Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)

Физтех-Школа Прикладной Математики
Кафедра Интеллектуальных Систем

Выпускная квалификационная работа бакалавра по направлению 010900
«Прикладные математика и физика»

Проверка гипотез условной независимости в тематических моделях

Студент 574 группы
Рогозина А. А.

Научный руководитель
Воронцов К. В.

Долгопрудный
2019

Содержание

Глава 1

Введение

Задача определения тематики текста имеет множество практических приложений:(цитата, цитата, цитата). 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)[цитата][цитата]: 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[цитата] 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). Таким образом, предложенный в работе критерий оценивания качества построенных тем

позволяет перевести однородность и непохожесть на остальные темы в количественную характеристику и избавляет от необходимости просматривать набор топ-слов каждой темы.

Глава 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:

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 с распределением 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, $\phi_{wt} = p(w \mid t)$ 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):

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

$$= \prod_{d \in D} \prod_{w \in d} (\sum_t \phi_{wt} \theta_{td})^{n_{wd}} p(d)^{n_w d} \to \max_{\Phi, \Theta}$$

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

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

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

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 критерии, оценивающие насколько для данной модели (Φ, Θ) в данной коллекции D выполняется гипотеза условной независимости. Основываясь на этих критериях, предлагается оценить качество тем, построенных моделью (Φ, Θ).

Глава 3

Статистические тесты проверки 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 $p(x)$:

$$H_0 : X = \{x_1, \ldots, _n\} \in p(x)$$
$$H_1 : X = \{x_1, \ldots, _n\} \notin p(x)$$

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* между двумя распределениями:

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

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 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.

3.2.1 Возможные приближения

В работах [надо][процитировать] предлагается группировать слова, увеличивая тем самым 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 разбиения, 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.

3.2.2 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 количества n_{td} вхождений слов темы t в документ d и от темы t.

Глава 4

Оценка качества тем в тематической модели с помощью тестов, проверяющих выполнение гипотезы условной независимости в коллекции

4.1 Сбалансированность тем

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(П. 3), 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 доку-

ментов d, для которых значение статистики S_{td} больше радиуса семантической однородности $R_t^\alpha(n_{td})$.

$$\text{SemHeterogeneity}(t) = \sum_{d\in D} \hat{p}(d\,|\,t)\big[S_{dt} < R_t^\alpha(n_{td})\big] = \sum_{d\in D} \frac{n_{td}}{n_t}\big[S_{dt} < R_t^\alpha(n_{td})\big]\,,$$

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':

$$\text{SemImpurity}(t) = \sum_{d\in D} p(d\,|\,t)\big[S_{dt} < R_t^\alpha(n_{td})\big]\,\big[S_{dt'} < R_{t'}^\alpha(n_{td})\big],$$

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

$$S_{dt'} = \min_{t'\in T\setminus t} \lambda\big(\hat{p}(u\,|\,d,t) : \hat{p}(u\,|\,t')\big).$$

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 1 показана иллюстрация к подсчету степеней загрязненности и неод-

