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 (бакалавриат)
Направленность (профиль) подготовки: Компьютерные технологии и интеллектуальный анализ данных

ПРОВЕРКА ГИПОТЕЗЫ УСЛОВНОЙ НЕЗАВИСИМОСТИ ДЛЯ ОЦЕНИВАНИЯ КАЧЕСТВА ТЕМАТИЧЕСКОЙ КЛАСТЕРИЗАЦИИ

(бакалаврская работа)

Студент:
Рогозина Анна Андреевна

_____
(подпись студента)

Научный руководитель:
Воронцов Константин Вячеславович,
д-р физ.-мат. наук

_____
(подпись научного руководителя)

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

Москва 2019

Аннотация

Рассматривается процесс построения тематической модели по текстовой коллекции документов. В работе предлагается метод оценки качества тем, построенных тематической моделью. Тематическое моделирование рассматривается как мягкая классификация документов по множеству тем, и качество тем оценивается с точки зрения качества мягкой классификации. Для определения принадлежности документов к кластеру тем проверяется гипотеза условной независимости с помощью семейства дивергенций Кресси-Рида. Проведены эксперименты, исследующие поведение предложенных мер качества для различного количества тем в моделях, различного количества итераций EM – алгоритма, проведенных для обучения модели. Так же показано влияние регуляризатора декоррелирования на предложенные меры качества.
Ключевые слова: тематическое моделирование; кластеризация; гипотеза условной независимости, статистика Кресси-Рида.

Содержание

1 Введение

Задачу определения тематики текста имеет множество практических приложений: анализ мнений в отзывах[1], выявление трендов в научных статьях[2], классификация последовательностей ДНК[3], векторные представления слов [4]. Один из способов определения тематики текста - тематическое моделирование. Вероятностное тематическое моделирование определяет набор тем в коллекции, для каждого документа в коллекции определяет дискретное распределение тем в документе $p(t \mid d)$, и для каждой темы - дискретное распределение слов в этой теме $(w \mid t)$.
Вероятностная модель тематического модели опирается на гипотезу условной независимости: предполагается, что распределения слов темы t во всех документах d совпадают с общим распределением $p(w \mid t)$ и не зависят от документа d. В естественном языке такое предположение может не выполняться: например, из-за явления повторяемости слов(word burstiness)[5]: если слово встретилось в тексте один раз, велика вероятность, что оно встретится в тексте еще раз. Это происходит потому, что, несмотря на наличие множества синонимов в теме, автор часто выбирает один предпочтительный термин(или небольшое множество терминов) и использует только их на протяжении всего написания текста. В работе приводится способ оценки выполнимости гипотезы условной независимости для коллекции D и построенной по ней тематической модели (Φ, Θ). На основе оценки выполнимости гипотезы условной независимости в работе предлагается критерий оценивания качества построенных тем.
Для оценивания качества тем в тематическом моделировании считается, например, когерентность[6] тем. Однако в большинстве случаев качество тем определяется по *топ-словам темы* - набору $|U|$ первых k слов из отсортированного по убыванию вектора $p(w \mid t)$. Таким образом, чтобы оценить качество тем, необходимо проверить набор топ-слов для каждой темы на интерпретируемость, однородность и убедиться, что эти наборы различны по смыслу для разных тем. Такой подход становится неэффективным, если мощность множества тем $|T|$ составляет несколько десятков. Такой подход становится невозможным, если коллекция документов написана на неизвестном вам языке, или словарь вообще не является множеством слов(например, множество кодов).
Предложенный в работе критерий оценивает не только выполнимость гипотезы условной независимости в коллекции, но и качество тематической классификации: насколько хорошо построенные темы описывают тематику коллекции и насколько темы отличаются друг от друга. Таким образом, появляется количественная характеристика качества темати-

ческой кластеризации, что избавляет от необходимости просматривать набор топ-слов каждой темы.

2 Постановка задачи

2.1 Задача тематического моделирования и гипотеза условной независимости

Пусть D — коллекция документов, W — множество токенов(слов или словосочетаний). Каждый документ $d \in D$ представляет собой последовательность n_d терминов $(w_1, \ldots, w_{nd})$ из словаря W. Обозначим частоту встречаемости слова w в документе d как n_{wd} Задача тематического моделирования основана на следующих предположениях[7]:

1. *Возможность разделения на темы*: предполагается, что существует определенный набор тем $T : (t_1, \ldots, t_T)$, для которого каждое слово в документе относится какой-то теме $t \in T$. Таким образом, коллекция D представляет множество троек (t, d, w), выбранных случайно и независимо из дискретного распределения $p(t, d, w)$ на множестве $|T| \times |D| \times |W|$. Слова w и документы d являются наблюдаемыми переменными,
   темы t — латентными.

2. *Гипотеза «мешка слов»* предполагает, что тематика документа описывается лишь частотой встречаемости слов в документе n_{wd}, но не их порядком. Тематика документа сохраняется даже при произвольной перестановке слов в документе. Порядок документов в коллекции так же неважен.

