Вероятностное тематическое моделирование несбалансированных текстовых коллекций.

Виктор Панкратов
научный руководитель: д.ф.-м.н. Воронцов К.В.

19 мая 2021 г.

1 Постановка задачи

1.1 Общая постановка

Пусть D - множество документов, T - конечное множество тем. Каждый из документов $d \in D$ задается его длиной n_d 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 эквивалентно представлению документа в виде неупорядоченного множества входящих в него слов, в которое каждое слово w входит n_{wd} раз.
- Гипотеза о существовании тем: каждое вхождение слова в документ связано с некоторой темой $t \in T$
- Гипотеза условной независимости: вероятность появления слова w в документе d по теме t не зависит от документа d и описывается распределением
$$p(w|d,t) = p(w|t)$$

При таких условиях вероятность появления слова w в документе d описывается распределениями $p(w|t) = \phi_{wt}, p(t|d) = \theta_{td}$. Задача тематическо-

го моделирования заключается в нахождении этих распределений. Это эквивалентно задаче получения матричного разложения

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

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

Данную задачу решают максимизацией с помощью EM-алгоритма функции правдоподобия

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

Задача (1) поставлена некорректно: в общем случае ее множество решений бесконечно. Чтобы уменьшить множество решений, в функцию (2)добавляют один или несколько регуляризаторов, зависящих от матриц Φ, Θ. 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

Вышеописанная модель склонна выделять равномощные темы [**?**], то есть при n_t определенном как $n_t = \sum\limits_{d\in D} p(t|d)n_d \Rrightarrow \forall\ t_i, t_j \in T \to \frac{n_{t_1}}{n_{t_2}} \approx 1$, что получило название "проблема несбалансированности". Такой эффект возникает из-за изначальной постановки задачи: 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 модель и в таком случае корректно выделила темы, в нее предлагается добавить регуляризатор R:

$$R = \sum_{d\in D}\sum_{w\in d} \beta_{dw} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} \tag{4}$$

$$\beta_{dw} = \sum_{t\in T} \frac{p(t|d,w)}{p_t} \qquad p_t = \frac{n_t}{n}$$

Введение данного регуляризатора эквивалентно требованию минимизации суммарной семантической неоднородности тем.

$$\sum_{t\in T} S_t = \sum_{d\in D}\sum_{w\in d}\left(\sum_{t\in T}\frac{n_{tdw}}{n_t}\right)\ln\frac{\hat{p}(w|d)}{p(w|d)} \to \min_{\Phi,\Theta} \tag{5}$$

В данной работе будет описана мера качества работы модели и с ее помощью оценено влияние R на получаемое решение задачи тематического моделирования для синтетических коллекций с различным балансом тем.

2 Эксперимент

2.1 Генерация коллекции

Для эксперимента будем генерировать синтетическую коллекцию данных. Процесс генерации можно разделить на 2 этапа: генерация матриц Φ, Θ и построение документов по ним.

2.1.1 Генерация матриц

Столбцы матриц Φ, Θ порождаются симметричными распределениями Дирихле. Параметр распределения определяется из соображений реалистичности коллекции и берется малым для разреженности получаемых матриц. Для матрицы Φ он берется равным ≈ 0.02, для матрицы $\Theta \approx 0.2$. Чтобы регулировать баланс тем будем на этом этапе генерации менять наибольшие значения в столбцах Θ с необходимыми для эксперимента. После этого в обе матрицы добавляется еще одна "фоновая"тема, порожденная несимметричным распределением Ципфа. Матрица Θ перед этим перенормируется в зависимости от желаемой доли фоновой темы в документах. В дальнейших экспериментах, если это не обговорено, доля фоновой темы принимается равной 0.5.

2.1.2 Генерация документов

Для генерации очередного слова w_i сначала генерируется тема t_i документа из соответствующего этому документу столбцу матрицы Θ. Затем