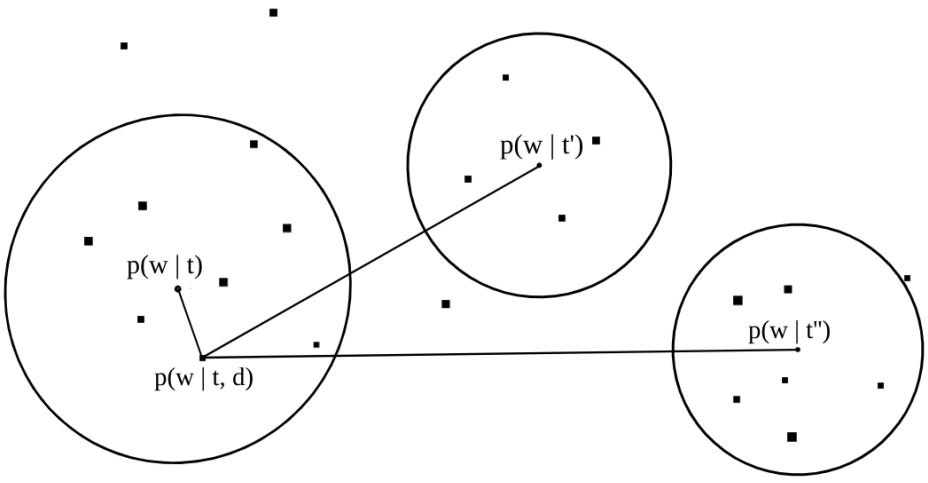


Рис. 1. Иллюстрация работы кластерной структуры распределений

нородности: 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 $S_{t'd} = CR_\lambda\big(\hat{p}(w \mid d,t) : p(w \mid t')\big)$ и выбрать минимальное расстояние $S_{t_{min}d}$ и сравнить его с $R^\alpha_{t_min}(n_{td})$. Если тема не является ни семантически неоднородной, ни семантически загрязненной, назовем ее *сбалансированной*.

Глава 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,
$\{U \subseteq W : \forall u \in Up(u \mid t) > \frac{1}{W},\ n_{tdu} \geq 0\}$, и считать $S_{dt} = CR_\lambda\big(\hat{p}(u \mid d, t) : \hat{p}(u \mid t)\big)$.
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, $\{U \subseteq W : \forall u \in Up(u \mid T) \geq \frac{1}{W}, n_{tdu} > 0\}$.

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.

Algorithm 5.2.1 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 $S_{dt} = CR_\lambda\big(\hat{p}(u\,|\,d,t) : \hat{p}(u\,|\,t)\big)\big)$
    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 $S_{dt} = CR_\lambda\big(\hat{p}(u\,|\,d,t) : \hat{p}(u\,|\,t')\big)\big)$
        Найти $t_{min} =_{t'\neq t} S_{dt'}$
        Сравнить $S_{dt_{min}}$ и $R_{t_{min}}^\alpha(n_{td}))$
  Вычислить SemH, SemI по формулам(), ()

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 2 представлена эта зависимость. Видно, что при $\lambda ge 0$ степень загрязненности становится нерепрезентативной и практически нулевой, а при $\lambda le -1$ степень неоднородности становится практически 1, что эквивалентно стягиванию кластеров тем в точку. Поэтому рекомендуется выбирать $-1 < \lambda < 0$. В дальнейших экспериментах будем выбирать $\lambda = \frac{1}{30}$.

5.2.3 Проверка сбалансированности предобученной модели

Предобученная модель содержит 20 тем, причем двадцатая тема — фоновая, то есть содержит общеупотребительные слова, стоп-слова и связывающие обороты. Для наглядности на гистограммах ниже будет показываться $1 - \text{SemHeterogeneity}(t)$ и $\text{SemImpurity}(t)$: таким образом, гистограмма разделится на три секции: $\text{SemImpurity}(t)$ - доля документов, содержащих как минимум две темы, $1-\text{SemImpurity}(t)-\text{SemHeterogeneity}(t)$ - доля документов, сожержащих тему t и только её, и $\text{SemHeterogeneity}(t)$ - доля документов, не содержащих тему t. На Рис. 3 слева показана гистограмма $1-\text{SemHeterogeneity}(t)$ и $\text{SemImpurity}(t)$ для предобученной модели: видно, что доля тем, содержащих только одни тему, очень мала. Если пересчитать степени загрязненности, исключив фоновую тему t_{back} из множества тем, среди которых ищется минимальное расстояние $S_{dt'}$:

