Оценивание способов борьбы с несбалансированостью тем в тематическом моделировании

Рогозина Анна

Московский Физико-технический институт

Факультет управления и прикладной математики

Кафедра интеллектуальных систем

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

2019

План

Постановка задачи

Задача тематического моделирования

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: $p(w \mid d) = \sum_{t \in T} \varphi_{wt} \theta_{td}$

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

Принцип максимума правдоподобия

Оптимизируемый функционал:

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

- Мощность темы $\hat{p}(t) = \frac{n_t}{n}$
- Гипотеза:

  Отношение максимальной к минимальной мощности $\frac{\hat{p}_{max}(t)}{\hat{p}_{min}(t)} \leq 3 - 4$ раза

Задача

Научиться выявлять «несбалансированность» тем в построенной модели. Сравнить способы борьбы с появлением «несбалансированности» тем.

Кластерная структура распределений слов

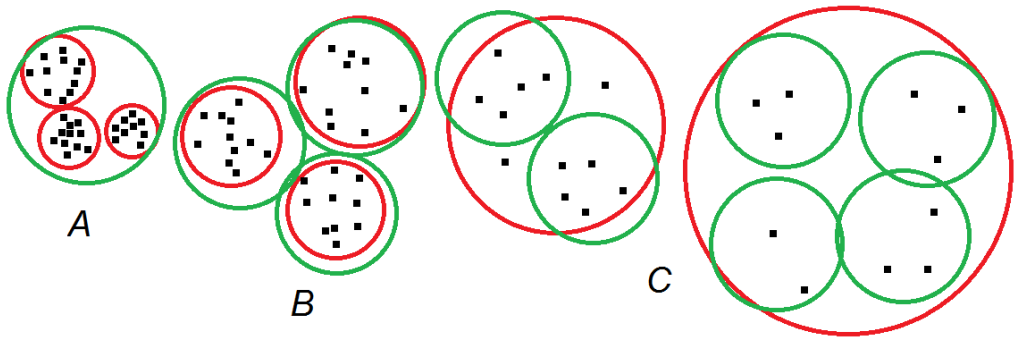


Рис.: Иллюстрация работы E-M алгоритма для несбалансированных тем

- Точки на графике - 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:

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

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

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

Обозначения

- S_{dt} - значение дивергенции CR_λ для документа d и темы t
- Радиус семантической однородности $R_t^\alpha(n_{td})$ темы t с уровнем значимости α

Степень семантической неоднородости

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

Степень семантической загрязненности

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

Итеративная балансировка тем

Локальный экстремум задачи тематического моделирования:

$$p_{tdw} = \underset{t \in T}{\text{norm}}\left(\varphi_{wt}\theta_{td}\right); \tag{1}$$

$$\varphi_{wt} = \underset{w \in W}{\text{norm}}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right); \quad n_{wt} = \sum_{d \in D} n_{dw} p_{tdw}; \tag{2}$$

$$\theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \quad n_{td} = \sum_{w \in d} n_{dw} p_{tdw}. \tag{3}$$

$n_{tdw} = n_{dw} p_{tdw}$ - оценка числа употреблений слова w в документе d по теме t.

Идея:

Домножим n_{tdw} на величину, обратно пропорциональную n_t.

Изменение вероятностей:

$$p'_{tdw} = \frac{n'_{tdw}}{\sum_s n'_{sdw}} = \frac{\frac{1}{n_t} n_{tdw}}{\sum_s \frac{1}{n_s} n_{sdw}} = \frac{1}{n_t} p_{tdw} \frac{\sum_s n_{sdw}}{\sum_s \frac{1}{n_s} n_{sdw}} = \underset{t \in T}{\text{norm}}\left(\frac{1}{n_t} p_{tdw}\right).$$

Изменение формул E-M алгоритма:

$$p_{tdw} = \underset{t \in T}{\text{norm}}\left(\varphi_{wt}\theta_{td}\right); \qquad n_t = \sum_{d \in D} \sum_{w \in d} n_{dw} p_{tdw}.$$

$$Z_{dw} = \sum_{t \in T} \frac{p_{tdw}}{n_t}.$$

$$\varphi_{wt} = \underset{w \in W}{\text{norm}}\left(n_{wt} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}}\right); \qquad n_{wt} = \sum_{d \in D} \frac{n_{dw}}{n_t Z_{dw}} p_{tdw};$$

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

Проверка качества борьбы с несбалансированностью

Алгоритм

- Генерируем список коллекций с разной степенью сбалансированности тем $\{D_i\}_{i=1}^n$
- Для каждой коллекции D_i обучаем модель ARTM $\{(\Phi_i, \Theta_i)\}_{i=1}^n$
- Для каждой модели оцениваем качество всех тем с помощью SemHeterogeneity(t) и SemImpurity(t)

