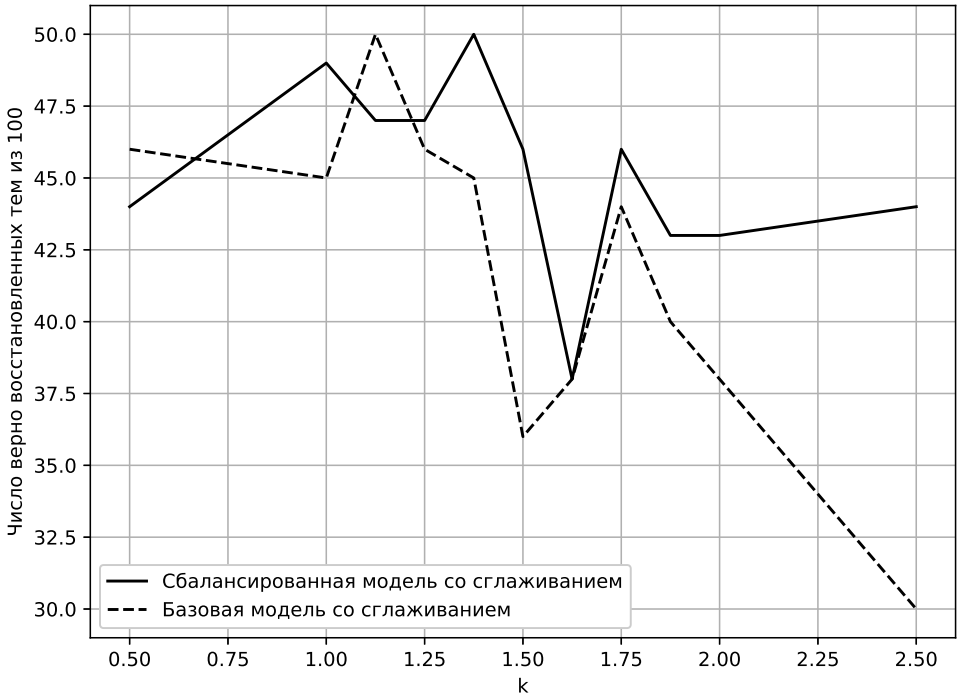

Рис. 23: Зависимость количества верно восстановленных тем от k при $T = 100$ (использовано разреживание)

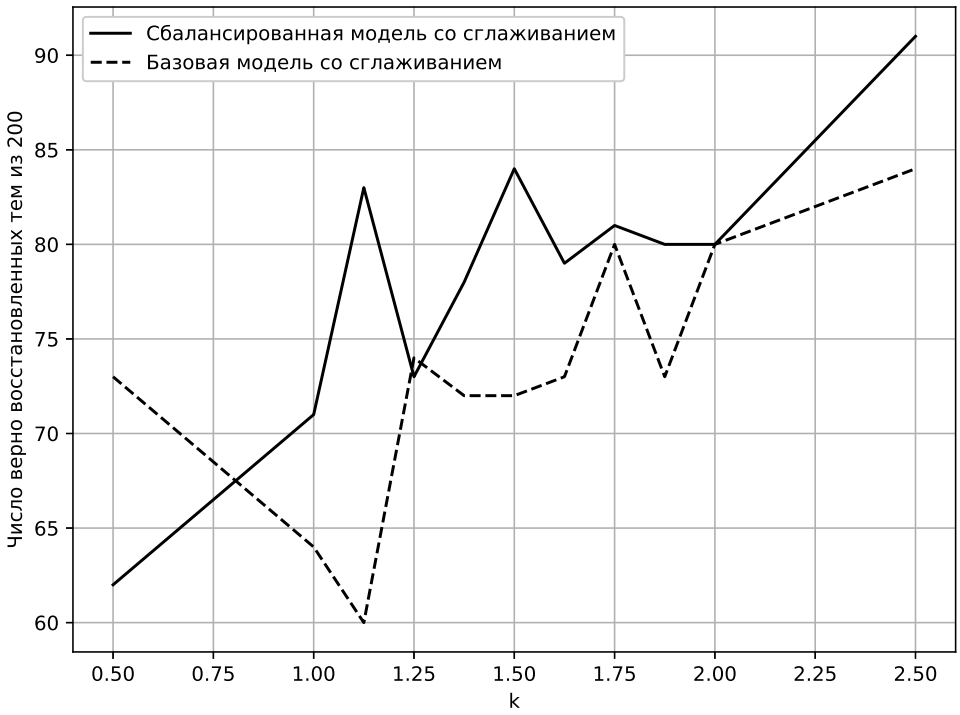

Рис. 24: Зависимость количества верно восстановленных тем от k при $T = 200$ (использовано разреживание)

и выделим T кластеров, подбирая параметр ε.

