



МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ
имени М.В.Ломоносова

Факультет вычислительной математики и кибернетики



Спецкурс

«Вероятностные тематические модели»
Поиск оптимальных методов применения
регуляризаторов при тематическом
моделировании
Отчет

о выполненном задании

студента 201 учебной группы факультета ВМК МГУ
Долгушева Глеба Дмитриевича

гор. Москва

2024

Содержание

Введение	2
Цель и задачи работы	2
Постановка задачи	3
Описание используемых метрик	4
Построение baseline	5
Информация о данных	5
Предобработка текстов	5
Описание базовой модели	6
Определение методики применения регуляризаторов	7
Описание используемых регуляризаторов	7
Сравнение работы регуляризаторов при их включении в начале обучения модели	8
Декоррелирование распределения термов в темах	8
Разреживание или сглаживание распределений термов в темах	9
Неподходящие регуляризаторы	9
Поиск изменений в работе регуляризаторов при их последовательном включении	10
Второй регуляризатор	10
Третий регуляризатор	12
Сравнение построенной методики с более простыми методами	13
Выводы	15
Литература	16

Введение

Данная работа посвящена подбору гиперпараметров для построения тематической модели с использованием библиотеки BigARTM.

Тематическое моделирование ставит перед собой задачу сопоставления имеющихся документов набору тем фиксированной длины. Кроме того, в процессе обучения для каждой темы определяется набор слов, присущих ей. В основе же моделирования лежит алгоритм, основанный на вероятностном подходе.

В ходе прохождения спецкурса «Вероятностные тематические модели» была показана методика использования регуляризаторов для улучшения информационного поиска. Результат ее применения заключался в возможности построить последовательность включения регуляризаторов таким образом, чтобы улучшить показания сразу нескольких, казалось бы независимых, метрик. Автор попытается повторить это наблюдение и ответить на следующие вопросы:

- Является ли полученный метод единственным верным, или для каждого конкретного набора документов методика (и не только параметры регуляризаторов) может меняться?
- Какого влияние выбранного количества эпох для обучения модели на возможность построения такой методики?
- Как можно с точки зрения математической модели объяснить полученные результаты?

Цель и задачи работы

Используя подход аддитивной регуляризации, реализованный в BigARTM построить тематическую модель для набора научных статей. Провести исследование поведения модели при различных параметрах и методах применения регуляризаторов. Получить последовательность регуляризаторов и набор параметров, позволяющих улучшать одновременно несколько метрик.

Выполнение цели разобъем на представленные задачи:

- Составить простое решение, используя рекомендации, полученные в ходе прохождения спецкурса
- Исследовать поведение регуляризаторов в отдельности
- Построить оптимальную с точки зрения улучшения сразу нескольких метрик последовательность использования регуляризаторов
- Сравнить работу построенного метода с простым решением

Постановка задачи

Пусть D - конечный набор документов, называемый коллекцией, а W набор слов, из которых состоят тексты (словарь). При этом появление каждого терма w в документе d связано с некоторой скрытой переменной $t \in T$, где T - конечное множество тем, к которым могут принадлежать термы. Так как данное множество не известно до начала моделирования, то оно задается исключительно своей мощностью. Коллекция документов рассматривается как множество троек (d, w, t) , взятых случайно и независимо из дискретного распределения $p(d, w, t)$, заданного на конечном множестве $D \times W \times T$.

Таким образом, используется гипотеза *bag of words* (порядок слов в документе не важен для определения его темы). Документ $d \in D$ в этом случае можно представлять как набор частот n_{dw} встречаемости терма w в нем.

Построить тематическую модель коллекции документов D - значит найти множество тем T , распределения $p(w|t)$ для всех тем $t \in T$ и распределения $p(t|d)$ для всех документов $d \in D$.

Подробное описание алгоритма приведено в [1], здесь же обратим внимание только на основные моменты. Приведем список допущенных гипотез:

- **Гипотеза условной независимости.** Полагаем, что появление слов в документе d , относящихся к теме t не зависит от документа. Это предположение допускает переходы:

$$\begin{aligned} p(w|d, t) &= p(w|t); \\ p(d|w, t) &= p(d|t); \\ p(d, w|t) &= p(d|t)p(w|t); \end{aligned}$$

- **Гипотеза разреженности.** Естественно предполагать, что каждый документ из коллекции связан лишь с небольшим числом тем, а каждая тема содержит в себе далеко не все термы. Следовательно, для фиксированной темы t и большой части термов w и документов d условные вероятности $p(w|t)$ и $p(t|d)$ должны обращаться в ноль.
- **Частотные оценки условных вероятностей.** Вероятности, связанные с наблюдаемыми переменными d и w , можно оценивать по выборке как частоты (здесь и далее выборочные оценки вероятностей будем обозначать через \hat{p}). Аналогичный подход можно применить и к вероятностям, связанным с латентными переменными.

$$\hat{p}(w, d) = n_{dw}/n, \quad \hat{p}(w|t) = n_{tw}/n_t$$

и т.д.

