Банк тем: сбор интерпретируемых тем с помощью множественного обучения тематических моделей и их дальнейшее использование для оценки качества тематических моделей

В. А. Алексеев

Научный руководитель: Воронцов Константин Вячеславович

Московский физико-технический институт (МФТИ)

64-я научная конференция МФТИ
4 декабря 2021

Содержание

Тематическое моделирование

- W — конечное множество слов
- D — конечное множество документов
- T — конечное множество тем
- $\Phi_{W \times T}$ — матрица вероятностей слов в темах
  $(\varphi_{wt} = p(w \mid t),\ \sum\limits_{w \in W} \varphi_{wt} = 1,\ \varphi_{wt} \geqslant 0)$
- $\Theta_{T \times D}$ — матрица вероятностей тем в документах
  $(\theta_{td} = p(t \mid d),\ \sum\limits_{t \in T} \theta_{td} = 1,\ \theta_{td} \geqslant 0)$
- $F_{W \times D}$ — матрица частот слов в документах

Задача стохастического матричного разложения:

$$F = \Phi\Theta$$

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

Задача *некорректно поставлена*: решение не единственно.

Работа

Проблема

- Тематические модели *неполны*.
- Тематические модели *неустойчивы*.
- Часть тем неинтерпретируемые.
- Необходим подбор параметров тематической модели.
- *Много времени уходит на поиск модели, лучше всего описывающей данные*.

Решение

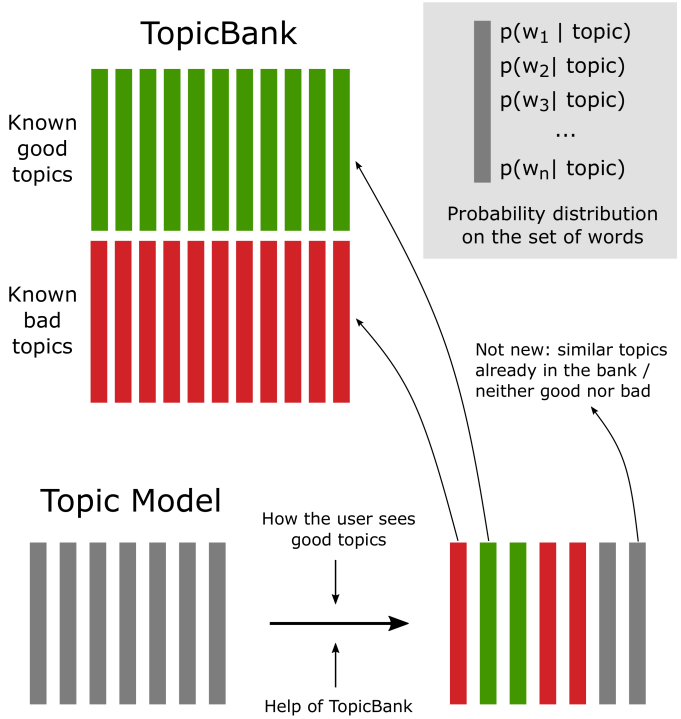
Предложить и реализовать алгоритм, позволяющий сохранять интерпретируемые темы, найденные в процессе поиска лучшей тематической модели, и использовать их для оценки качества вновь обученных тематических моделей.

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

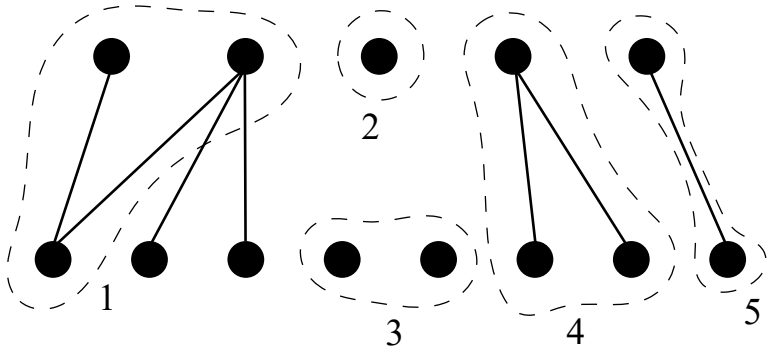
Обучить тематическую модель. Извлечь хорошие темы из модели и сохранить в банк тем. Повторить N раз.