После разбиения на кластеры, составим матрицы «распределения» тем над документами, применив one-hot encoding. Применив венгерский алгоритм, получим соответствие между истинными и построенными темами, переставим строки матрицы $\hat{\Theta}'$ в соответствии с исходными темами.

На Рис.25 построено распределение $p(t)$ для несбалансированной и балансированной моделей с разреживающим регуляризатором и алгоритмом кластеризации при $T = 50$.

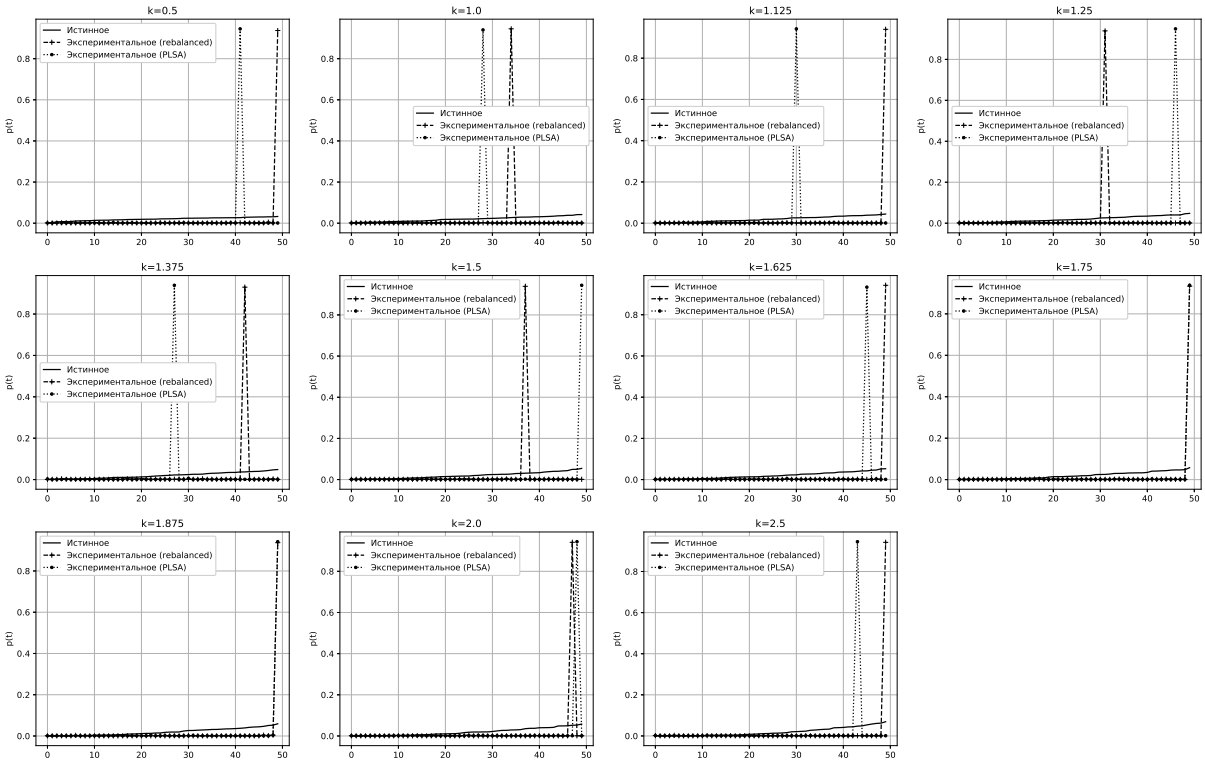

Рис. 25: Распределение p(t) для небалансированной и балансированной моделей с кластеризацией при $T = 50$, построенные темы были соотнесены с истинными темами

В обоих случаях, алгоритм выделяет по одному документу в представители каждого из $T-1$ классов-тем, остальные $N-T+1$ документа алгоритм выделяет в оставшуюся тему. Стоит отметить, что в большинстве случаев алгоритм с балансировкой тем выделяет большую истинную тему в качестве самой мощной, чем алгоритм без балансировки.

