

Вероятностные тематические модели на основе данных о со-встречаемости слов

Михаил Солоткий

Московский государственный университет им. М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математических методов прогнозирования

Выпускная квалификационная работа бакалавра

Научный руководитель — д.ф-м.н. Воронцов К. В.

Москва 2019 г.

Тематическое моделирование, модель PLSA

Тематическое моделирование — один из подходов к статистическому анализу текстов.

Дано: коллекция текстовых документов D , словарь токенов W , счётчики вхождения токенов в документы n_{dw} . Каждый токен в каждом документе описывается некоторой скрытой темой $t \in T$.

Найти: вероятностные распределения $\Phi = P(w|t)$, $\Theta = P(t|d)$ методом максимального правдоподобия:

$$\frac{n_{dw}}{\sum_w n_{dw}} \approx P(w|d) = \sum_{t \in T} P(w|t) P(t|d)$$

$$\mathcal{L}(\Phi, \Theta) = \sum_{d, w} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta}$$

АРТМ — один из подходов к регуляризации log-правдоподобия:

$$\mathcal{L}(\Phi, \Theta) = \sum_{d,w} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + \sum_{i=1}^k \tau_i R_i(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

- Желаемые свойства полученных тем можно формализовать в виде регуляризаторов $R_i(\Phi, \Theta)$
- Можно оптимизировать ЕМ-алгоритмом

- Перплексия:

$$\mathcal{P}(D) = \exp \left(-\frac{1}{n} \sum_{d \in D} \sum_{w \in d} n_{wd} \ln P(w|d) \right)$$

- Средняя (по темам) когерентность:

$$C = \frac{1}{|T|} \frac{2}{m(m-1)} \sum_{t \in T} \sum_{i=1}^{m-1} \sum_{j=i+1}^m \text{PMI}(w_i, w_j)$$

где PMI — поточечная взаимная информация

$$\text{PMI}(w_i, w_j) = \ln \frac{P(w_i, w_j)}{P(w_i) P(w_j)}$$

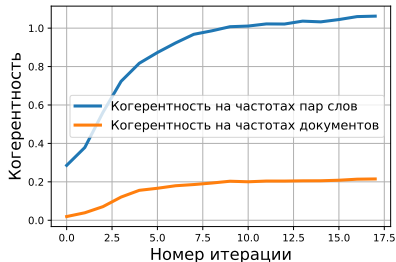
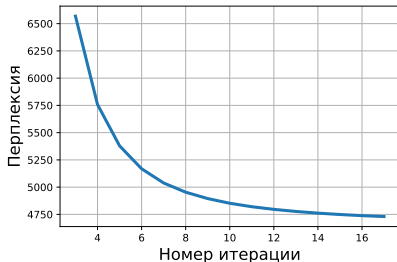
Обоснование PMI как меры интерпретируемости тем

Newman et al показали, что средняя PMI хорошо коррелирует с человеческими оценками интерпретируемости.

Resource	Method	Median	Mean
WordNet	HSO	0.15	0.59
	JCn	-0.2	0.19
	LCh	-0.31	-0.15
	Lesk	<u>0.53</u>	<u>0.53</u>
	Lin	0.09	0.28
	Path	0.29	0.12
	Res	0.57	0.66
	Vector	-0.8	0.27
	WuP	0.41	0.26
Wikipedia	RACO	0.62	0.69
	MiW	0.68	0.70
	DocSim	0.59	0.60
	PMI	<u>0.74</u>	<u>0.77</u>
Google	Titles	<u>0.51</u>	
	LogHits	-0.19	
Gold-standard	IAA	0.82	0.78

Newman, D., Lau, J.H., Grieser, K., Baldwin, T., 2010: Automatic evaluation of topic coherence

Обучение модели PLSA на коллекции статей журнала «NY Times»



- Перплексия падает, она явно минимизируется
- Когерентность растёт, но явно она нигде в функционале не участвует

- Придумать и реализовать эффективный алгоритм подсчёта статистики со-встречаемости по большим коллекциям
- Сравнить разные статистики со-встречаемостей по качеству построенных тематических моделей
- Показать, что можно без серьёзного ухудшения перплексии существенно увеличивать когерентность модели