3. *Гипотеза условной независимости* заключается в предположении, что распределения слов, относящихся к теме t в документе d совпадают сраспределением слов
   в теме t, $p(w \mid t, d) = p(w \mid t)$

Сделанные предположения позволяют записать распределение слов в документе через распределение слов в теме в компактной форме: $p(w \mid d) = \sum_t p(w \mid t)p(t \mid d)$
Задача тематического моделирования заключается в нахождении по известным $p(w \mid d) = \frac{n_{wd}}{n_d}$ множества тем T, дискретных распределений $p(w \mid t)$ слов в теме и дискретных распределений тем в документе $p(t \mid d)$ для всех $d \in D$, $w \in W$, $t \in T$.
Обозначим за Φ матрицу $w \times t$, в которой каждый элемент φ_{wt} равен вероятности слова w в теме t, $\varphi_{wt} = p(w \mid t)$ и за Θ матрицу $t \times d$, в которой каждый элемент θ_{td} равен вероятности встретить тему t в документе d, $\theta_{td} = p(t \mid d)$. Запишем правдоподобие выборки, применив новые

обозначения и предположения (1 - 3):

$$\mathscr{L}((d_i, w_i)_{i=1}^n, \Phi, \Theta) = \prod_{i=1}^{n} p(d_i, w_i) = \prod_{d\in D}\prod_{w\in d} p(w \mid d)^{n_{wd}} p(d)^{n_{wd}} =$$

$$= \prod_{d\in D}\prod_{w\in d} (\sum_t \varphi_{wt}\theta_{td})^{n_{wd}} p(d)^{n_{wd}} \to \max_{\varphi,\Theta} \tag{2.1}$$

Учитывая, что член $p(d)^{n_{wd}}$ является константой и не зависит от параметров модели и логарифмируя правдоподобие, получаем следующую задачу минимизации с ограничениями:

$$\sum_{d\in D}\sum_{w\in d} n_{wd} \ln(\sum \varphi_{wt}\theta_{td}) \to \max_{\Phi,\Theta} \tag{2.2}$$

$$\sum_{w\in W} \varphi_{wt} = 1; \;\; \varphi_{wt} \geq 0 \qquad\qquad \sum_{t\in T} \theta_{td} = 1; \;\; \theta_{td} \geq 0 \tag{2.3}$$

Добавление регуляризаторов.
Задача тематичекого моделирования имеет бесконечное количество решений: Если (Φ, Θ) —- решение задачи (2.1).Тогда возьмем любую невырожденную матрицу S, и $(\Phi S, S^{-1}\Theta)$ — тоже решение задачи (2.2) при условии, что ограничения (2.3) на новые матрицы сохраняются. Чтобы уменьшить множество решений и наделить его новыми полезными свойствами, вводят *регуляризацию*. В общем случае вводится набор регуляризаторов $R_i(\Phi, \Theta)$ и коэффициентов регуляризации $\tau_i \geq 0$. Таким образом, задача минимизации записывается как:

$$\sum_{d\in D}\sum_{w\in d} n_{wd} \ln(\sum \varphi_{wt}\theta_{td}) + R(\Phi, \Theta) \to \max_{\Phi,\Theta}; \qquad R(\Phi, \Theta) = \sum_{i}^{k} \tau_i R_i(\Phi, \Theta), \tag{2.4}$$

с ограничениями (2.3)
В работе используются тематические модели без регуляризаторов(PLSA) и модели с *регулризатором декоррелирования*, для которого

$$R(\Phi, \Theta) = -\frac{\tau}{2}\sum_{t\in T}\sum_{s\in T\setminus t}\sum_{w\in W} \varphi_{wt}\varphi_{ws}. \tag{2.5}$$

2.2 Постановка задачи

Дана коллекция документов D. По этой коллекции построена тематическая модель (Φ, Θ). Построение тематической модели основывается на гипотезе условной независимости распределения слов w в теме t от документа d: $p(w \mid t, d) = p(w \mid t)$. Предлагается разработать автоматически вычислимые критерии, оценивающие, насколько для данной модели (Φ, Θ) и в данной коллекции D выполняется гипотеза условной независимости. Основываясь на этих критериях, предлагается оценить качество тем, построенных моделью (Φ, Θ) и качество тематической кластеризации в целом.

3 Проверка гипотезы условной независимости

3.1 Дивергенция Кресси-Рида

Дана выборка $X = \{x_1, \ldots, x_n\}$ реализаций независимы одинаково распределенных случайных величин, принимающих значения из конечного множества Ω. Проверяется гипотеза о том, что данная выборка X была получена из известного нам распределения $p(x)$:

$$\begin{aligned} H_0 &: X = \{x_1, \ldots, _n\} \in p(x) \\ H_1 &: X = \{x_1, \ldots, _n\} \notin p(x) \end{aligned} \tag{3.1}$$

Критерии, проверяющие гипотезу о равенстве распределений, называются *критериями согласия*. К таким, например, относится критерий Хи-квадрат Пирсона, дивергенция Кульбака–Лейблера, расстояние Хеллингера. Все они являются частым случаем семейства *дивергенций Кресси-Рида*[8] между двумя распределениями:

$$CR_\lambda\big(\hat{p}(w\,|\,d,t) : \hat{p}(w\,|\,t)\big) = \frac{2n_{td}}{\lambda(\lambda+1)} \sum_{w\in W} \hat{p}(w\,|\,d,t)\left(\left(\frac{\hat{p}(w\,|\,d,t)}{\hat{p}(w\,|\,t)}\right)^{\lambda} - 1\right) = \tag{3.2}$$

$$= \frac{2}{\lambda(\lambda+1)} \sum_{w\in W} n_{tdw}\left(\left(\frac{n_{tdw}n_t}{n_{td}n_{wt}}\right)^{\lambda} - 1\right).$$

При $\lambda = 1$ дивергенция Кресси-Рида преходит в статистику хи-квадрат Пирсона, при $\lambda \to 0$ в дивергенцию Кульбака–Лейблера, при $\lambda = -\frac{1}{2}$ - в расстояние Хеллингера. Все эти статистики в условии истинности нулевой гипотезы асимптотически стремятся к распределению χ^2 с $k = |\Omega| - 1$ степенями свободы $_\lambda \sim \chi^2(k)$

3.2 Симметричность статистики Кресси-Рида

Исследуем симметричность статистики Кресси-Рида в зависимости от параметра λ.

