Проверка гипотезы условной независимости для оценивания качества тематической кластеризации

Рогозина Анна

Московский Физико-технический институт

Физтех-Школа Прикладной математики и Информатики

Кафедра интеллектуальных систем

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

13 июня 2019

Вероятностное тематическое моделирование

Дано:

- Множество токенов W, коллекция текстовых документов D, множество тем T
- n_{wd} — частоты токенов в документах
- $D \times W \times T$ — дискретное вероятностное пространство

Предположение:

- Гипотеза условной независимости: $p(w \mid d, t) = p(w \mid t)$

Найти параметры модели: $p(w \mid d) = \frac{n_{wd}}{n_d} = \sum_{t \in T} \varphi_{wt} \theta_{td}$

- $\varphi_{wt} = p(w \mid t)$ - вероятность токенов w в теме t
- $\theta_{td} = p(t \mid d)$ - вероятность тем t в документе d

Постановка задачи

Проблема

Оценивание качества отдельных тем

Существующее решение

Когерентность тем (Newman D. et all, (2011), Optimizing Semantic Coherence in Topic Models)

Цель работы

Построить критерий, характеризующий выполнимость гипотезы условной независимости для каждой темы

- Без экспертных оценок
- Эффективно вычислимый
- Имеющий итерпретируемое значение

Кластерная структура распределений слов

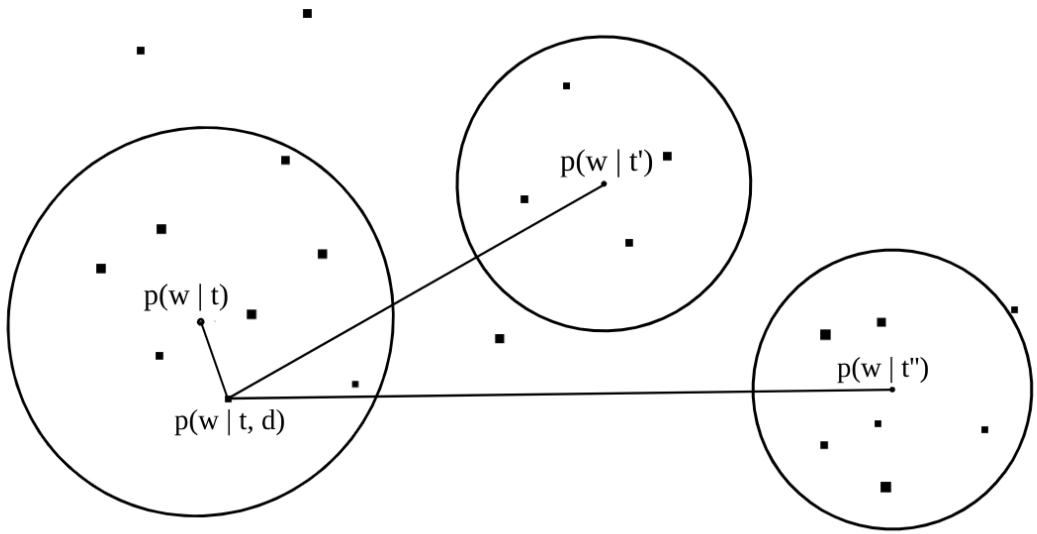


Иллюстрация кластерной структуры распределений

- Точки—распределения слов в документах $p(w \mid t, d)$
- Центры кластеров—распределение слов в теме $p(w \mid t)$

Формирование кластеров

Для каждой темы t и документа d проверяем гипотезу:

$$H_0: \; p(w \,|\, d,t) = p(w \,|\, t)$$
$$H_1: \; p(w \,|\, d,t) \neq p(w \,|\, t)$$

Дивергенция Кресси-Рида между двумя распределениями:

$$\mathrm{CR}_\lambda\big(\hat{p}(w \,|\, d,t) : \hat{p}(w \,|\, t)\big) =$$

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

$$= \frac{2}{\lambda(\lambda+1)} \sum_{w \in W} \frac{n_{dw}\varphi_{wt}\theta_{td}}{\sum_{s \in T} \varphi_{ws}\theta_{sd}} \left(\left(\frac{n_{wd}}{n_d \sum_{s \in T} \varphi_{ws}\theta_{sd}} \right)^\lambda - 1 \right) \qquad (2)$$

Обозначения

- S_{dt} — значение CR_λ для документа d и темы t
- Радиус семантической однородности $R_t^\alpha(n_{td})$ темы t — $(1-\alpha)$ квантиль распределения S_{dt}.

