Банк тем

Сбор интерпретируемых тем с помощью множественного обучения тематических моделей и их дальнейшее использование для оценки качества тематических моделей

В. А. Алексеев, К. В. Воронцов

Московский физико-технический институт (МФТИ)

20-я конференция «Математические методы распознавания образов» (ММРО-2021)

Москва
7 декабря 2021

Содержание

Суть тематического моделирования

Ветер взметнул её чёрные волосы, и Люку мгновенно вспомнилась когда-то виденная картина — «Ведьма» Невинсона. Удлинённое бледное лицо с тонкими чертами, взметнувшиеся к звёздам чёрные волосы. Он буквально представил себе, как эта черноволосая красотка подлетает на помеле к луне...

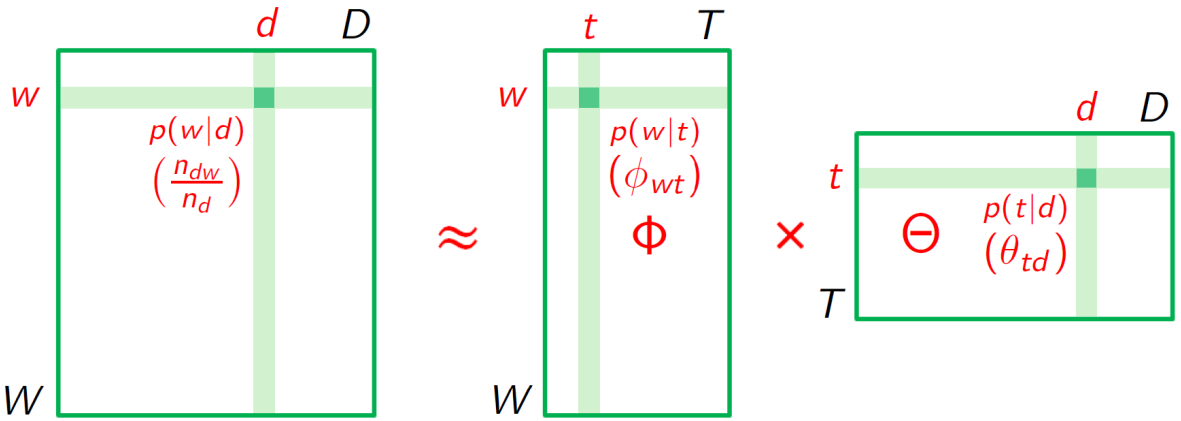
Темы: *человек* *природа* *фэнтези* *искусство*

Возможный пример того, что может получаться в результате работы тематической модели: темы как совокупности слов и сочетаний слов, разметка документа по темам. (Отрывок из книги «Убить легко» Агаты Кристи. Перевод на русский Маргариты Юркан.)

Интерпретация задачи тематического моделирования

- W — слова; D — документы; T — темы
- n_{dw} — частота слова $w \in W$ в документе $d \in D$

Задача стохастического матричного разложения:



(Иллюстрация из лекции К. В. Воронцова
по тематическому моделированию: `https://bit.ly/1bCmE3Z`.)

$$\Phi, \Theta \text{ — решение} \Rightarrow (\Phi S), (S^{-1}\Theta) \text{ — тоже решение}$$

Задача некорректно поставлена: решение не единственно.

Работа

Проблема

- Тематические модели *неполны*.
- Тематические модели *неустойчивы*.
- Часть тем неинтерпретируемы.
- Необходим подбор параметров тематической модели.

Много времени уходит на поиск модели, лучше всего описывающей данные.