Утверждение

Статистика Кресси-Рида симметрична только для $\lambda = -\frac{1}{2}$ (Расстояние Хеллингера)

Доказательство. Статистика симметрична, если $\mathrm{CR}_\lambda(p,q) = \mathrm{CR}_\lambda(q,p)$, или

$$\frac{2n_{td}}{\lambda(\lambda+1)} \sum_{w \in W} p_w \left(\left(\frac{p_w}{q_w} \right)^\lambda - 1 \right) = \frac{2n_{td}}{\lambda(\lambda+1)} \sum_{w \in W} q_w \left(\left(\frac{q_w}{p_w} \right)^\lambda - 1 \right)$$

Заметим, что так как $\sum_{w \in W} p_w = \sum_{w \in W} q_w = 1$(это дискретные распределения), то достаточно проверить, при каких λ

$$\sum_{w \in W} p_w \left(\frac{p_w}{q_w} \right)^\lambda = \sum_{w \in W} q_w \left(\frac{q_w}{p_w} \right)^\lambda$$

Так как это равенство должно выполняться для любых (p,q), необходимо, чтобы $\forall w \in W$

$$p_w \left(\frac{p_w}{q_w} \right)^\lambda = q_w \left(\frac{q_w}{p_w} \right)^\lambda$$

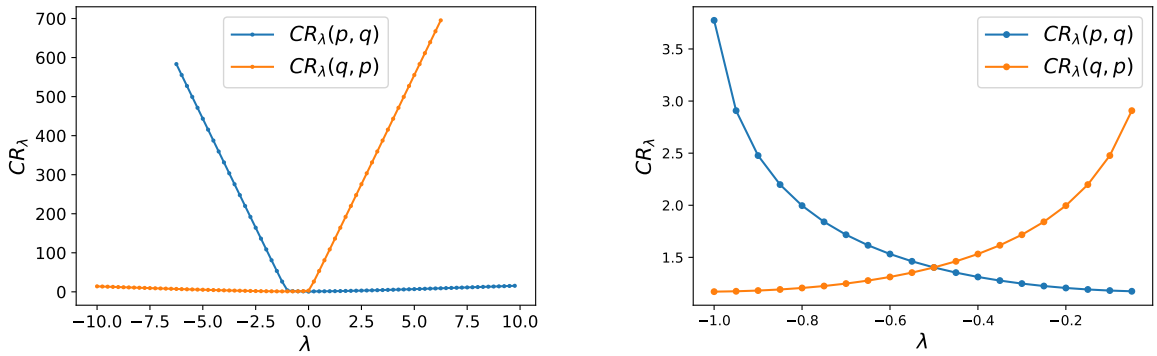
Не умаляя общности, считаем, что $p_w, q_w > 0$. Приводим подобные слагаемые, получаем $p_w^{2\lambda+1} = q_w^{2\lambda+1}$. Соответственно, $2\lambda+1 = 0$ и единственная точка симметричности - $\lambda = -\frac{1}{2}$
(При сокращении коэффициента $\frac{2}{\lambda(\lambda+1)}$ в этом доказательстве упущены случаи $\lambda = 0$ - KL-дивергенция, не симметричная, и $\lambda = -1$ - модифицированная статистика Пирсона, не симметричная).

□

На рисунке (1) в качестве иллюстрации к утверждению приведена зависимость $CR_\lambda(p,q)$ и $CR_\lambda(q,p)$ от λ, где (p,q) - нормальные распределения и p вложено в q. По данным графикам также можно предположить, что для $\lambda > -\frac{1}{2}$ статистика $CR_\lambda(p,q)$ измеряет вложенность распределения p в q, а для $\lambda < -\frac{1}{2}$ статистика $CR_\lambda(p,q)$ измеряет, наоборот, вложенность q в p.

3.3 Применимость для разреженных распределений

Асимптотика χ^2 применима для проверки равенства распределений, если размер выборки ≥ 50 и наблюдений $np(x) \geq 5$ для всех $x \in \Omega$. Если же вероятности $p(x)$ малы для многих x или $|\Omega| \gg n$, асимптотика не выполняется. Распределения слов в теме $p(w \mid t)$ и слов в документе $p(w \mid t,d)$ разреженные, так как размер словаря как правило гораздо больше длины документа $|W| \gg n$, кроме того, $p(w)$ мала для многих w, поэтому асимптотика χ^2 неприменима для сравнения распределений слов. Необходимо ослабить статистические тесты.



a)Общий график b)Диапазон $-1 \leq \lambda \leq 0$

Рис. 1: Симметричность статистики Кресси-Рида в зависимости от λ

3.3.1 Возможные приближения

В качестве сужения множества альтернатив предлагается вместо проверки гипотезы для всех слов W ввести некоторую группировку слов U, и проверять гипотезу уже для векторов, получившихся в качестве такой группировки. Таким образом увеличиваются вероятности $p(x)$, а так же количество разбиений для каждого наблюдения $np(x)$. Однако, такой способ оказывается неустойчивым, так как результаты сильно зависят от способа группировки, выбираемого произвольно. Предлагается также фильтровать словарь и проводить тесты для вектора из слов, относящихся к теме t, игнорируя нетематические слова, вероятность встретить которые в этой теме меньше равномерного распределения, $p(w \mid t) < \frac{1}{W}$. Кроме того, предлагается проводить тесты равенства $p(w \mid t, d)$ и $p(w \mid t)$ только для слов, которые встретились в документе d.

