Автоматическая настройка параметров BigARTM под широкий класс задач

Гришанов А. В.

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

Задачу поставил к.ф.-м.н., н.с. ВЦ РАН К. В. Воронцов
Консультант Мурат Апишев

Москва,
2019 г.

Цель работы

Проблема

BigARTM — продвинутая бибиотека для тематического моделирования. В ней реализовано много регуляризаторов, что повышает гибкость, но при этом усложняет настройку параметров. В результате на практике популярнее более простые методы, такие как LDA.

Цель работы

Проверить гипотезу о существовании конфигураций, хорошо работающих на широком классе задач.

Базовые алгоритмы

1. Логистическая регрессия с L_1 регуляризацией
2. PLSA
   David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan):993–1022, 2003.
3. LDA Thomas Hofmann. Probabilistic latent semantic analysis. In Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, UAI'99, pages 289–296, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

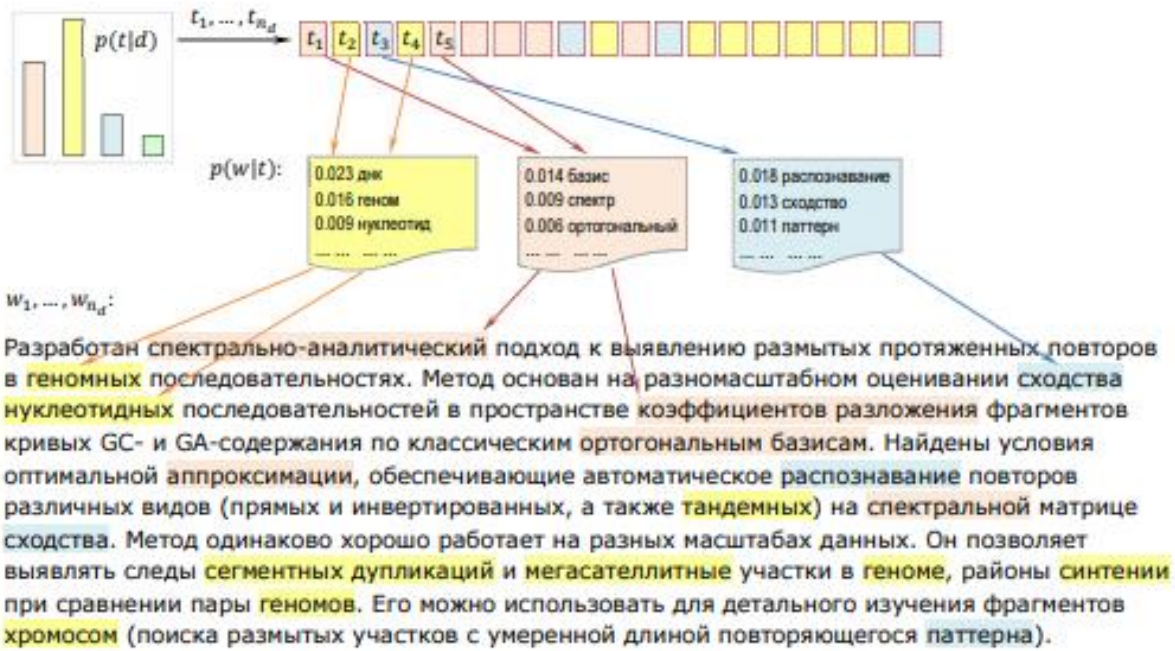
Стартовые обозначения

- D — коллекция текстовых документов, состоящая из документов d
- W — словарь, состоящий из терминов w;
- T — множество тем, состоящее из тем t.

Согласно формуле полной вероятности и гипотезе условной независимости, распределение термов в документе $p(w|d)$ описывается вероятностной смесью распределений термов в темах $\varphi_{wt} = p(w|t)$ с весами $\theta_{td} = p(t|d)$