Введем следующие обозначения: $\phi_{wt} = p(w|t)$ и $\theta_{td} = p(t|d)$. Если число тем $|T|$ много меньше числа документов $|D|$ и мощности словаря $|W|$, то можно рассматривать поставленную задачу, как задачу разложения матрицы $F = (\hat{p}_{wd})_{W \times D}$ в произведение двух неизвестных матриц

$$\Phi = (\phi_{wt})_{W \times T} \quad \Theta = (\theta_{td})_{T \times D} : \quad F \approx \Phi\Theta$$

Нахождение матриц Φ и Θ осуществляется методом максимального правдоподобия:

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

При использовании подхода аддитивной регуляризации максимизируемый функционал приобретает следующий вид:

$$\log L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in W} n_{dw} \log \sum_{t \in T} \phi_{wt} \theta_{td} + \sum_i \tau_i R_i(\Theta, \Phi) \longrightarrow \max_{\Phi, \Theta},$$

где $R_i(\Theta, \Phi)$ - некоторый регуляризатор с параметром τ_i . Для удобства $R(\Theta, \Phi) = \sum_i \tau_i R_i(\Theta, \Phi)$.

Это позволяет решить проблему не единственности матричного разложения и поставить задачу корректно по Адамару.

Решить данную задачу можно используя итеративный подход, а именно EM -алгоритм:

E -шаг:

$$p_{tdw} = \text{norm}_{t \in T}(\phi_{wt} \theta_{td})$$

M -шаг:

$$\begin{aligned} \phi_{wt} &= \text{norm}_{w \in W}(n_{wt} + \phi_{wt} \frac{\partial R(\Theta, \Phi)}{\partial \phi_{wt}}) \\ \theta_{td} &= \text{norm}_{t \in T}(n_{td} + \theta_{td} \frac{\partial R(\Theta, \Phi)}{\partial \theta_{td}}) \end{aligned}$$

где

$$\text{norm}(x_i) = \frac{\max(x_i, 0)}{\sum_k \max(x_k, 0)}$$

Описание используемых метрик

Описанные гипотезы и поставленная задача определяют набор критериев, позволяющих оценивать сходимость процесса и качество тематического моделирования. Безусловно, наилучший критерий - это интерпретируемость полученных тем, но его проверка возможна только с участием ассессоров, что затруднительно. Поэтому выделяются более простые условия корректности тематического моделирования:

- Темы должны быть различимы в смысле наиболее частотных для них слов
- Матрицы Φ и Θ должны быть разрежены (В данном исследование не предполагается построение фоновых тем, а потому слова, не играющие роли в определении тем будут удалены на этапе предобработки).

В соответствии с этими условиями выбраны следующие метрики:

1. Разреженность матриц Φ и Θ - доля нулевых элементов. Как уже сказано выше, чем больше, тем лучше (до порока примерно в 95%).

2. Перплексия - это величина, которая может быть интерпретирована как способность модели предсказывать следующее слово в тексте. Чем меньше перплексия - тем лучше способность. В тематических моделях предсказание следующих слов не является целью, тем не менее эта метрика позволяет судить о сходимости процесса обучения. Вычисление перплексии происходит по формуле:

$$\text{Perp}(\Phi, \Theta) = \exp \left(-\frac{1}{n} \sum_{d \in D} \sum_{w \in W} n_{dw} \log p(w | d) \right)$$

$$p(w | d) = \sum_{t \in T} \phi_{wt} \theta_{td}, \quad n \equiv \sum_{d \in D} \sum_{w \in W} n_{dw}.$$

Построение baseline

Информация о данных

В качестве исходных данных для обучения тематической модели был выбран открытый датасет с платформы kaggle, который представляет собой набор англоязычных статей взятых с ресурса Medium. Статьи затрагивают довольно широкий диапазон областей. Для данной работы взяты 10000 из доступных 190000 статей в целях более быстрого процесса обучения и обработки датасета.