Степень семантической неоднородости

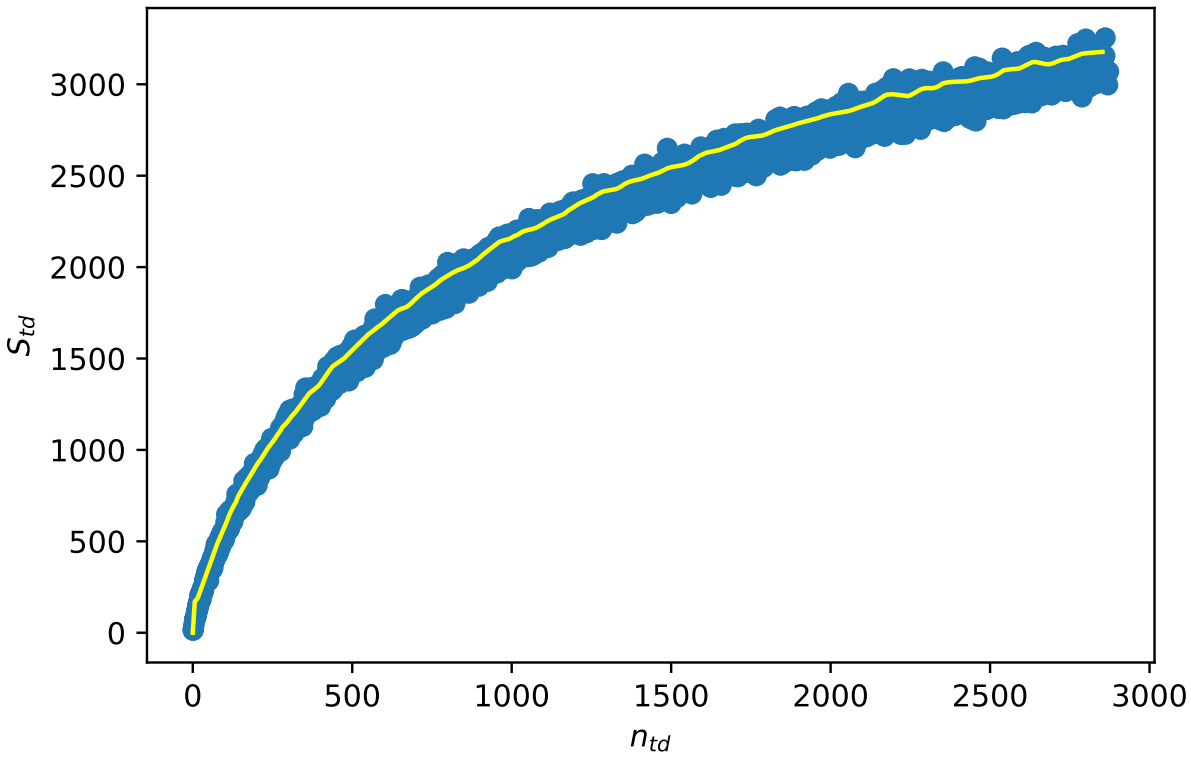
$$\mathrm{SemH}(t) = \sum_{d \in D} p(d|t) \left[S_{dt} > R_t^\alpha(n_{td}) \right] \tag{3}$$

Степень семантической загрязненности

$$\mathrm{SemI}(t) = \sum_{d \in D} p(d|t) \left[S_{dt} < R_t^\alpha(n_{td}) \right] \left[S_{dtt'} < R_{t'}^\alpha(n_{td}) \right] \tag{4}$$

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

Input: $\Phi,\ \Theta,\ \lambda$
Result: SemI, SemH
for *тема* $t \in T$ **do**
  Сгенерировать коллекцию документов D *из* $p(w\,|\,t)$ *с различными* n_{td}, *получить* $\{(n_{tdw}, n_{td})\}_{d\in D}$
  $(n_{tdw}, p(w\,|\,t)) \to (n_{tdu}, p(u\,|\,t))$, *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* $R_t^\alpha(n_{td}))$;
end
for *тема* $t \in T$ **do**
  for *документ* $d \in D$ **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 $t_{min} = \arg\min\limits_{t'\neq t} S_{dt'}$;
        Сравнить $S_{dt_{min}}$ *и* $R^\alpha_{t_{min}}(n_{td}))$;
      end
    end
  end
  Вычислить SemH, SemI по формулам (3), (4)
end