Количество верно восстановленных документов в зависимости от k показано на Рис.26. Как видно, сбалансированная модель справляется лучше, чем модель без балансировки, но их результаты все равно гораздо хуже, чем при неиспользовании кластеризации.

Это понятно, как видно, например, на Рис27, большинство истинных тем были слиты в одну сгенерированную, поэтому процент верно восстановленных тем очень мал. Все графы соответствий есть в Приложении

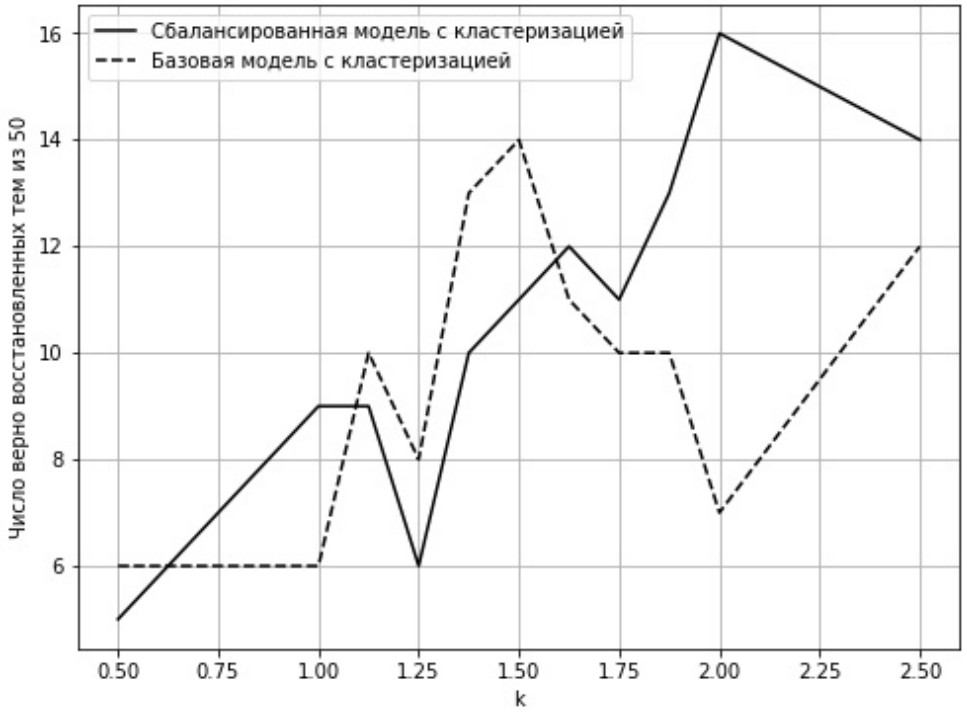

Рис. 26: Зависимость количества верно восстановленных тем от k при $T = 50$ (использована кластеризация)

4 Выводы

Экспериментально была проиллюстрирована и изучена проблема несбалансированности тем. Было изучено влияние регуляризатора разреживания, итерационного алгоритма и кластеризации как в отдельности, так и совместно. Несмотря на то, что исследуемые методы практически не решают проблему, было получено улучшение построения несбалансированных тем, используя предложенный итеративный алгоритм совместно с разреживающим регуляризатором. Однако, рассматриваемая проблема требует дальнейшего изучения посредством других методов.