3.3.2 Эмпирическое распределение статистики Кресси-Рида

Для проверки равенства распределений $\hat{p}(w \mid t, d)$ и $p(w \mid t)$ на уровне значимости α необходимо вычислить $(1-\alpha)$ квантиль распределения статистики Креси-Рида CR_λ. Однако экспериментально показано, что распределение статистики Кресси-Рида для условных распределений слов в документах $p(w \mid t, d)$ в условиях истинности нулевой гипотезы зависит от длины документа n_{td}. Поэтому для вычисления $(1-\alpha)$ квантиля предлагается семплировать документы с разными значениями n_{td}, вычислять по ним эмпирическое распределение статистики Кресси-Рида S_{td} и проводить непараметрическую квантильную регрессию $(1-\alpha)$ квантиля от n_{td}. На рисунке (2) представлен пример построения непараметрической квантилоной регрессии по эмпирическому распределению

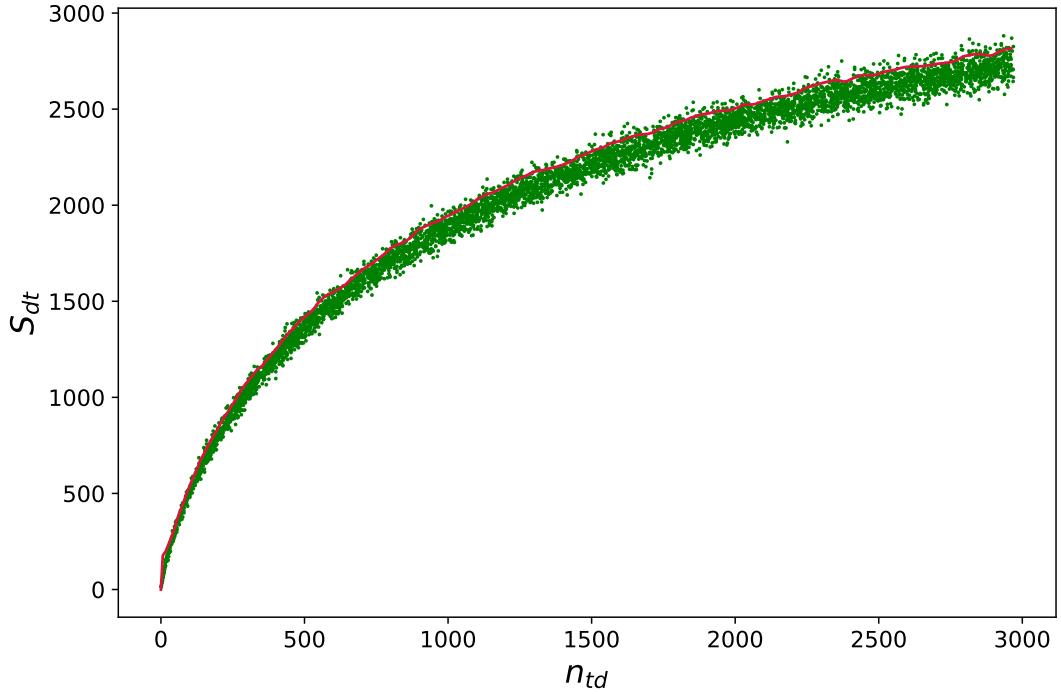


Рис. 2: Пример непараметрической регрессии $(1-\alpha)$ квантиля статистики CR_λ

статистики Кресси-Рида для одной из тем.

4 Оценка качества тематической кластеризации

Дана коллекция документов D. По коллекции построена тематическая модель Φ, Θ. Рассмотрим пространство дискретных распределений слов из словаря W, $p(w)$. В условиях истинности гипотезы условной независимости(раздел 2.1), для любой темы t и документа d, распределения $p(w \mid t)$ и $p(w \mid t, d)$ совпадают. Это означает, что в пространстве распределений $p(w)$ множество $p(w \mid t, d)$ представляет собой t точек, совпадающий с $p(w \mid t)$.
В действительности, в естественном языке гипотеза условной независимости не выполняется. Кроме того, нам доступны только частотные оценки распределений $p(w \mid t, d)$ (в дальнейшем обозначаются как $\hat{p}(w \mid t, d)$). Вместо гипотезы условной независимости вводится *гипотеза компактности*: предполагается, что для каждой темы t распределения $\hat{p}(w \mid t, d)$ представляют собой кластер, центром которого является распределение $p(w \mid t)$. Границы кластера оцениваются с помощью проверки гипотезы том, что эмпирическое распределение $\hat{p}(w \mid t, d)$ было сгенерировано из распределения $p(w \mid t)$.

Радиусом семантической неоднородности $R_t^\alpha(n_{td})$ темы t на уровне значимости α назовем $(1 - \alpha)$ квантиль распределения статистики Кресси-Рида $S_{dt} = CR_\lambda\big(\, \hat{p}(u \mid d, t) : \hat{p}(u \mid t)\big)$. Он показывает, насколько точка $p(w \mid d, t)$ может удалиться от центра кластера, не нарушая при этом нулевую гипотезу. Радиус семантической однородности зависит от размера выборки n_{td}, темы t и уровня значимости α.
Степенью семантической неоднородности темы t назовем взвешенную долю документов d, для которых значение статистики S_{td} больше радиуса семантической однородности $R_t^\alpha(n_{td})$.

$$\mathrm{SemH}(t) = \sum_{d \in D} \hat{p}(d \mid t)\big[S_{dt} < R_t^\alpha(n_{td})\big] = \sum_{d \in D} \frac{n_{td}}{n_t}\big[S_{dt} < R_t^\alpha(n_{td})\big], \quad (4.1)$$

Степень семантической неоднородности изменяется от 0 до 1 и показывает, какая доля точек кластера темы t находится за пределами радиуса семантической однородности и нарушает нулевую гипотезу. Если для темы t степень семантической неоднородности больше α, назовем ее *семантически неоднородной*.