Статьи представлены в достаточно хорошо обработанном виде, без аномалий, связанных с переносом строк, с малым количеством спецсимволов. Формат данных: csv-файл.

Предобработка текстов

Для того чтобы приступить к тематическому моделированию, необходимо провести предобработку текстов, которые могут содержать опечатки, слова, не значимые для тематического моделирования, спецсимволы, пунктуацию и т.д.

Для решения перечисленных проблем при составлении словаря производятся следующие операции:

1. Выполняется приведение всех слов к нижнему регистру.
2. Выделение слов из текстов при помощи регулярных выражений из библиотеки `re` языка `python`. Позволяет избавиться от спецсимволов и знаков препинания.
3. Удаление слишком часто встречающихся термов (более чем в 60% документов) и слишком редко (менее чем 5 раз).
4. Удаление стоп-слов при помощи библиотеки `nltk`. Вместе с предыдущим пунктом позволяет оставить только значимые для определения тематической структуры текстов термы.

5. Лемматизация слов по средствам уже упомянутой `nltk`. Позволяет корректно применять гипотезу мешка слов, так как предполагается, что форма слова не оказывает существенного влияния на его принадлежность к теме.

В результате описанной фильтрации и обработки в словарь вошло порядка 15000 уникальных термов (для исследования была выбрана униграмная модель). Для проверки корректности выполненных операций приведем распределение количества термов в зависимости от их частоты встречаемости на рис. 1.

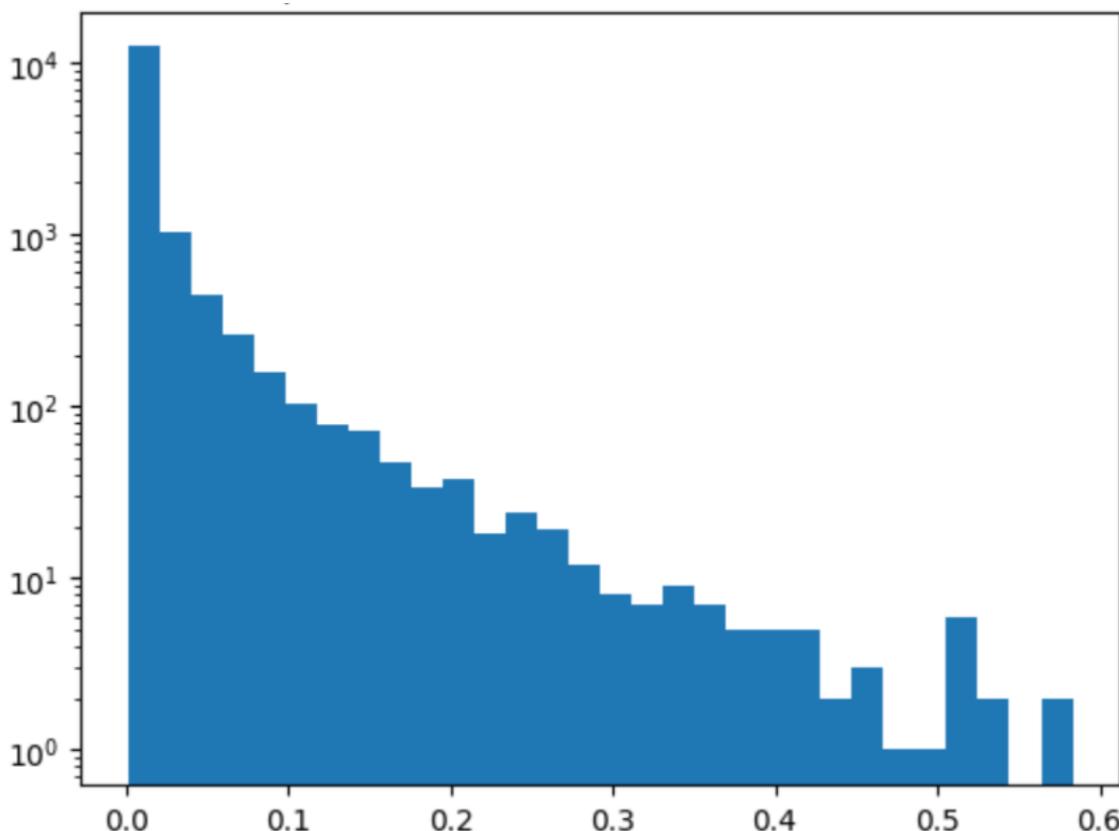


Рис. 1 Распределение относительных частот термов

Хорошо видно, что полученная зависимость удовлетворяет распределению Ципфа. Это можно считать подтверждением корректной работы с данными.