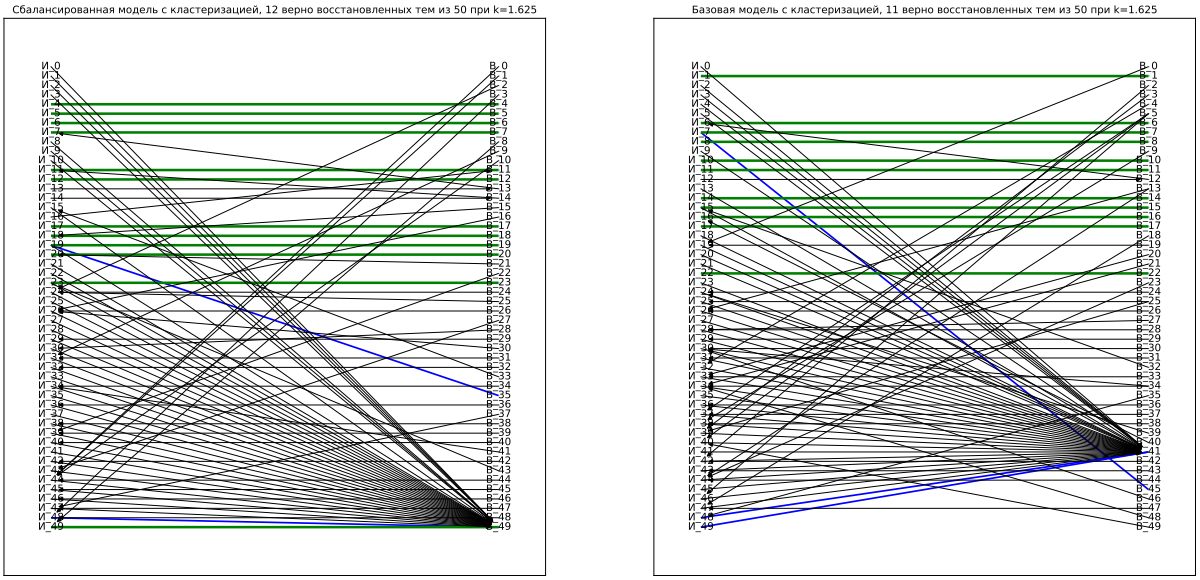

Рис. 27: Граф соответствий тем при $T = 50, k = 1.625$, была использована кластеризация

4.1 Результаты, выносимые на защиту

- Показано, что тематическое моделирование сталкивается с проблемой несбалансированности тем
- Предложен итеративный алгоритм для балансировки тем
- Изучена работа итеративного алгоритма совместно с регуляризацией и подходом кластеризации, но эти методы не решают проблему полностью

Список литературы

[1] Ф.А., . Применение мультимодальных тематических моделей к анализу транзакционных данных банков / Никитин Ф.А., Воронцов К.В., Матвеев И.А. — http://www.machinelearning.ru/wiki/images/b/bd/Nikitin18bsc.pdf.

[2] Н., . . Методы решения некорректных задач / Тихонов А. Н., Арсенин В. Я // Наука. — 1986.

[3] Blei, D. M. Latent dirichlet allocation / David M. Blei, Andrew Y. Ng, Michael I. Jordan // J. Mach. Learn. Res. — 2003. — Mar. — Vol. 3. — P. 993–1022. — http://dl.acm.org/citation.cfm?id=944919.944937.

[4] Dempster, A. P. Maximum likelihood from incomplete data via the EM algorithm / A. P. Dempster, N. M. Laird, D. B. Rubin // J. of the Royal Statistical Society, Series B. — 1977. — no. 34. — P. 1–38.

[5] Fast and modular regularized topic modelling / D. Kochedykov, M. Apishev, L. Golitsyn, K. Vorontsov // 2017 21st Conference of Open Innovations Association (FRUCT). — 2017. — Nov. — P. 182–193.

[6] Hofmann, T. Probabilistic latent semantic indexing / Thomas Hofmann // Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. — SIGIR '99. — New York, NY, USA: ACM, 1999. — P. 50–57. — http://doi.acm.org/10.1145/312624.312649.

[7] N., T. A. Solution of ill-posed problems / Tikhonov A. N., Arsenin V. Y // W. H. Winston, Washington, DC. — 1977.

[8] Non-bayesian additive regularization for multimodal topic modeling of large collections / Konstantin Vorontsov, Oleksandr Frei, Murat Apishev et al. // Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications. — TM '15. — New York, NY, USA: ACM, 2015. — P. 29–37. — http://doi.acm.org/10.1145/2809936.2809943.

[9] Vorontsov, K. Bigartm: Open source library for regularized multimodal topic modeling of large collections / Konstantin Vorontsov, Alexander Frei,

Murat Apishev // International Conference on Analysis of Images, Social Networks and Texts. — 2015. — P. 370–381.

[10] Vorontsov, K. Additive regularization of topic models for topic selection and sparse factorization / Konstantin Vorontsov, Anna Potapenko, Alexander Plavin // Machine Learning. — 2014. — machinelearning.ru/wiki/images/e/e3/Voron15slds.pdf.