Генерация коллекций

Дано:

- Коллекция документов из «ПостНауки»(3404 документа)
- Предобученная на этой коллекции матрица Φ на 20 тем

Алгоритм генерации коллекции длины N:

- Из предобученной модели берем распределение $p(w \mid t)$
- Задаем дискретное распределение тем в коллекции $p(t)$
- Для каждого документа $\{d_i\}_{i=1}^{N}$:
  - Из списка длин реальных документов «ПостНауки» равновероятно выбираем длину генерируемого документа n_i
  - Генерируем распределение $p(t \mid d) \in Dir(p(t))$
  - Генерируем слова $\{w_j^i\}_{j=1}^{n_i}$:
    - Выбираем тему $t_j \in p(t \mid d)$
    - Добавляем в документ слово $w_j \in p(w \mid t_j)$

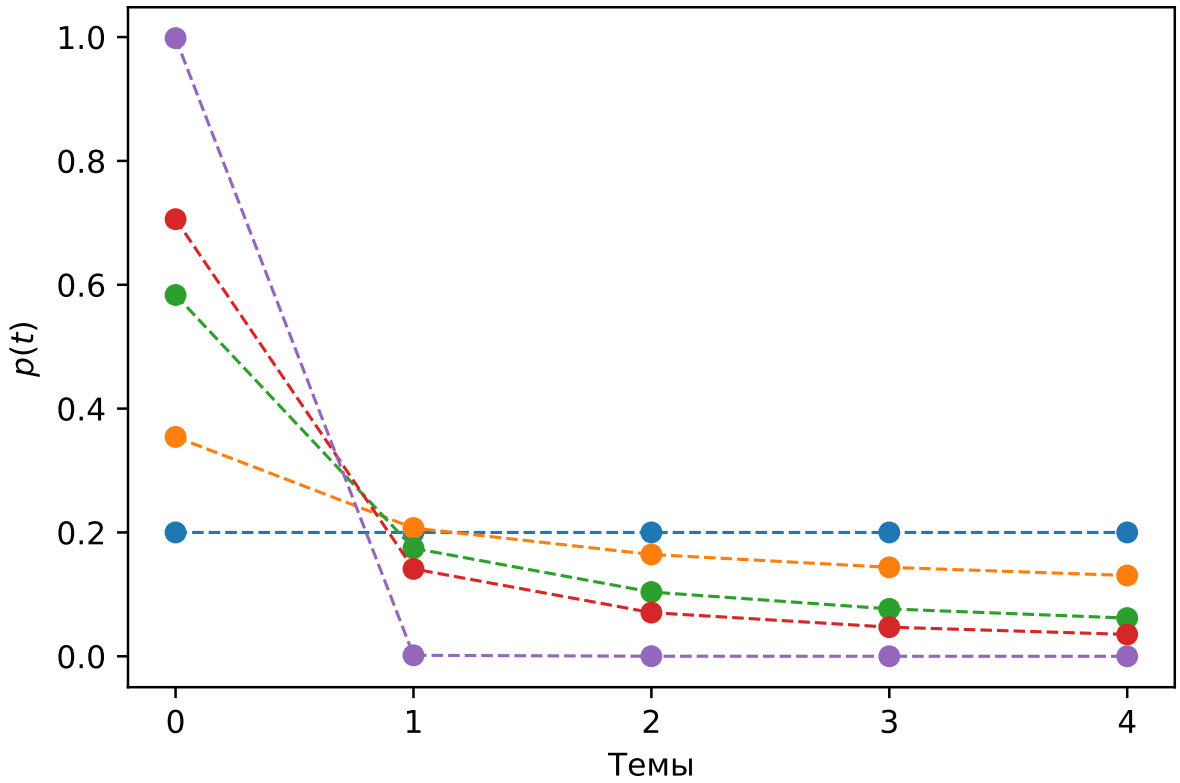


Рис.: Пример градации $p(t)$ для 5 тем

Вычисление $R_t^\alpha(n_{td})$

Проблема:

Распределение статистики Кресси-Рида $\mathrm{CR}_\lambda \in \chi^2(|W|)$, но только если $\forall w, d : n_{wd} \geq 5$