Описание базовой модели

В качестве простого решения была выбрана униграмная модель без регуляризаторов. Так как статьи довольно разнообразны, было принято решение остановиться на 50 темах для тематического моделирования. Результаты обучения после 24 эпох показаны на рис. 2 (большое количество эпох было выбрано с целью возможности сравнения с усложненными моделями).

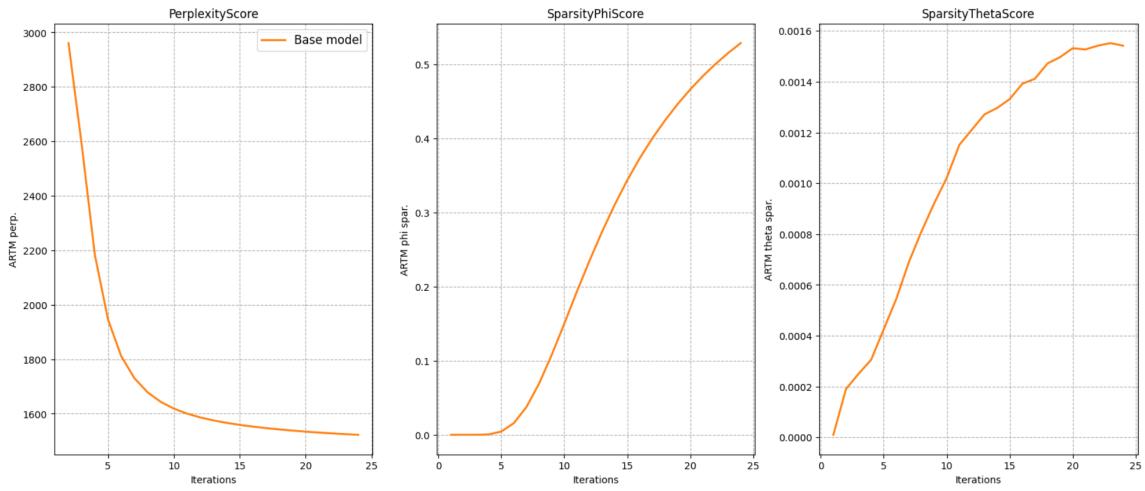


Рис. 2 Значения метрик для базовой модели

Также приведем несколько хорошо интерпретируемых тем из полученного списка:

- topic1: ['writing', 'write', 'story', 'writer', 'work', 'read', 'want', 'creative', 'people', 'first'] - написание, чтение книг
- topic27: ['mental', 'feel', 'health', 'brain', 'people', 'anxiety', 'body', 'person', 'help', 'stress'] - ментальное здоровье
- topic12: ['people', 'public', 'pandemic', 'health', 'government', 'social', 'country', 'many', 'virus', 'political'] - пандемия
- topic28: ['learning', 'machine', 'deep', 'based', 'image', 'neural', 'detection', 'network', 'algorithm', 'recognition'] - машинное обучение
- topic37: ['plot', 'import', 'axis', 'title', 'chart', 'date', 'matplotlib', 'column', 'figure', 'dataframe'] - библиотека matplotlib

Как мы видим, даже базовая модель дает неплохую интерпретируемость.

Определение методики применения регуляризаторов

Описание используемых регуляризаторов

Исходя из перечисленных в начале работы требований к матрицам распределения термов в темах и распределения тем в документах для проведения эксперимента были выбраны следующие регуляризаторы:

- Декоррелирование распределения термов в темах. Необходим для уменьшения схожести столбцов матрицы Φ .

$$R(\Phi) = -\tau \sum_{s,t \in T} \sum_{w \in W} \phi_{wt} \phi_{ws} \quad (1)$$

- Разреживание или сглаживание распределений термов в темах

$$R(\Phi) = \pm \alpha \sum_{w,t} \ln \phi_{wt} \quad (2)$$

- Разреживание или сглаживание распределений тем в документах

$$R(\Theta) = \pm \beta \sum_{t,d} \ln \theta_{td} \quad (3)$$

- Отбор тем

$$R(\Theta) = -\tau \sum_{t \in T} \ln \sum_{d \in D} p(d) \theta_{td} \quad (4)$$

