1 Постановка задачи

1.1 Общая постановка задачи тематического моделирования

Пусть D - множество документов, T - конечное множество тем, W - множество термов. Каждый из документов $d \in D$ задается его длиной n_d, $\sum_{d \in D} n_d = n$ и последовательностью термов $\{w_i \in W\}_{i=1}^{n_d}$, элементы которой в дальнейшем будем называть словами. Вероятностная модель порождения коллекции вводится при следующих дополнительных предположениях:

- Гипотеза мешка слов: вышеописанное представление документа d эквивалентно представлению документа d в виде неупорядоченного множества входящих в него слов, в которое каждое слово w входит n_{dw} раз.
- Гипотеза условной независимости: вероятность появления слова w в документе d по теме t не зависит от документа d и описывается распределением
$$p(w|d,t) = p(w|t)$$

С учетом данных предположений вероятность появления слова w в документе d описывается распределениями $p(w|t) = \phi_{wt},\ p(t|d) = \theta_{td}$. Задача тематического моделирования заключается в нахождении этих распределений. Это эквивалентно задаче получения матричного разложения

$$F = \Phi\Theta \tag{1}$$

$$F = \left(\frac{n_{wd}}{n_d}\right)_{W \times D} \quad \Phi = (\phi_{wt})_{W \times T} \quad \Theta = (\theta_{td})_{T \times D}$$

Ставится задача максимизации функции правдоподобия. Данную задачу решают с помощью EM-алгоритма.

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} \to \max_{\Phi,\Theta} \tag{2}$$

Задача (1) поставлена некорректно: в общем случае ее множество решений бесконечно. Для выделения из этого множества решений одного в функцию (2) добавляют один или несколько регуляризаторов, зависящих

от матриц Φ, Θ. Вид регуляризаторов определяется тем, какие свойства ожидаются от результата. Функция правдоподобия при этом принимает следующий вид:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + \sum_{i} \tau_i R_i(\Phi, \Theta) \to \max_{\Phi, \Theta} \tag{3}$$

1.2 Проблема несбалансированности

Результатом, который выдает вышеописанная модель являются две матрицы Φ, Θ. Они описывают порождение коллекции: матрица Φ показывает вероятность конкретного слова в документе с заданной темой, а матрица Θ показывает распределение тем между документами. Данный алгоритм уже исследовался в литературе, в частности в [?] показывалось, что устойчивость результата зависит от априорного представления о матрицах Φ, Θ. Выданные такой моделью матрицы Θ часто свидетельствуют о равенстве мощностей всех тем, то есть распределение $p(t)$ определенное как $p(t) = \sum\limits_{d \in D} p(t|d) n_d$ не сильно отличается от равномерного, а именно $\forall\ t_i, t_j \in T \to \frac{p(t_1)}{p(t_2)} \approx 1$ вне зависимости от истинного вида $p(t), t \in T$, что было упомянуто ранее как "проблема несбалансированности". Такой эффект возникает из-за используемого алгоритма: при максимизации правдоподобия модели выгодно использовать все свои параметры.В свою очередь, сокращение долей отдельных тем приводит к неполному использованию, а в пределе - к уменьшению числа параметров. В реальных же коллекциях темы могут оказаться несбалансированными. Чтобы повысить качество решения для несбалансированных коллекций, в модель предлагается добавить регуляризатор.

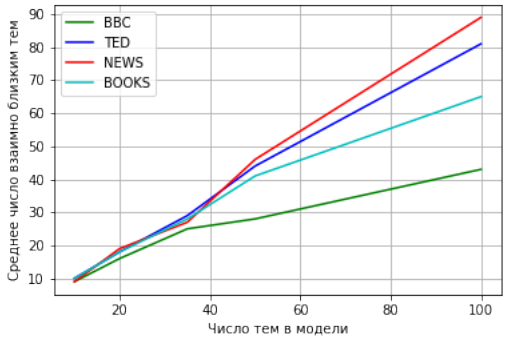
1.3 Неоднозначность матричного разложения

Даже в случаях, когда алгоритм достигает максимума функции правдоподобия, решение может оказаться не единственным. В таком случае говорят о неустойчивости решения задачи. Даже учитывая регуляризацию матрицы Φ', Θ' могут существенно отличаться от исходных. Для корректного отображения результатов эксперимента необходимо, чтобы все возможные матрицы Φ', Θ' получались перестановкой тем исходных матриц.

Предполагается, что это будет выполнено при достаточной разреженности исходных матриц. Для того чтобы в процессе обработки коллекции свойство разреженности модель была дополнена регуляризатором.

2 Эксперимент

Чтобы проверить факт того, что модель с регуляризатором выдает одни и те же результаты с точностью до перестановки было выбрано несколько реальных коллекций: новости BBC(BBC), ted talks (TED), 20newsgroups(NEWS) и описания книг(books). На каждой из них была построена модель по 5 раз для каждого числа тем: 10,20,35,50,100. Рассматривалось среднее число взаимно близких тем между каждой парой результатов для одной коллекции и одного числа тем. Результаты приведены на графике ниже.



Из графика видно, что не на всех коллекциях модель выдала близкие результаты. Однако это связано с тем, что не все коллекции разбиваются на 100 интерпретируемых тем. Например, оптимальное число тем для коллекции BBC меньше 10.