Степенью семантической загрязненности темы t назовем долю документов d, для которых нулевая гипотеза не отвергается не только для темы t, но и еще для какой-то темы t':

$$\text{SemI}(t) = \sum_{d \in D} \hat{p}(d \mid t) \big[S_{dt} < R_t^{\alpha}(n_{td}) \big] \, \big[S_{dtt'} < R_{t'}^{\alpha}(n_{td}) \big], \qquad (4.2)$$

где дивергенция $S_{dtt'}$ измеряет расстояние от распределения $\hat{p}(w \mid d, t)$ до центра ближайшего чужого кластера $\hat{p}(w \mid t')$:

$$S_{dtt'} = \min_{t' \in T \setminus t} \text{CR}_{\lambda}\big(\hat{p}(w \mid d, t) : \hat{p}(w \mid t')\big). \qquad (4.3)$$

Степень семантической загрязнённости принимает значения от 0 до 1 и показывает, какая доля точек кластера относится также и к другим кластерам. Тему, в которой степень семантической загрязненности больше α, назовем *семантически загрязненной*.

На рисунке (3) показана иллюстрация к подсчету степеней загрязнен-

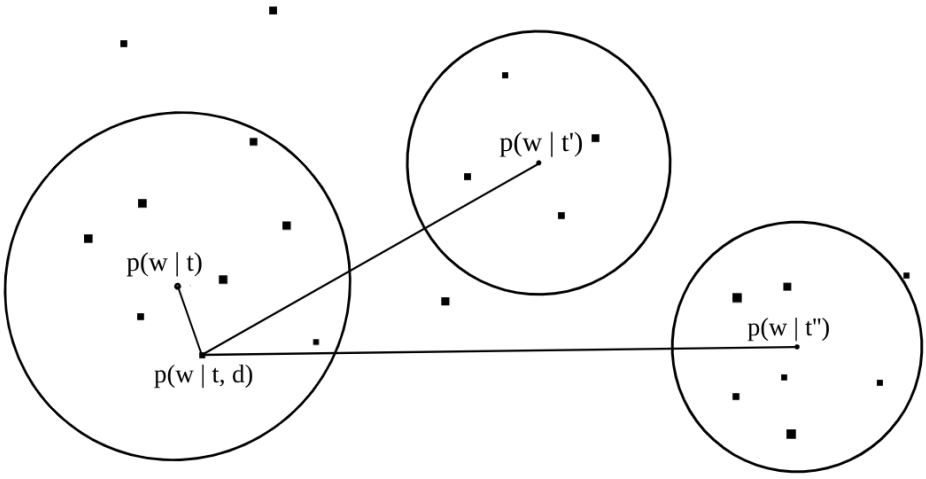


Рис. 3: Иллюстрация кластерной структуры распределений

ности и неоднородности: для подсчета степени неоднородности нужно найти расстояние $S_{td} = CR_{\lambda}\big(\hat{p}(w \mid d, t) : p(w \mid t)\big)$ и сравнить его с $R_t^{\alpha}(n_{td})$, а для подсчета степеней загрязненности нужно сравнить все $S_{t'd} = CR_{\lambda}\big(\hat{p}(w \mid d, t) : p(w \mid t')\big)$, выбрать минимальное расстояние $S_{t_{min}d}$ и сравнить его с $R_{t_{min}}^{\alpha}(n_{td})$.

Если тема не является ни семантически неоднородной, ни семантически загрязненной, назовем ее *сбалансированной*.

5 Эксперименты

5.1 Данные

В качестве данных берется коллекция документов из «Постнауки». Она содержит 3404 документа и состоит из небольших заметок на какую-то научно-популярную тему.

5.2 Сбалансированность

5.2.1 Общая постановка эксперимента

Необходимо для модели (Φ, Θ) и коллекции документов D определить несбалансированность тем, то есть для каждой темы t определить ее степень неоднородности и степень загрязненности. В качестве сужения множества альтернатив H_1 для применимости статистики Кресси-Рида к разреженным распределениям предлагается для каждого документа t и темы d выбирать подмножество слов U,
$\{U \subseteq W \ : \ \forall\, u \in U \ \ p(u\,|\,t) > \frac{1}{W},\, n_{tdu} > 0\}$, и считать
$S_{dt} = CR_\lambda\big(\hat{p}(u\,|\,d,t) : \hat{p}(u\,|\,t)\big)$. Ниже представлен алгоритм, вычисляющий SemH и SemI.

Эксперимент состоит из следующих частей:

1. Выбор способа группировки слов U для сужения множества альтернатив H_1 и значения λ в статистике Кресси-Рида.
   Предлагается в качестве сужения проверять гипотезу о равентсв распределений для подожества U слов W,
   $\{U \subseteq W : \forall u \in U p(u\,|\,T) \geq \frac{1}{W},\, n_{tdu} > 0\}$.

2. Выяснение зависимости радиуса семантической однородности $R_t^\alpha(n_{td})$ для всех тем t.

3. Подсчет степеней неоднородности и загрязненности для всех тем t.