Сравнение работы регуляризаторов при их включении в начале обучения модели

Для комфорtnого проведения был реализован интерфейс, позволяющий максимально быстро проводить тестирование регуляризаторов на различных грубых сетках и получать результаты в удобном формате. Тестирование проводилось при следующих условиях: создавалась новая модель с одним выбранным регуляризатором, проводилось обучение модели для нескольких значениях параметра и достаточном числе эпох, результаты выводились в виде графиков (при этом регуляризаторы разреживания и сглаживания считались различными).

В работе отражены наиболее важные результаты.

Декорелирование распределения термов в темах

Один из двух хорошо показавших себя регуляризаторов на начальном этапе обучения модели.

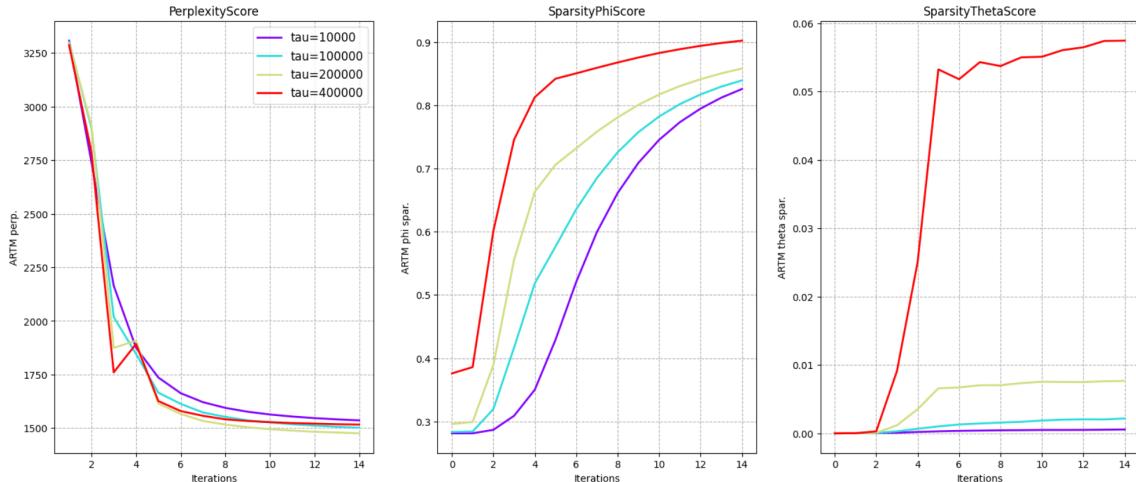


Рис. 3 Значения метрик для модели со включенным регуляризатором (1)

Видно улучшение всех трех метрик, хотя и не столь идеальное, как ожидалось. Можно пронаблюдать, что до определенного порога увеличение параметра ведет к улучшению всех трех метрик, после чего перплексия начинает ухудшаться. Тем не менее полученный результат можно объяснить. Так как из (1) регуляризатор для декорелирования распределения термов в темах представляет собой скалярное произведение строк матрицы Φ , то при слишком больших

значениях параметра регуляризатор теряет свой смысл и начинает обнулять элементы матрицы Φ (Для нулевой Φ значение регуляризатора действительно максимально и равно 0). Можно интерпретировать полученный результат так: перплексия помогает определить, при каких значениях параметра регуляризатор выполняет свою задачу и при каких - ухудшает качество модели.

Разреживание или сглаживание распределений термов в темах

В данном случае значения метрик также получилось увеличить одновременно.