слово генерируется из столбца Φ, соответствующего теме t_i. Таким образом, процесс генерации документов описывается как

$$t_i \sim Dir(t|d) \;\; w_i \sim Dir(w|t_i), i \in 1 \dots n_d$$

2.2 Стандартная модель

Сгенерируем вышеописанным образом несколько коллекций с различной степенью несбалансированности: число документов, соответствующих одной из тем в ней будет значительно превышать число, соответствующее остальным темам. Для каждой из них используем стандартную модель для находения матриц Φ, Θ. Чтобы оценить сходство полученных матриц Φ_{exp} с используемыми при генерации в данной работе считается количество взаимно ближайших по некоторой метрике столбцов матриц Φ, Φ_{exp}, то есть пар столбцов $\Phi[i]$,$\Phi_{exp}[j]$:

$$\arg\min_k \;\; (dist(\Phi[i], \Phi_{exp}[k])) = j$$

$$\arg\min_k \;\; (dist(\Phi[k], \Phi_{exp}[j])) = i$$

$dist$ в формулах выше - расстояние по заданной метрике. На данный момент в качестве метрики выбрано расстояние Йенсена-Шеннона. Будем полагать, что если вышеописанная оценка сходства низкая, исходная матрица Φ плохо восстанавливается и, напротив, если она высока, то и матрица Φ восстановлена хорошо.

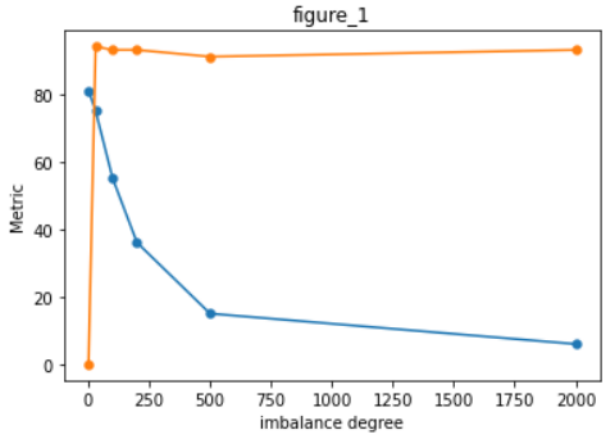
На графиках синей линией представлены результаты описанного эксперимента для различных долей слов из фоновой темы [1]. По оси ординат отложено число взаимно ближайших пар(100 - идеальный результат), по оси абсцисс - степень несбалансированности коллекции, то есть отношение наибольшего и наименьшего числа документов, соответствующих одной теме. Видно, что при увеличении степени несбалансированности качество решения в описанном выше смысле падает.

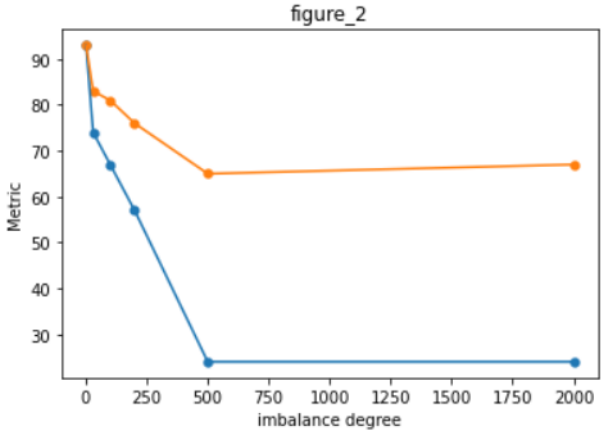
2.3 Добавление регуляризатора

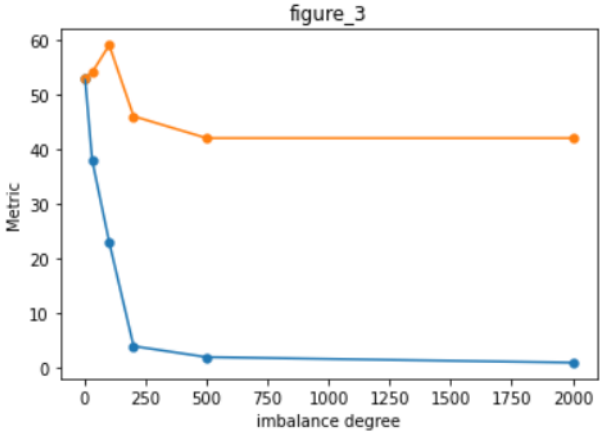
Теперь рассмотрим, как введение регуляризатора R улучшает качество модели. Добавим регуляризатор R в модель и повторим предыдущий

[1] 0.3 для первого графика, 0.5 для второго графика, 0.7 для третьего графика

эксперимент. В данном эксперименте использовался постоянный коэффициент регуляризации $\tau = 0.3$. Результаты представлены на графиках оранжевой линией.