Решение

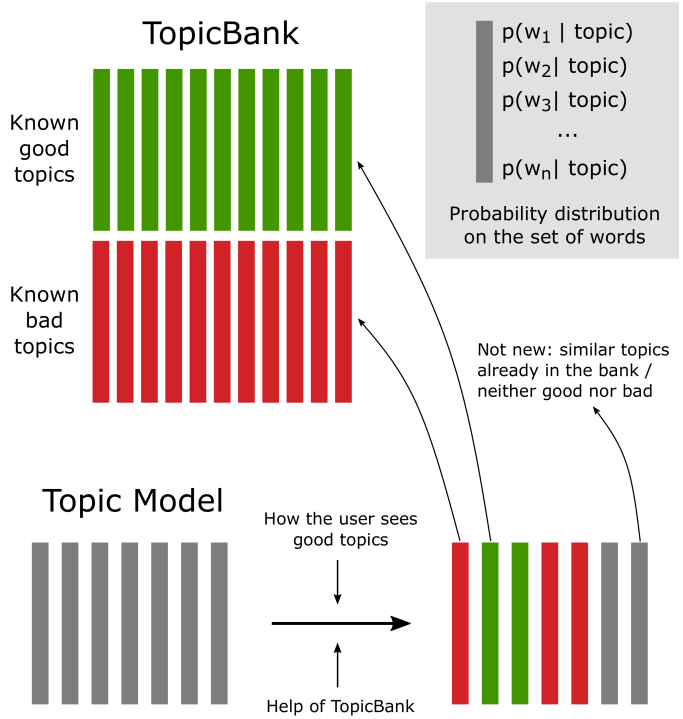
Предложить и реализовать алгоритм, позволяющий сохранять интерпретируемые темы, найденные в процессе поиска лучшей тематической модели, и использовать их для оценки качества вновь обученных тематических моделей.

Содержание

Идея банка тем

Использование банка тем:

1. Создание банка тем с помощью множественного обучения тематических моделей.
2. Оценка качества новых тематических моделей с помощью банка тем.



В банке тем сохраняются хорошие темы. В качестве оценки “хорошести” темы может использоваться функция когерентности темы.

Создание банка тем

Алгоритм

Обучить тематическую модель. Извлечь *хорошие темы* из модели и *сохранить* в банк тем. Повторить N раз.

Ограничения на темы, содержащиеся в банке тем

- Темы интерпретируемые.
- Темы различные.
- Темы составляют хорошую тематическую модель.

Создание банка тем

Алгоритм

Обучить тематическую модель. Извлечь *хорошие темы* из модели и *сохранить* в банк тем. Повторить N раз.

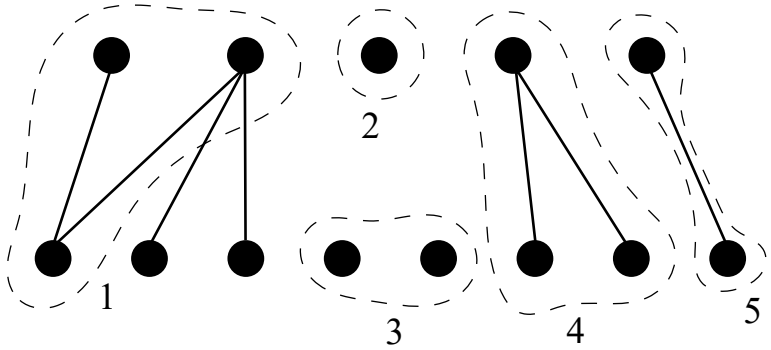
Ограничения на темы, содержащиеся в банке тем

- Темы интерпретируемые.
- Темы различные.
- ~~Темы составляют хорошую тематическую модель~~[1].

[1] Не учитывается в текущей работе.

Добавление темы в банк тем

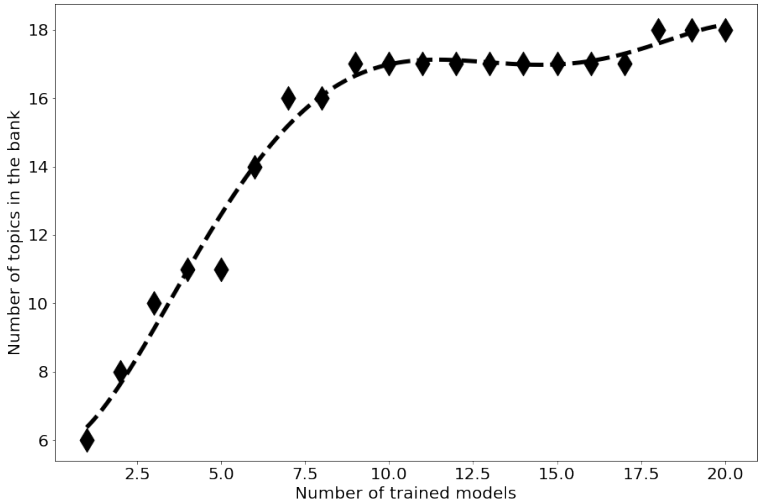
1. Оценка качества тем вновь обученной модели с помощью *когерентности*[2].
2. Оценка зависимостей между темами модели и темами банка тем с помощью построения *двухуровневой иерархической тематической модели*.
3. Хорошие темы *могут быть добавлены* в банк тем в том случае, если темы банка будут оставаться различными.