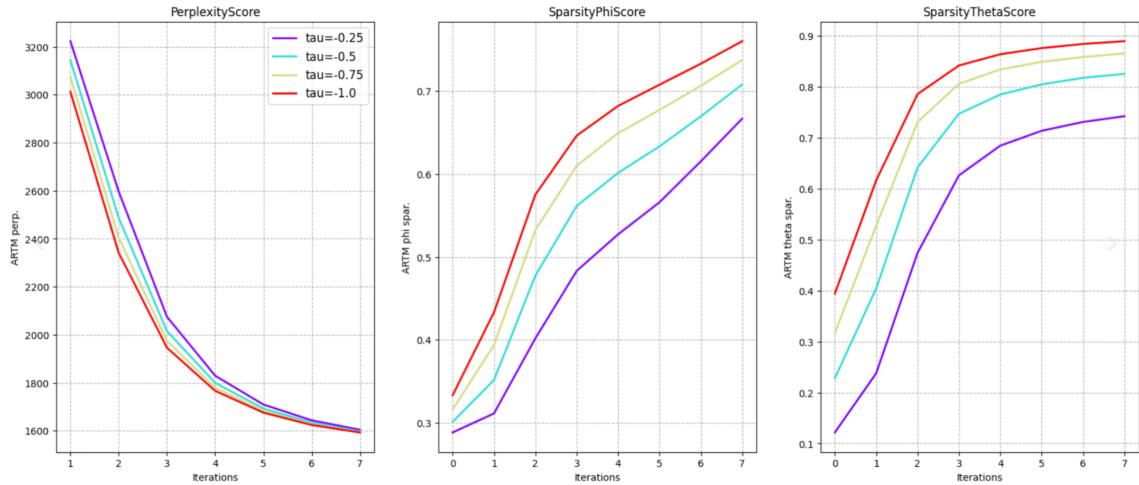


Рис. 4 Значения метрик для модели со включенным регуляризатором (3)

Наблюдается схожая с (1) проблема, но получение еще одного подходящего в начале обучения регуляризатора позволяет предположить, что последовательность, рекомендуемая в ходе курса не является единственной.

Неподходящие регуляризаторы

Приведем также динамику метрик (для дальнейшего сравнения), полученную при обучении с разреживанием распределений термов в темах и отбором тем.

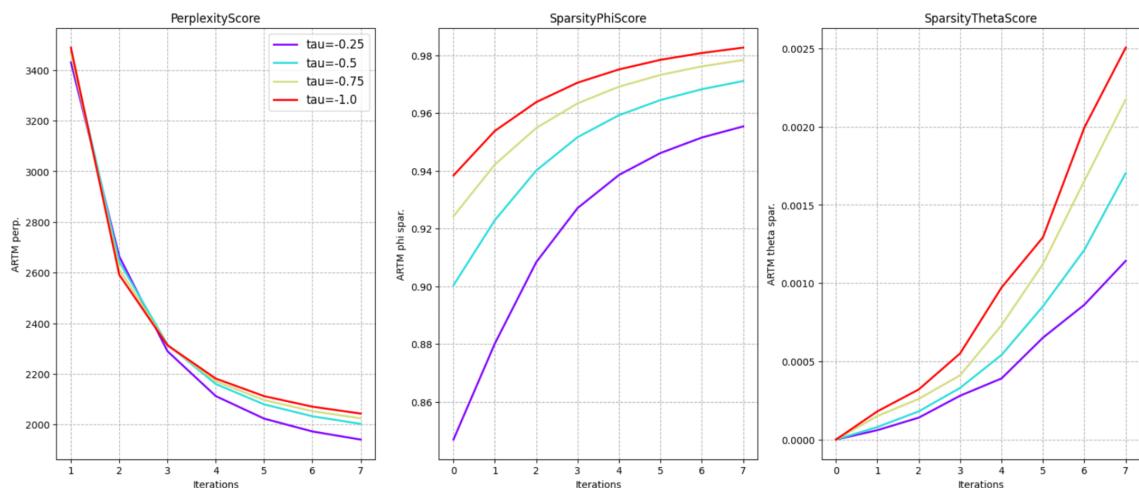


Рис. 5 Значения метрик для модели со включенным регуляризатором (2)