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

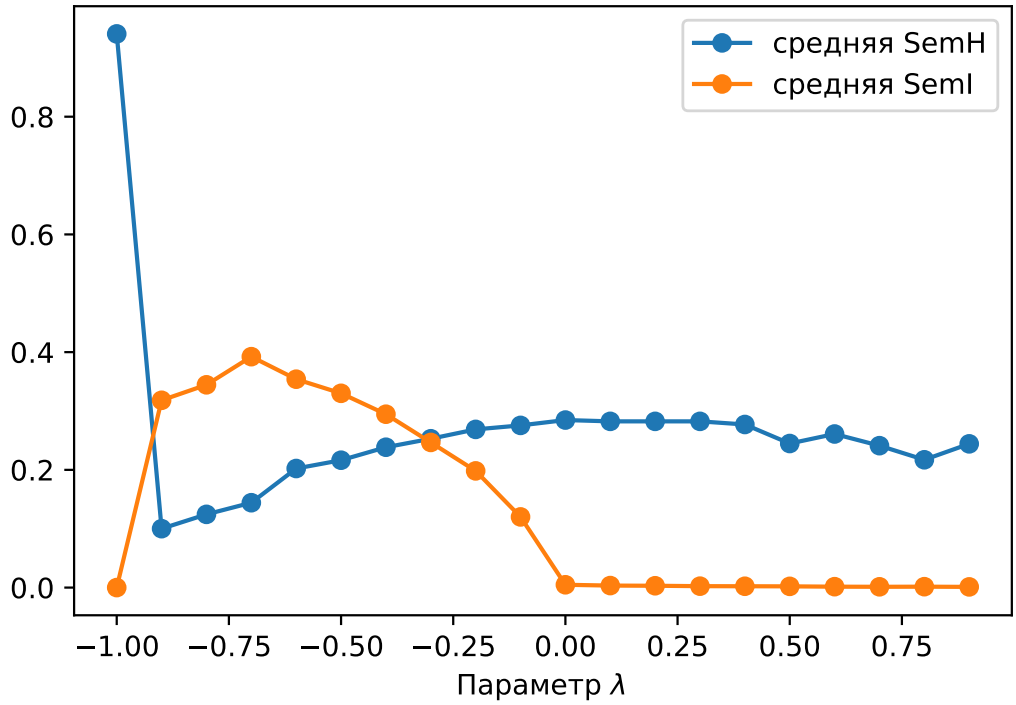
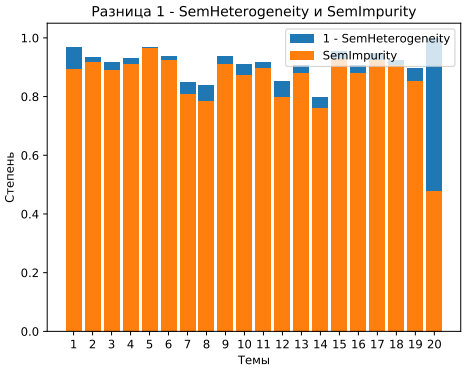
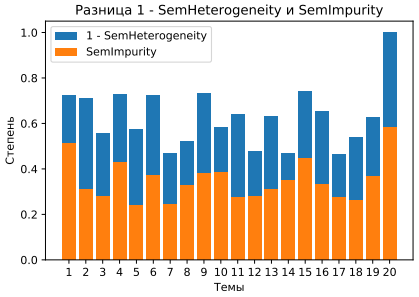


Рис. 2. Зависимость степеней загрязненности и неоднородности от параметра λ

получим(на Рис. 3, 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.



a)Фоновая тема не исключена



b)Фоновая тема исключена

Рис. 3. Сравнение 1 – SemH и SemI для предобученной модели.

5.2.4 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 рис. 4 представлена зависимость средних SemH и SemI от числа тем. Видно, что при увеличении числа тем и SemH, и SemI увеличиваются.