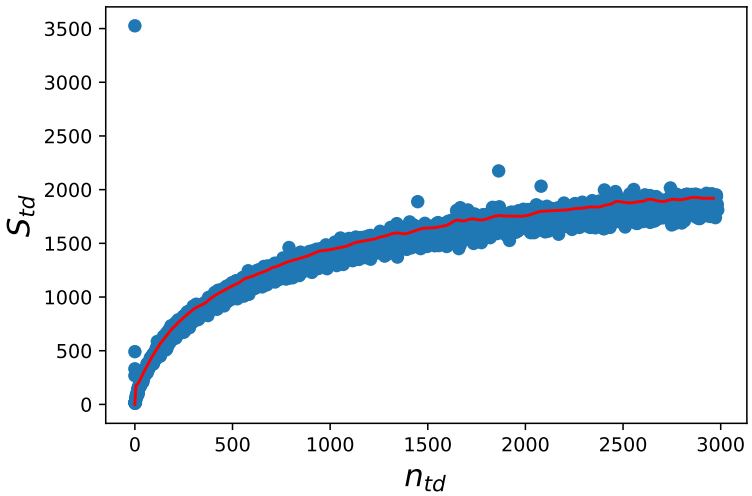
На самом деле

Распределение статистики зависит от t и n_{td}

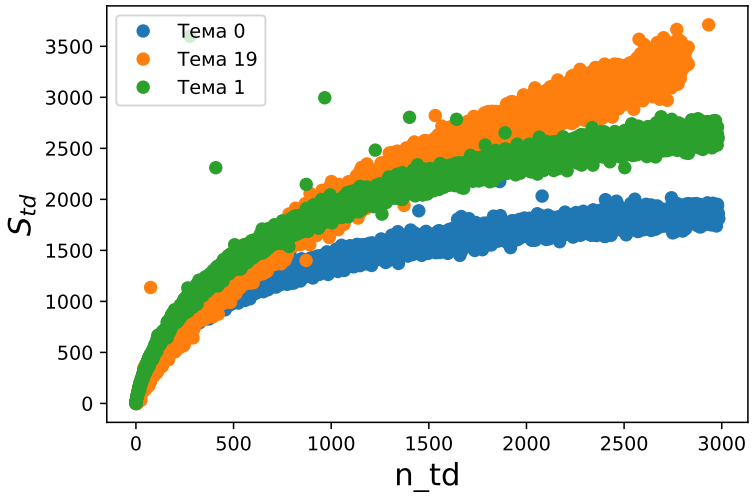
Предложение

Для каждой модели (Φ_i, Θ_i), для каждой темы t_j^i

- Генерируем коллекцию $D_{t_j^i}$, состоящую из документов d_{t_i} разной длины, содержащих только тему t_j^i
- Вычисляем значение статистики $S_{t_j^i d}$
- По полученной эмпирической зависимости $S_{t_j^i d}(n_{td})$ считаем непарметрическую квантильную регрессию $R_t^\alpha(n_{td})$.



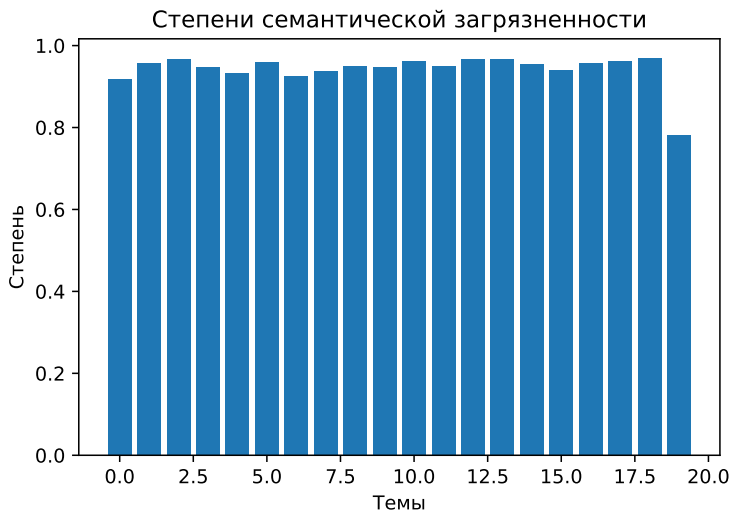
а) Квантильная регрессия для темы 0



б) Распределение S_{td} для разных тем

Проверка качества предобученной модели

Подсчитаны степени семантической однородности и загрязненности



Выводы

- Видим, что степень загрязненности ~ 1 для каждой темы
- Значит, либо каждый документ содержит больше двух тем
- Либо критерий в данном виде слишком маломощный и вообще ничего не отвергает(Сейчас статистика считается для всех слов $w : p(w \mid t) > \frac{1}{|W|}$)

Возможные модификации

- Из всех 20 тем отобрать те, что не пересекаются в смысле степеней загрязненности(если это возможно)
- Начать считать статистику S_{td} только для слов, входящих в документ d

Заключение

Что сделано:

- Поставлена задача
- Предложен способ оценивания качества моделей
- Разработан дизайн эксперимента
- Проведена оценка качества предобученной модели

Нужно:

- Увеличить мощность используемого критерия CR_λ
- Провести эксперимент для итеративной балансировки тем
- Реализовать оптимизацию гиперпараметров сглаживания
- Провести эксперимент и оптимизации гиперпараметров