5.2.2 Выбор параметра λ в статистике Кресси-Рида

Была обучена модель на 80 тем на «Постнауке» исследована зависимость средних SemH, SemI от параметра λ в статистике Кресси-Рида. На рисунке (4) представлена эта зависимость. Видно, что при $\lambda \geq 0$ степень загрязненности становится нерепрезентативной и практически нулевой, а при $\lambda \leq -1$ степень неоднородности становится практически 1, что эквивалентно стягиванию кластеров тем в точку. Кроме того, заметим,

Algorithm 5.1 Подсчет SemH и SemI для тематической модели

for $t \in T$ **do**
  Сгенерировать коллекцию документов D из $p(w\,|\,t)$ с различными n_{td}, получить $\{(n_{tdw}, n_{td})\}_{d \in D}$
  Преобразовать $(n_{tdw},\ p(w\,|\,t)) \to (n_{tdu},\ p(u\,|\,t))$,
  в которых $\forall\ u\ \in\ U:\ p(u\,|\,t \geq \frac{1}{W},\quad n_{tdu} \geq 0)$
  По $(n_{tdu},\ n_{td},\ p(u\,|\,t))$ построить непараметрическую квантильную регрессию $R_t^\alpha(n_{td}))$
for $t \in T$ **do**
  for $d \in D$ **do**
    $(n_{tdw}, p(w\,|\,t)) \to (n_{tdu},\ p(u\,|\,t))$
    Вычислить $S_{dt} = CR_\lambda\big(\hat{p}(u\,|\,d,t) : \hat{p}(u\,|\,t)\big))$
    Сравнить S_{dt} и $R_t^\alpha(n_{td}))$
    if $S_{dt} \leq R_t^\alpha(n_{td}))$ **then**
      for $t' \in T$ **do**
        Вычислить $S_{dtt'} = CR_\lambda\big(\hat{p}(u\,|\,d,t) : \hat{p}(u\,|\,t')\big))$
        Найти $t_{min} = argmin_{t' \neq t} S_{dtt'}$
        Сравнить $S_{dt_{min}}$ и $R^\alpha_{t_{min}}(n_{td}))$
  Вычислить SemH, SemI по формулам(2), (3)

что, при подсчете $CR_\lambda(\hat{p}(w\,|\,t,d)\ :\ p(w\,|\,t))$, необходимо измерять вложенность $\hat{p}(w\,|\,t,d)$ в $p(w\,|\,t)$, а не наоборот. Значит, как показано в разделе(3.2), необходимо брать $\lambda\ >\ -\frac{1}{2}$. Таким образом, рекомендуется выбирать $-\frac{1}{2}\ <\ \lambda\ <\ 0$. В дальнейших экспериментах будем выбирать $\lambda = \frac{1}{30}$.

5.2.3 Зависимость SemH и SemI от количества итераций при обучении

В данном эксперименте бралась модель PLSA на 20 и на 150 тем. Модели постепенно дообучались и строились графики зависимости SemH и SemI от количества итераций EM - алгоритма, проведенных для каждой модели. Кроме этого, на графиках так же показана перплексия модели. На рисунке (5), слева, представлен график зависимости для модели на 20 тем, справа — для модели на 150 тем. Видно, что, несмотря на то, что перплексии модели практически не меняется, степень загрязненности SemI продолжает падать. Степень неоднородности же, SemH, устанавливается уже после небольшого количества итераций. Это можно проинтерпретировать тем, что изначально у модели формируются большие кластеры

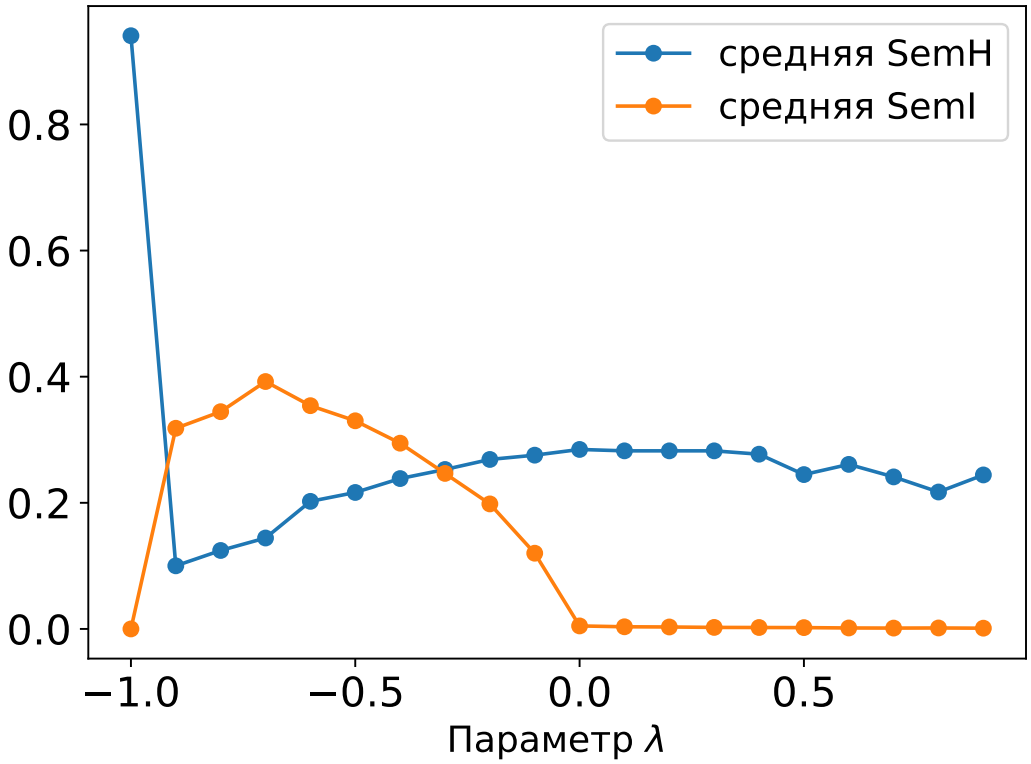
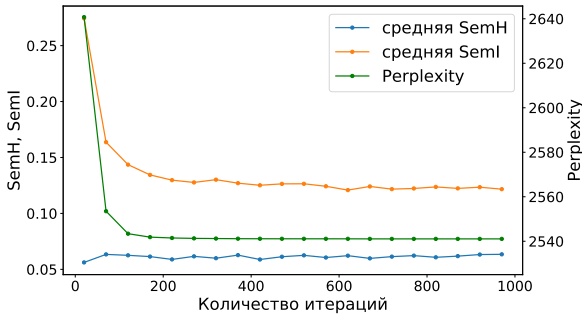