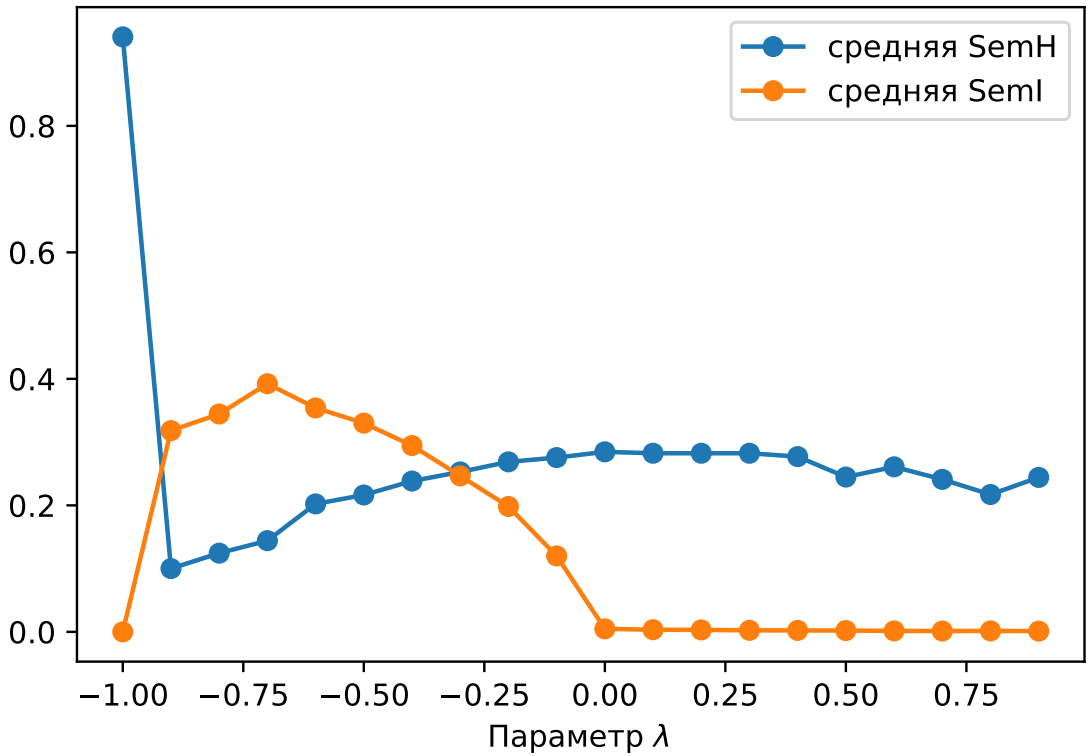
Пример непараметрической квантильной регрессии $R_t^\alpha(n_{td})$

Эксперименты

Данные

- Коллекция «Постнаука»
- $\sim$ 3500 документов
- Документы на научно-популярную тематику