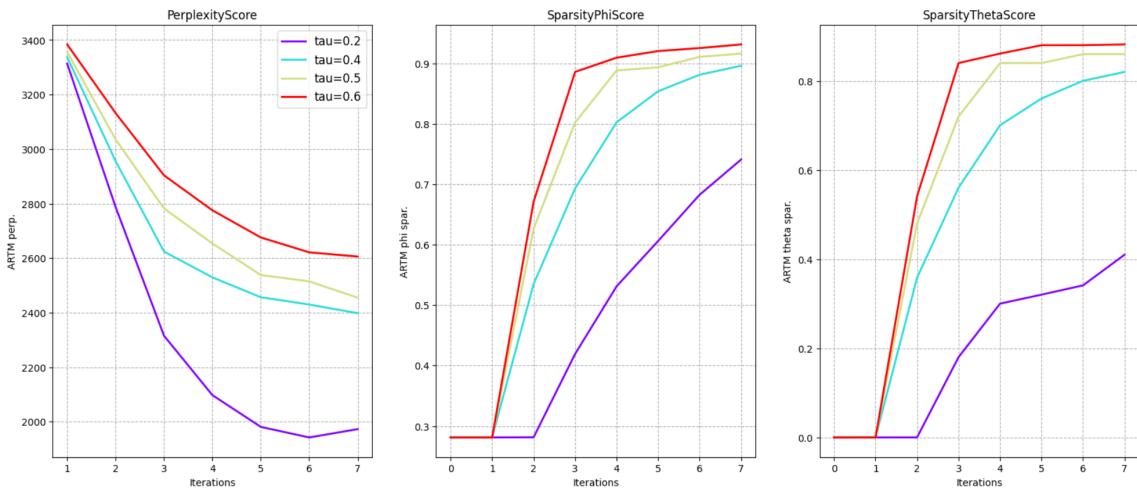


Рис. 6 Значения метрик для модели со включенным регуляризатором (4)

В обоих случаях четко видна корреляция уменьшения перплексии с одновременным уменьшением разреженности матриц Φ и Θ (а также наоборот). Значит, данные регуляризаторы нежелательно применять в начале обучения модели.

Поиск изменений в работе регуляризаторов при их последовательном включении

По результатам экспериментов первым был выбран регуляризатор для декоррелирования с параметром $\tau = 2e5$. А так же оптимальное количество эпох - от 6 до 8, когда матрица Φ еще не слишком разрежена.

Второй регуляризатор

Лучшим вторым по счету регуляризатором действительно оказался (3), причем необходимое для получения результата количество эпох оказалось равным 8-ми.

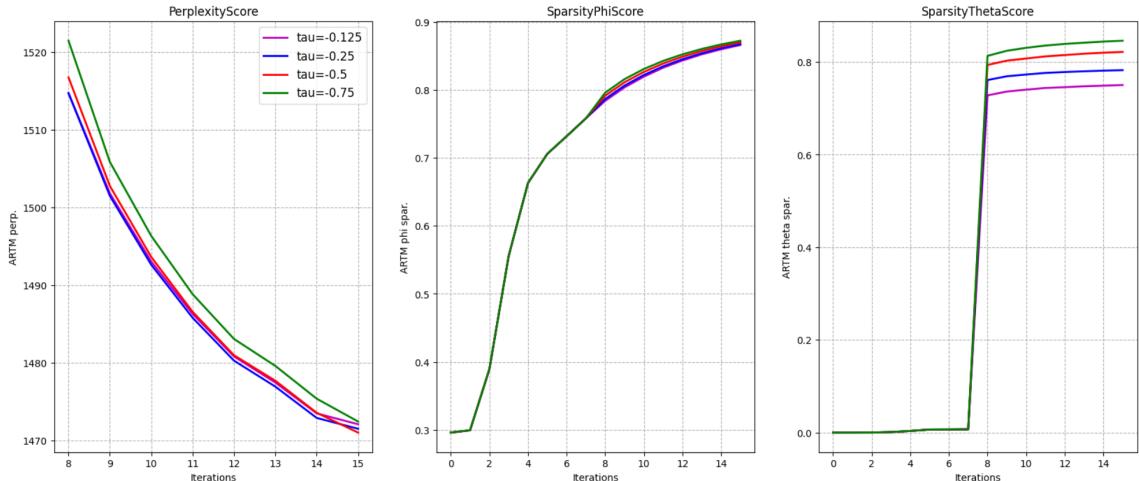


Рис. 7 Значения метрик для модели со включенным регуляризатором (3) при включении его вторым.

Вновь оказалось, что существует порок для параметра, до достижения которого увеличиваются все три метрики, а после - только разреженность Θ . Что является логичным следствием из устройства (3) и дает нам некий индикатор того, когда регуляризатор работает на пользу тематическому моделированию.

Важным оказалось исследование двух вопросов:

1. Выключать ли декоррелятор распределений термов в темах?
2. Имел ли смысл последовательное включение?

Ответ на первый вопрос неоднозначен - результаты приблизительно равны и показаны на рис. 8.