Распределение термов в документах

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td}. \qquad (1)$$

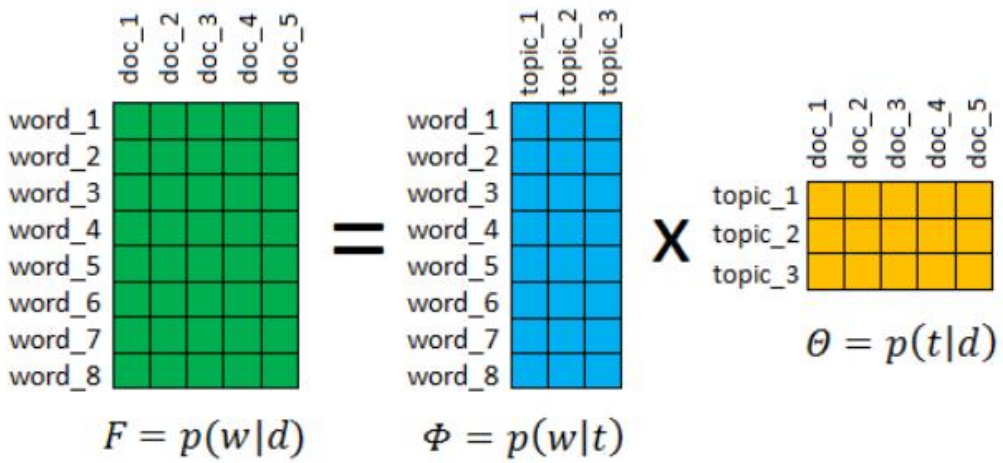


Распределение термов в документах

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td}. \qquad (2)$$

Постановка задачи тематического моделирования

Ставится задача разложения матрицы F в произведение двух матриц Φ и Θ меньшего размера



Поставленная задача ($F \approx \Phi\Theta$) эквивалентна поиску матриц Φ и Θ, максимизирующих следующий функционал:

Задача

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \sum_{t \in T} \phi_{wt} \theta_{td} \to \max_{\Phi, \Theta} \qquad (3)$$

Сложности, возникающие при решении задачи

Проблема

Разложение матрицы F в произведение матриц Φ и Θ не единственно. В частности, для любой невырожденной матрицы S размера $T \times T$ верно, что $F = (\Phi S)(S^{-1}\Theta)$.

Таким образом, из-за сложившейся неопределённости, невозможен поиск произвольного матричного разложения, нужно уточнять модель. Можно наложить дополнительные ограничения, что приведёт к сокращению произвольности выбора или же сделать некоторые предположения о вероятностном распределении коллекции. Рассматриваются следующие два подхода к решению проблемы:

Подходы к устранению неопределённости

Латентное размещение Дирихле

Тематическая модель латентного размещения Дирихле (latent Dirichlet allocation, LDA) основана на дополнительном предположении, что векторы документов $\theta_d = (\theta_{td}) \in \mathbb{R}^{|T|}$ и векторы тем $\phi_t = (\phi_{wt}) \in \mathbb{R}^{|W|}$ порождаются распределениями Дирихле с параметрами $\alpha \in \mathbb{R}^{|T|}$ и $\beta \in \mathbb{R}^{|W|}$.

Аддитивная регуляризация

Тематическая модель аддитивной регуляризации (additive regularization of topic models, ARTM) получается при наложении на модель дополнительных требований (регуляризаторов).

$$L(\Phi, \Theta) + \sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta) \to \max_{\Phi, \Theta} \tag{4}$$

Решение

- Рассмотрим набор датасетов $\{\mathfrak{D}_{ex}, \mathfrak{D}_{in}\}$, где $\mathfrak{D}_{ex}$ имеют внешний критерий качества, а $\mathfrak{D}_{in}$ — только внутренние.
- Необходимо проверить гипотезу о том, что существуют коэффициенты регуляризации $\tau_{general}$, которые можно считать «универсальными», т.е. для которых метрики качества отличаются от лучших на том же датасете не более чем на 5%.
- Для каждого из первых найдём лучшие параметры, затем будем искать общие.
- В конце проверим выполнение гипотезы на всех данных, для неразмеченных будем сравнивать внутренние критерии качества.

Цель работы — построить модель, которая **не хуже** чем PLSA и **лучше** PLSA по нескольким критериям.

Результаты эксперимента

1. **20news groups**
   Best f1_score: 0.9155
   General params f1_score: 0.9148
2. **Victorian Era**
   Best f1_score: 0.9777
   General params f1_score: 0.9777
3. **Toxic comments**
   Best f1_score: 0.9539
   General params f1_score: 0.9582