Со-встречаемость пар токенов:

$$n_{uv} = \sum_{d=1}^{|D|} \sum_{i=1}^{n_d} \sum_{j=1}^{n_d} [0 < |i - j| \leq k] [w_{di} = u] [w_{dj} = v]$$

$$n_u = \sum_{v \in W} n_{uv} \quad n_v = \sum_{u \in W} n_{uv}$$

$$n = \sum_{(u,v) \in W^2} n_{uv}$$

$$\text{PMI}(u, v) = \ln \left[\frac{n_{uv} n}{n_u n_v} \right]$$

Документная со-встречаемость:

$$n_{uv} = \left| \left\{ d \in D \mid \exists (i, j) : w_{di} = u, w_{dj} = v, 0 < |i - j| \leq k \right\} \right|$$

n_u — количество документов, в которых встретился токен u

n_v — количество документов, в которых встретился токен v

n — количество документов всего

$$\text{PMI}(u, v) = \ln \left[\frac{n_{uv}n}{n_u n_v} \right]$$

$$R(\Phi) = - \sum_{t \in T} n_t \text{KL}(\hat{P}(u|t) \parallel \phi_{ut})$$

По формуле полной вероятности:

$$\hat{P}(u|t) = \sum_{v \in W} P_{cooc}(u|v) P(v|t)$$

Формула М-шага:

$$\phi_{wt} = \text{norm}_{w \in W} \left(n_{wt} + \tau \sum_{v \in W \setminus w} P_{cooc}(u|v) n_{vt} \right)$$

Алгоритм сбора статистики со-встречаемостей

3 этапа:

- ❶ Обработка входной коллекции:
 - Загрузка в оперативную память *batch_size* документов
 - Параллельная обработка
 - Сохранение статистики со-встречаемостей по пакетам во внешнюю память в отсортированном формате
- ❷ Слияние файлов с помощью k-Way Merge
 - Если файлов слишком много, многошаговое слияние
- ❸ Вычисление PMI

Реализован в библиотеке тематического моделирования BigARTM

Время работы: $\mathcal{O}(|D| + |W|^2)$

Оперативная память: $\mathcal{O}(1)$

Внешняя память: $\mathcal{O}(|D| \log |D|)$

<https://github.com/bigartm/bigartm>

Эксперимент на корпусе Википедии

Обработка полного текста англоязычной Википедии:

- ≈ 20 Гб текста
- ≈ 8.5 млн статей
- ≈ 8.2 млн уникальных токенов
- ≈ 3.8 млрд токенов всего
- *window_width* = 10

В результате:

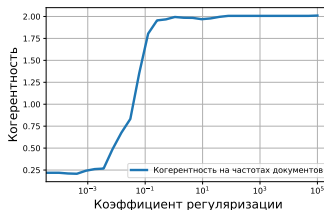
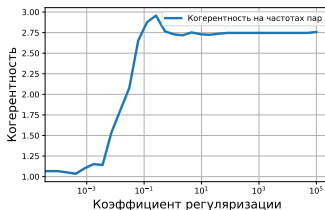
- время работы: 4 часа 18 минут на ноутбуке с 8-ядерным процессором
- около 130 Гб занимают промежуточные файлы

KL регуляризатор, разные типы когерентности

Максимизация средней PMI на частотах пар токенов:



Максимизация средней PMI на частотах документов:



Эксперимент с регуляризацией KL

- 1 Экспоненциальное уменьшение коэффициента регуляризации
- 2 Достаточно большое число тем



Возможные применение данных со-встречаемости в ВТМ

- измерение когерентности тематической модели
- регуляризаторы когерентности
- модель битермов BitermTM
- модель WNTM (Word Network Topic Model)

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Xueqi Cheng. A Biterm Topic Model for Short Texts. WWW 2013.

Yuan Zuo, Jichang Zhao, Ke Xu. Word Network Topic Model: a simple but general solution for short and imbalanced texts. 2014.

A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL-6, 2017.

- Предложен метод повышения когерентности тематических моделей и исследованы условия его применимости
- Предложен и реализован эффективный параллельный пакетный алгоритм для вычисления статистики совместной встречаемости токенов в больших текстовых коллекциях