Рис. 4: Зависимость степеней загрязненности и неоднородности от параметра λ

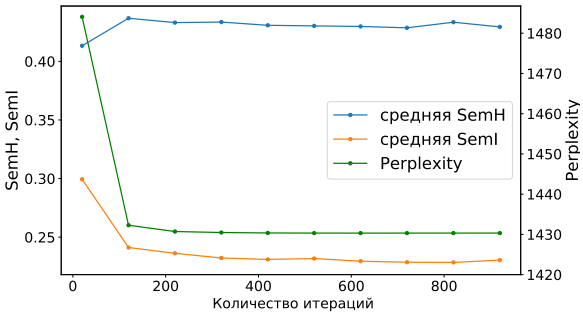
тем, и, по мере дообучения, кластеры становятся меньше и обособленней, не увеличивая при этом неоднородность. Кроме этого, по степени загрязненности видно, что модели с большим количеством тем нужно больше итераций, чтобы обучиться.

5.2.4 Проверка сбалансированности модели с фоновой темой

Предобученная модель содержит 20 тем, причем тема 20 — фоновая, то есть содержит общеупотребительные слова, стоп-слова и связывающие обороты. Для наглядности на гистограммах ниже будет показываться $(1 - \mathrm{SemH}(t))$ и $\mathrm{SemI}(t)$: таким образом, гистограмма разделится на три секции: $\mathrm{SemI}(t)$ — доля документов, содержащих как минимум две темы, $(1 - \mathrm{SemI}(t) - \mathrm{SemH}(t))$ — доля документов, содержащих тему t и только её, и $\mathrm{SemH}(t)$ — доля документов, не содержащих тему t. На рисунке (6), слева, показана гистограмма $(1 - \mathrm{SemH}(t))$ и $\mathrm{SemI}(t)$ для модели с фоновой темой: видно, что доля документов, относящихся только к одной



a)PLSA с 20 темами



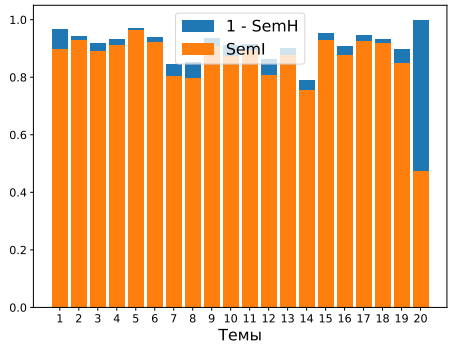
b)PLSA с 150 темами

Рис. 5: Зависимость степеней загрязненности и неоднородности количества итераций

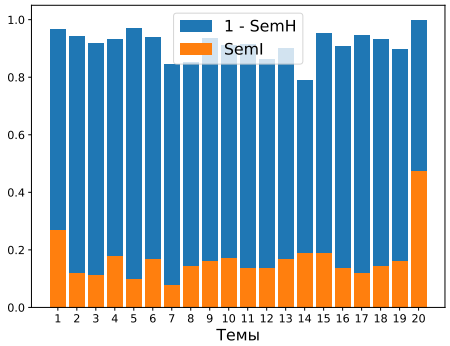
теме (то есть синяя часть гистограммы на рисунке), очень мала. Если пересчитать степени загрязненности, исключив фоновую тему t_{back} из множества тем, среди которых ищется минимальное расстояние $S_{dtt'}$:

$$S_{dtt'} = \min_{t' \in T \setminus t, t_{back}} CR_\lambda\big(\hat{p}(u \mid d, t) : \hat{p}(u \mid t')\big),$$

получим(на рисунке (6), справа), что степени загрязненности резко уменьшаются для всех тем. Это подтверждает предположение, что фоновая тема состоит из общеупотребительных слов и присутствует практически в каждом документе. Кроме того, это означает, что кластер фоновой темы «покрывает» все остальные кластеры.



a)Фоновая тема не исключена

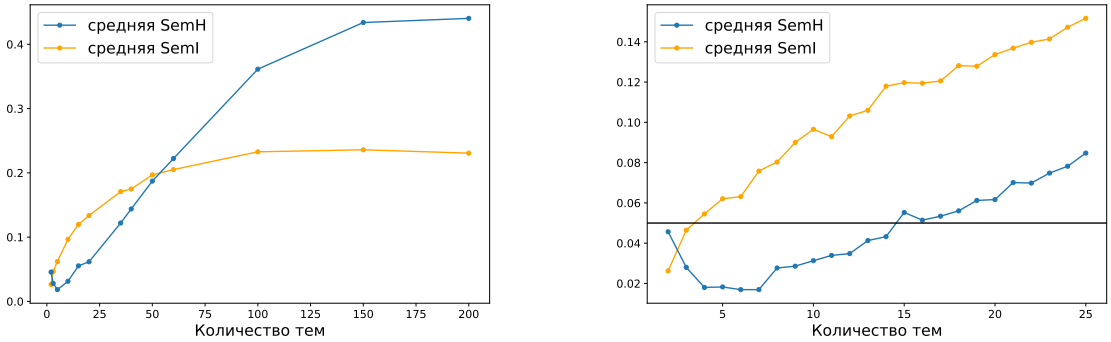


b)Фоновая тема исключена

Рис. 6: Сравнение 1 – SemH и SemI для предобученной модели.