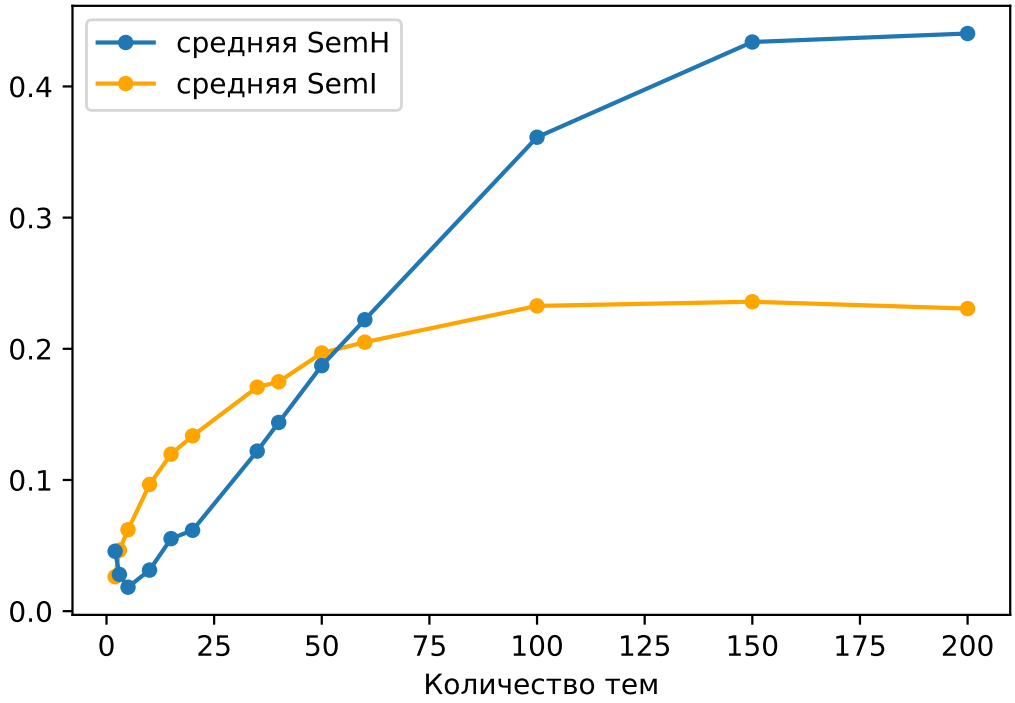


Рис. 4. Зависимость степеней загрязненности и неоднородности от числа тем в модели

5.2.5 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 в регуляризаторе декоррелирования. На рис. 5 представлена зависимость средних SemH и SemI от значения τ в модели. Кроме того, на графике изображена так же перплексия модели. Видим, что в при увеличении *tau* 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

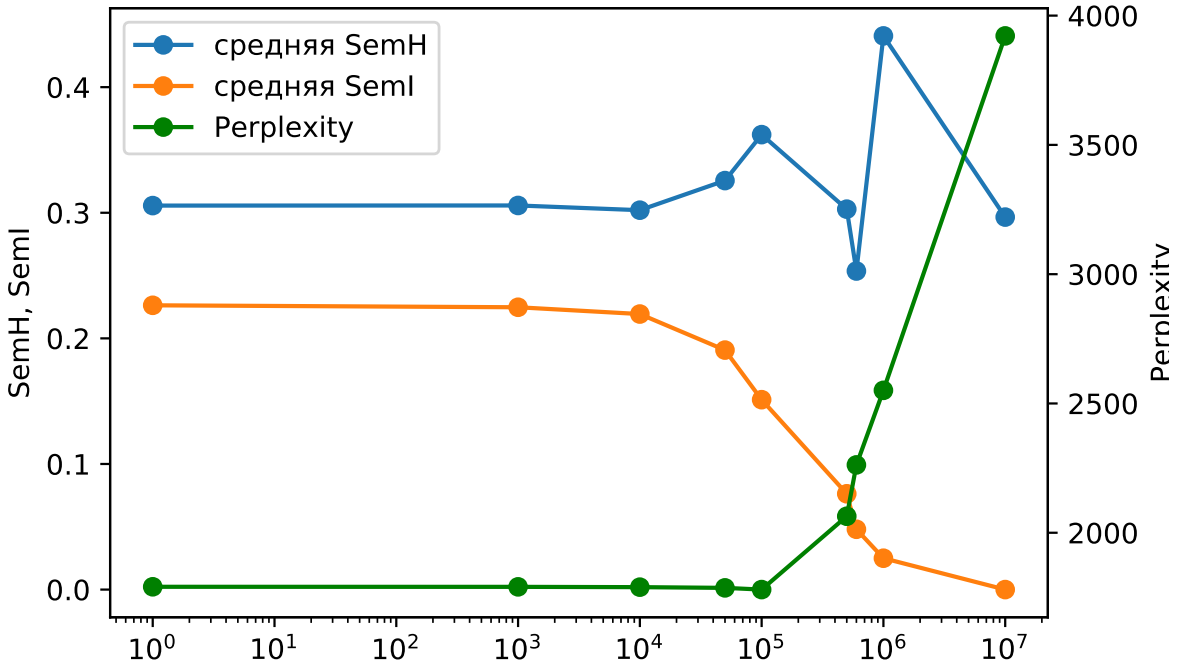


Рис. 5. Зависимость степеней загрязненности и неоднородности от параметра τ в регуляризаторе декоррелирования

вырастает(что свидетельствует о вырождении модели), видно, что степень загрязненности SemI становится практически нулевой, а неоднородность SemH претерпевает скачки. Значит, можно использовать SemH, и SemI для определения подходящего параметра τ в регуляризаторе декоррелирования.