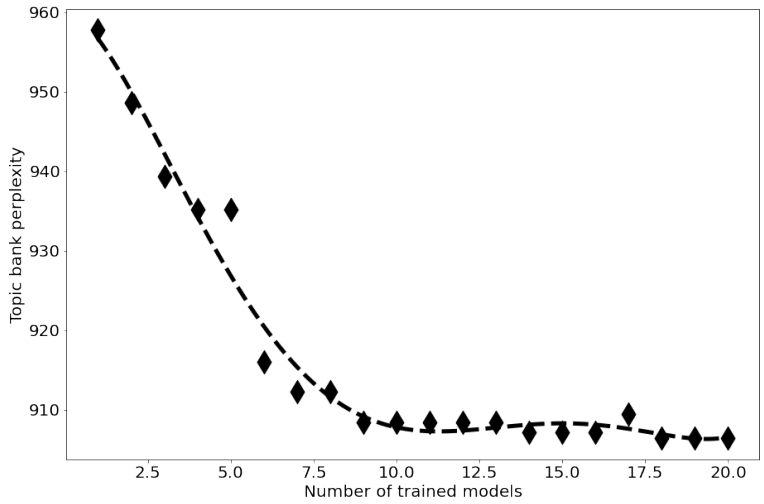


Ситуации, возникающие при добавлении темы в банк тем. Верхний уровень тем — темы банка тем. Нижний уровень тем — темы вновь обученной модели. Возможные ситуации: объединение тем (1), отсутствие дочерних тем (2), отсутствие родительских тем (3), расщепление темы (4), сохранение темы (5).

[2](Alekseev, Bulatov и Vorontsov 2018)

Полнота банка тем





Зависимости характеристик банка тем в зависимости от числа обученных тематических моделей: *слева* — число тем в банке тем; *справа* — перплексия банка тем как тематической модели.

Вывод: с некоторого момента процесс пополнения банка тем выходит на насыщение.

Применение банка тем

Оценка качества модели путём сравнения её тем T и тем B, сохранённых в банке тем:

$$\text{coherence@bank}(T, B) = \frac{\left|\{t \in T \mid \exists \tau \in B : \rho(t, \tau) < h\}\right|}{|T|}$$

Расстояние между темами t_1 и t_2 (мера различия Жаккара):

$$\rho(t_1, t_2) = 1 - \frac{\sum\limits_{w \in \text{Ker}_{12}} \min\limits_{i \in \{1,2\}} p(w|t_i)}{\sum\limits_{\substack{i,j \in \{1,2\} \\ i \neq j}} \sum\limits_{w \in \text{Ker}_i \setminus \text{Ker}_j} p(w|t_i) + \sum\limits_{w \in \text{Ker}_{12}} \max\limits_{i \in \{1,2\}} p(w|t_i)}$$

где $\text{Ker}_i \equiv \text{Ker}(t_i)$, $\text{Ker}_{12} \equiv \text{Ker}(t_1) \cap \text{Ker}(t_2)$, а $\text{Ker}(t)$ – ядро темы t, то есть: $\text{Ker}(t) = \left\{w \in W : p(w \mid t) > 1/|W|\right\}$.

Содержание

Эксперимент

Цель

Понять, можно ли использовать банк тем для оценки качества тематических моделей.