Ограничения на темы, содержащиеся в банке тем

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

Добавление темы в банк тем

1. Оценка качества тем вновь обученной модели с помощью *когерентности*.
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).

Содержание

Эксперимент

Цель

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

Задача

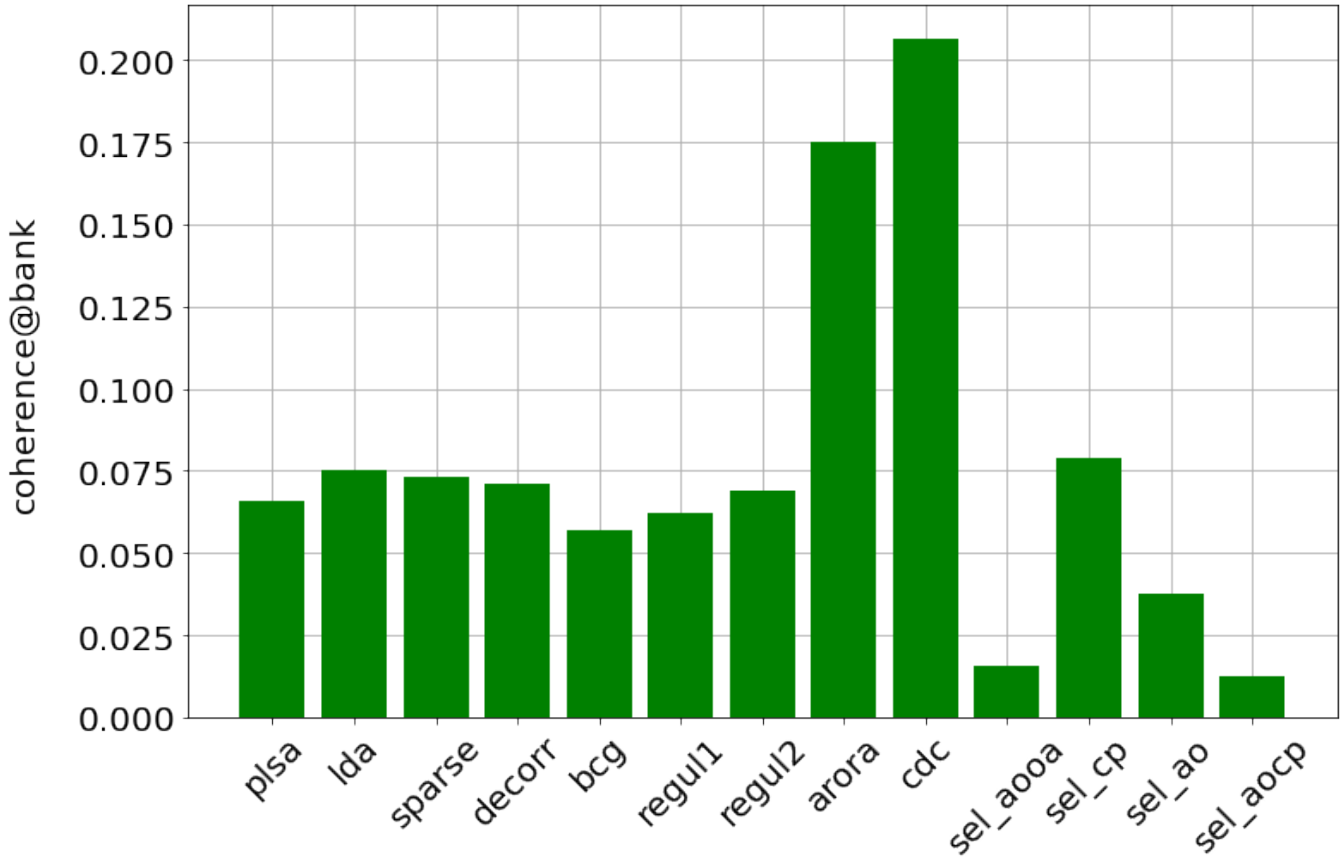
Проверить, позволяет ли банк тем найти лучшую модель из фиксированного множества моделей.

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[1], Arora, CDC.
- Оценка качества моделей с помощью банков тем.

[1] (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. *Модели arora и cdc выявлены Банком тем как модели, позволяющие найти наибольшее количество интерпретируемых тем.*

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