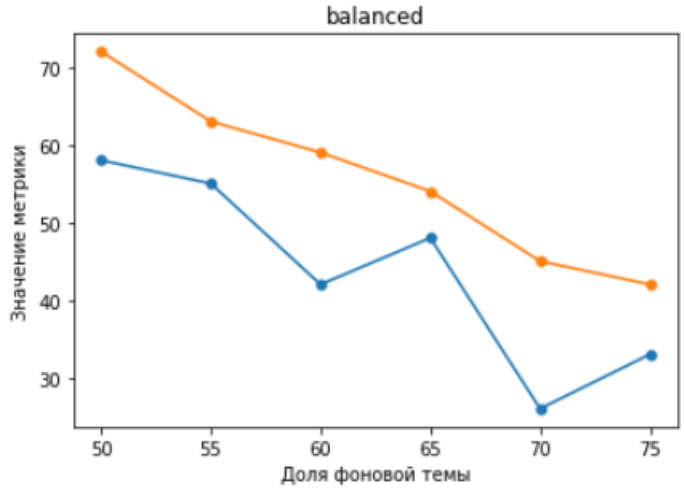




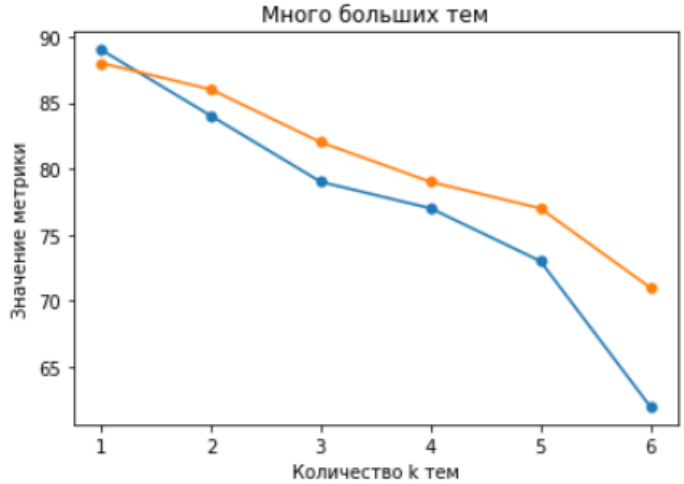
2.4 Дополнительные эксперименты

Далее сгенерируем вышеописанным способом несколько коллекций и проведем исследование для них.

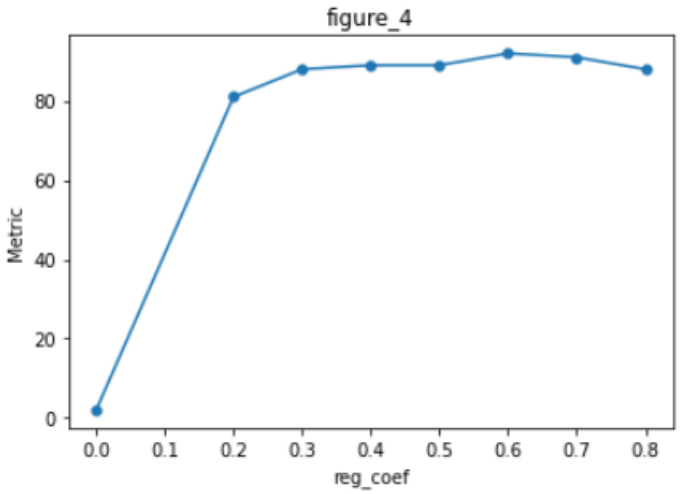
Проверим, что регуляризатор не уменьшает качество работы модели на сбалансированных коллекциях. Для этого сгенерируем несколько сбалансированных коллекций с различной долей фоновой темы в документах. Результаты представлены на графике 'balanced' ниже.



Проверим, как модель отработает, если больших по мощности тем будет несколько. Для этого построим коллекцию, где k темам принадлежит одинаковая доля документов:по 10-15 процентов на каждую из тем.Оставшиеся документы поровну распределена по остальным темам. Зависимость качества решения от k для нулевого и ненулевого τ представлена на графике ниже.



Проведем также эксперимент на коллекции, где юольшинтсво тем равномощные, но среди остальных ест как большие так и меньшие по мощности. Исследуем зависимость результата от коэффициента регуляризации. Результат представлен на графике ниже.



Наконец, проведем эксперимент, где мощность первых 50 тем равен i для темы с номером i, а остальные распределены равномерно и также исследуем завидимость результата от коэффициента регуляризации. Результат представлен на графике ниже.