Задача

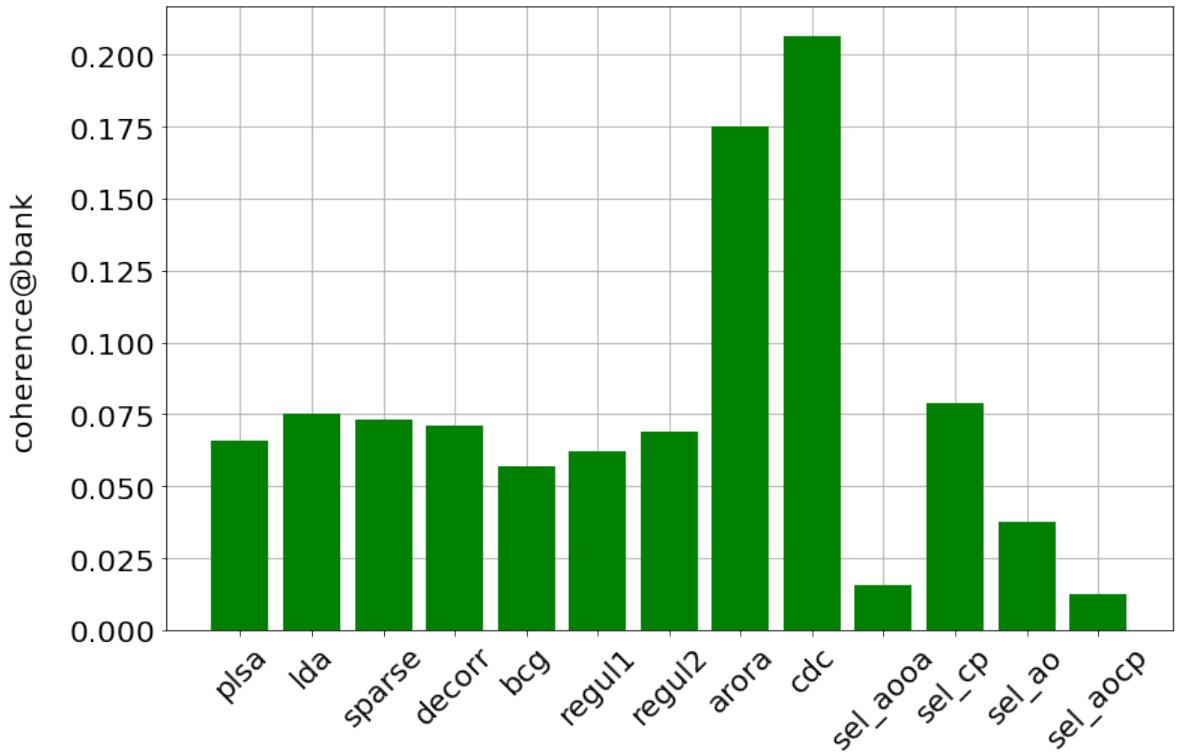
Проверить, позволяет ли банк тем найти лучшую модель из фиксированного множества моделей.

Постановка

- Несколько текстовых коллекций: PostNauka (RU), Reuters (EN), Brown (EN), Twenty Newsgroups (EN), AG News (EN), Habrahabr (RU), Watan2004 (AR).
- Создание банка тем для каждой текстовой коллекции.
- Набор моделей: PLSA, LDA, ARTM[3], Arora, CDC.
- Оценка качества моделей с помощью банков тем.

[3](Hofmann 1999; Blei, Ng и Jordan 2003; Vorontsov и др. 2015)

Результаты



Усреднённые оценки качества моделей, рассчитанные с помощью банков тем текстовых коллекций. Горизонтальная ось — тематическая модель. Вертикальная ось — средняя доля хороших тем модели, посчитанная с помощью банков тем (чем больше, тем лучше). *Модели arora и cdc*[4] *выявлены Банком тем как модели, позволяющие найти наибольшее количество интерпретируемых тем.*

[4](Arora и др. 2012; Dobrynin, Patterson и Rooney 2004)

Содержание

Заключение

Сделано в работе

- Представлен Банк тем: “обёртка” над тематическим моделированием, ускоряющая валидацию вновь обученных тематических моделей.
- Предложен и реализован алгоритм автоматического создания Банка по данному набору текстов.
- Проведён эксперимент на реальных данных, подтверждающий возможность применения Банка тем для оценки качества тематических моделей.

- **Публикация**: Alekseev V. et al. TopicBank: Collection of coherent topics using multiple model training with their further use for topic model validation //*Data & Knowledge Engineering*. – 2021. – Т. 135. – С. 101921. – DOI.
- **Репозиторий**: https://github.com/machine-intelligence-laboratory/OptimalNumberOfTopics.