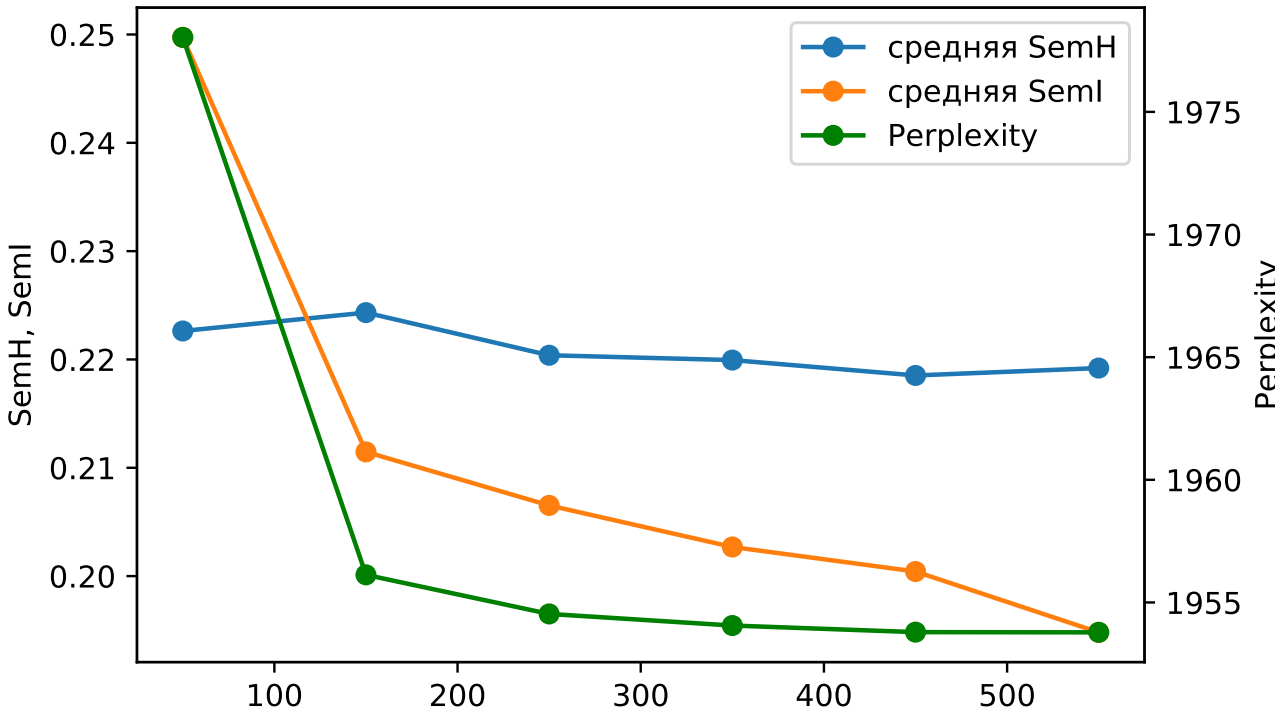
Зависимость от параметра λ



Зависимость SemH и SemI от параметра λ

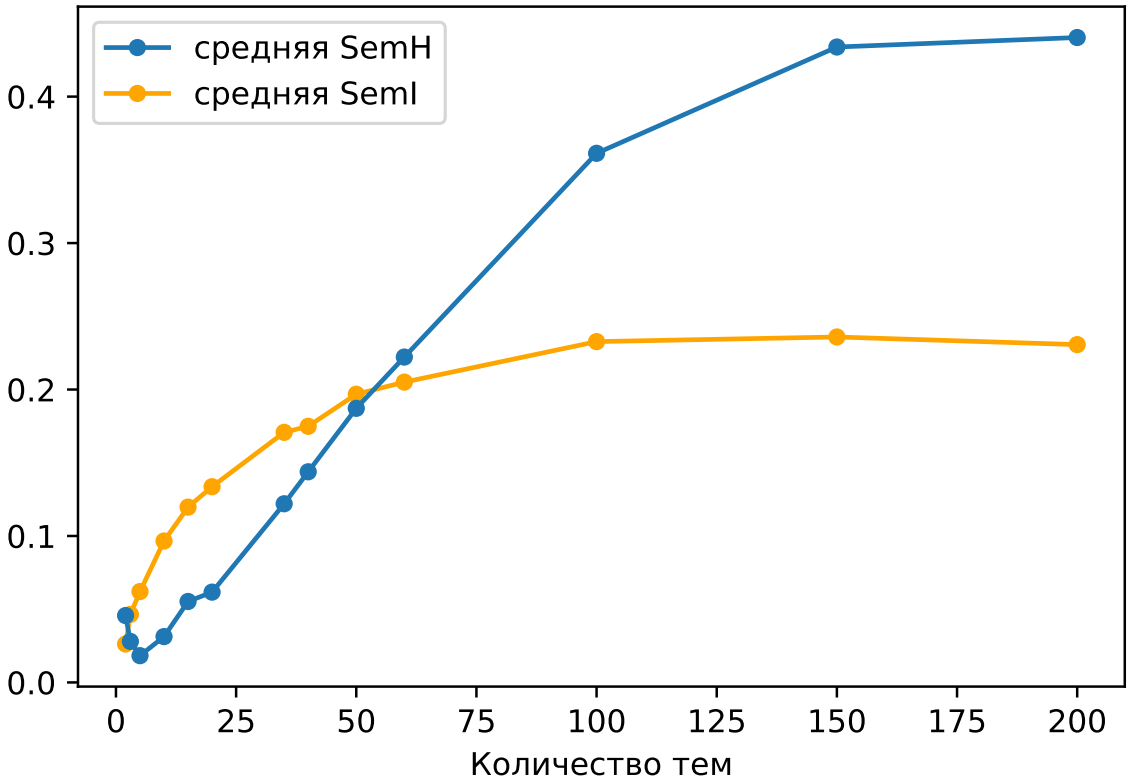
Вывод: рекомендуемый диапазон : $-0.8 \leq \lambda \leq -0.2$