5.2.5 Зависимость сбалансированности модели от количества тем

Для этого эксперимента обучался набор моделей $\{(\Phi, \Theta)\}_{i=1}^{n}$ по одной и той же коллекции, но с разным числом тем. Модели обучались без регуляризаторов. На рисунке (7) представлена зависимость средних SemH и SemI от числа тем. На рисунке (7), справа, черной линией отмечен уровень сбалансированности 0.05. Видно, что при увеличении числа тем и SemH, и SemI увеличиваются.



a)Общий график b)График для количества тем ≤ 25

Рис. 7: Зависимость степеней загрязненности и неоднородности числа тем в модели

5.2.6 Влияние регуляризатора декоррелирования на сбалансированность тем

Обучался набор моделей $\{(\Phi, \Theta)\}_{i=1}^{n}$ на 80 тем по одной и той же коллекции «Постнауки» и с разным значением τ при регуляризаторе декоррелирования. На рисунке (8) представлена зависимость средних SemH и SemI от значения τ при реуляризаторе. Кроме того, на графике изображена так же перплексия модели. Видим, что в при увеличении τ загрязненность SemI падает, а неоднородность SemH растет. Это легко интерпретируется: при добавлении регуляризатора декоррелирования темы становятся более непохожими друг на друга, а значит кластеры сужаются и становятся более обособленными. Кроме того, если обратить внимание на момент, когда перплексия резко вырастает(что свидетельствует о вырождении модели), видно, что степень загрязненности SemI становится практически нулевой, а неоднородность SemH претерпевает скачки.

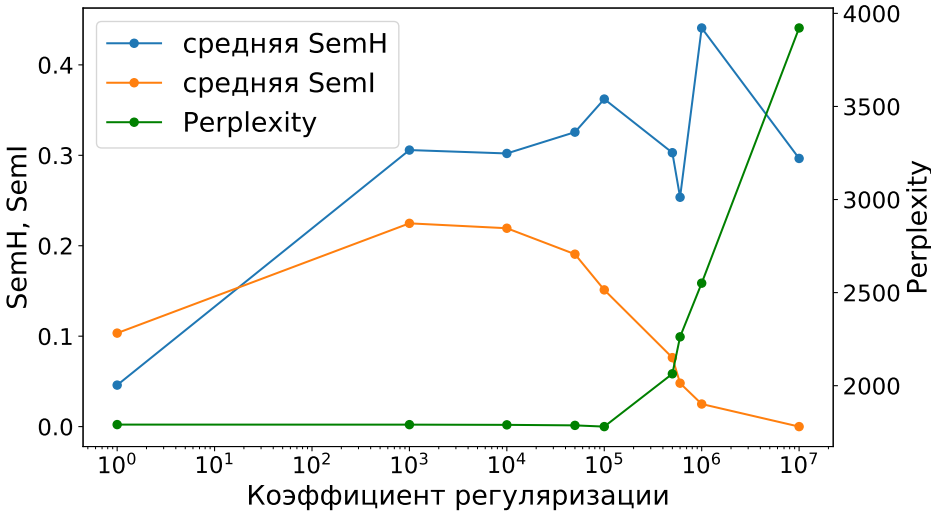


Рис. 8: Зависимость SemI и SemH от параметра τ в регуляризаторе декоррелирования

Значит, по SemH и SemI так же, как и по перплексии, можно проследить момент перерегуляризации.

6 Заключение

В работе был предложен новый критерий качества тем в тематических моделях, основанный на выполнимости гипотезы условной независимости в коллекции и характеризующий качество тематической кластеризации: степени неоднородности и степени загрязненности(SemH, SemI). Был разработан алгоритм по их вычислению.
Исследована зависимость SemH, SemI от параметра λ в статистике Кресси-Рида, дана рекомендация по выбору λ для поведения вычислений.
Исследована зависимость SemH, SemI от количества итераций EM - алгоритма, на котором была обучена модель. Показано, что по SemH, SemI, как и по перплексии, можно определить степень обученности модели.
Исследовано влияние регуляризатора декоррелирования на SemH, SemI. Показано, что при увеличении параметра регуляризации τ степень загрязненности уменьшается, а степень неоднородности увеличивается. Показано, что по SemH, SemI можно проследить момент перерегуляризации.

Список литературы

[1] *Ana Catarina Calheiros and Sérgio Moro and Paulo Rita*Sentiment Classification of Consumer-Generated Online Reviews Using Topic Modeling // Journal of Hospitality Marketing & Management, 2017, 26(7):675-693

[2] *W. Cui and S. Liu and L. Tan and C. Shi and Y. Song and Z. Gao and H. Qu and X. Tong* TextFlow: Towards Better Understanding of Evolving Topics in Text. // IEEE Transactions on Visualization and Computer Graphics, 2011, 17(12), Pp.2412–2421.

[3] *La Rosa, Massimo and Fiannaca, Antonino and Rizzo, Riccardo and Urso, Alfonso* Probabilistic topic modeling for the analysis and classification of genomic sequences // BMC Bioinformatics, 2015, 16(Suppl 6):S2

[4] *Qiang, Jipeng and Chen, Ping and Wang, Tong and Wu, Xindong"*Topic Modeling over Short Texts by Incorporating Word Embeddings // Advances in Knowledge Discovery and Data Mining, 2017,Pp. 363–374

[5] *Doyle, Gabriel and Elkan, Charles* HAccounting for Burstiness in Topic Models. // Proceedings of the 26th Annual International Conference on Machine Learning, ICML'09, 2009, Pp. 281–288

[6] *Newman, David and Lau, Jey Han and Grieser, Karl and Baldwin, Timothy* Automatic Evaluation of Topic Coherence. // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT '10, 2010 Pp. 100–108

[7] *Vorontsov K. V.* Additive regularization for topic models of text collections //Doklady Mathematics, 2014, 89(3):Pp. 301–304

[8] *Cressie, N. and Read, T. R.*
Multinomial Goodness-Of-Fit Tests. //Journal of the Royal Statistical Society: Series B (Methodological), 1984, 46:Pp. 440-464