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 д.ф.-м.н., К. В. Воронцов
Консультант Виктор Булатов

Москва,
2019 г.

Цель работы

Проблема

Настройка параметров BigARTM требует работы аналитика (эксперта). Требуется автоматизировать этот процесс.

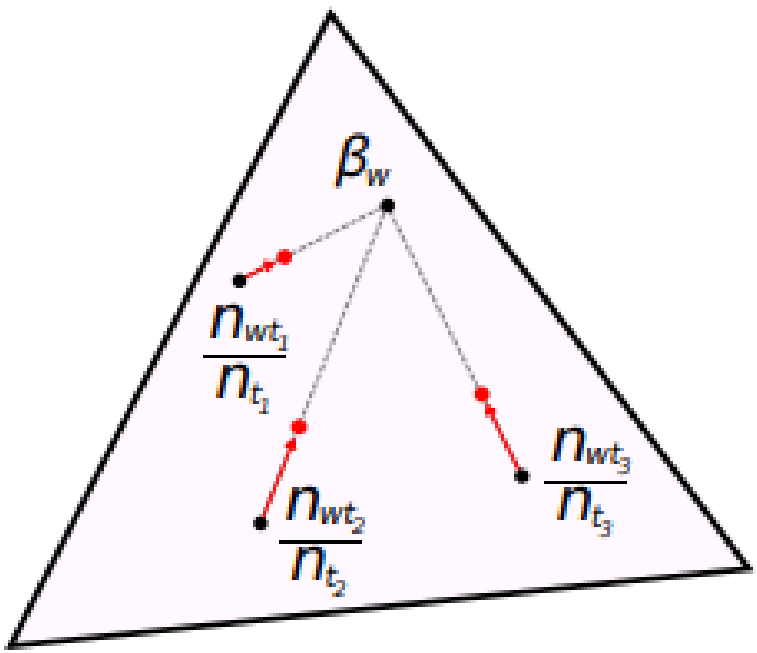
Цель работы

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

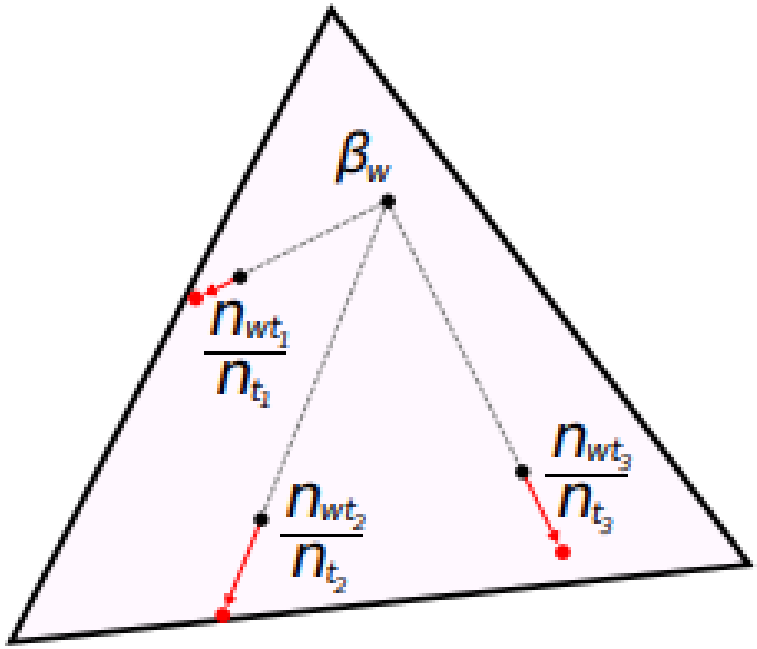
Метод решения

Предлагается использовать относительные коэффициенты регуляризации и автоматический подбор n-граммм.

Относительные коэффициенты регуляризации



(a) Сглаживание



(b) Разреживание

Обозначения

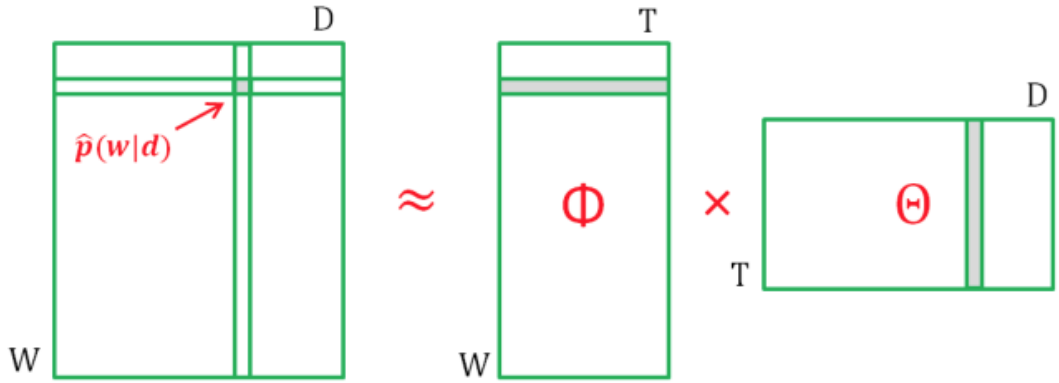
- $d \in D$ — текстовые документы
- $w \in W$ — слова
- $t \in T$ — темы

Распределение термов в документах

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

Задача 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

$$L(\Phi,\Theta) = \sum_{d\in D}\sum_{w\in d} n_{dw} \sum_{t\in T} \phi_{wt}\theta_{td} \to \max_{\Phi,\Theta} \qquad (2)$$

Постановка задачи ARTM

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

Аддитивная регуляризация

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

$$L(\Phi, \Theta) + \sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta) \to \max_{\Phi, \Theta} \tag{3}$$

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 общие коэффициенты регуляризации $\tau_{general}$, 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.

Критерий — 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