Зависимость от количества итераций при обучении



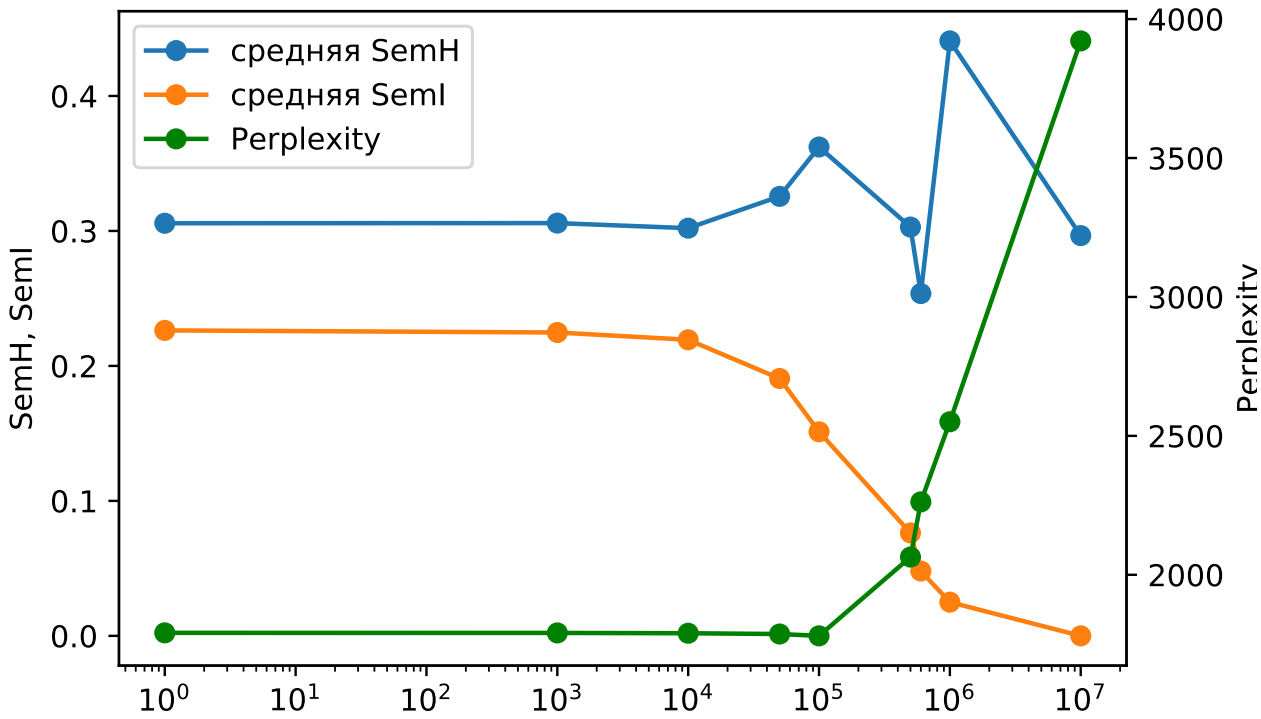
Зависимость SemH и SemI от количества итераций, PLSA на 80 тем

Зависимость от количества тем в модели



Зависимость SemH и SemI от числа тем в модели PLSA

Влияние регуляризатора декоррелирования



Зависимость SemH и SemI параметра регуляризатора декоррелирования τ, модель 60 тем

Результаты, выносимые на защиту

- Разработан алгоритм вычисления SemH и SemI на основе проверки гипотезы условной независимости
- Исследована зависимость SemH и SemI от количества тем в модели
- Установлены рекомендации по выбору параметра λ в статистике Кресси-Рида
- Исследовано влияние регуляризатора декоррелирования на SemH и SemI, установлены рекомендации по выбору параметра в регуляризаторе декоррелирования.