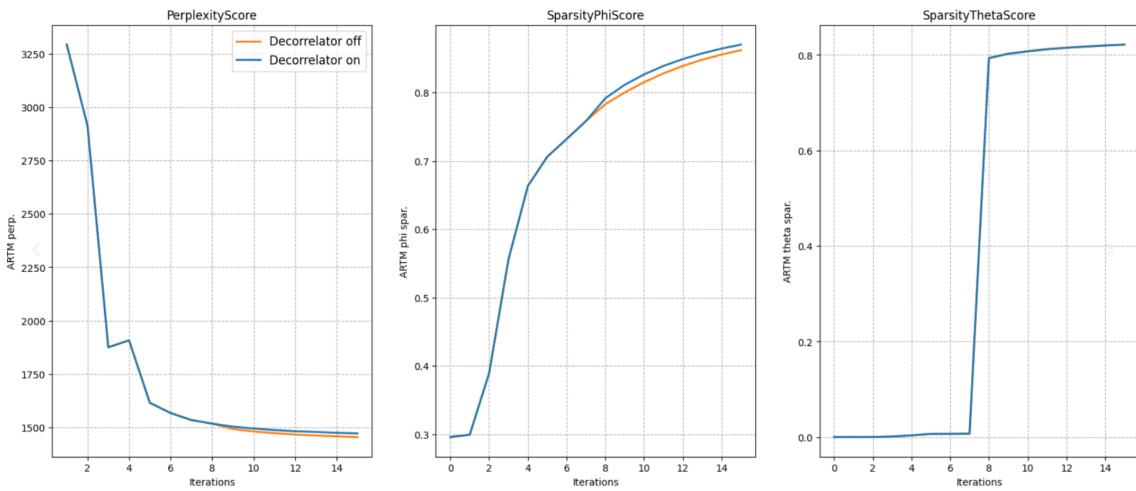


Рис. 8 Значения метрик при отключении декоррелятора и без него

Можно интерпретировать их следующим образом: так как порядок регуляризаторов подобран правильно, то декоррелятор уже не оказывает значительного влияния на качество модели.

Второй вопрос имеет более явное решение.

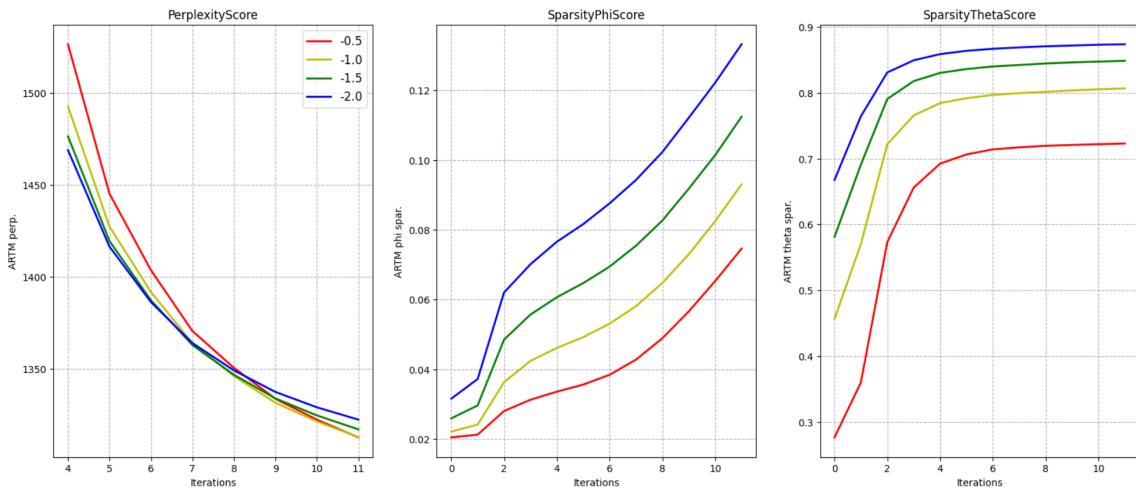


Рис. 9 Значения метрик при включении всех регуляризаторов в начале обучения

Даже без прямого сравнения видно, что добиться улучшения всех трех метрик не получается. Остается та же тенденция, что и при включении только регуляризатора для разреживания распределения тем в документах - улучшение разреженности за счет уменьшения перплексии.

Итак, наиболее эффективным вторым по счету регуляризатором оказался (3) с параметром $\tau = -0.5$ и количеством эпох 8.

Третий регуляризатор

В данном пункте не получилось добиться результата, показанных на спецсеминаре. Приведем значения метрик при включении сглаживания матрицы Φ .

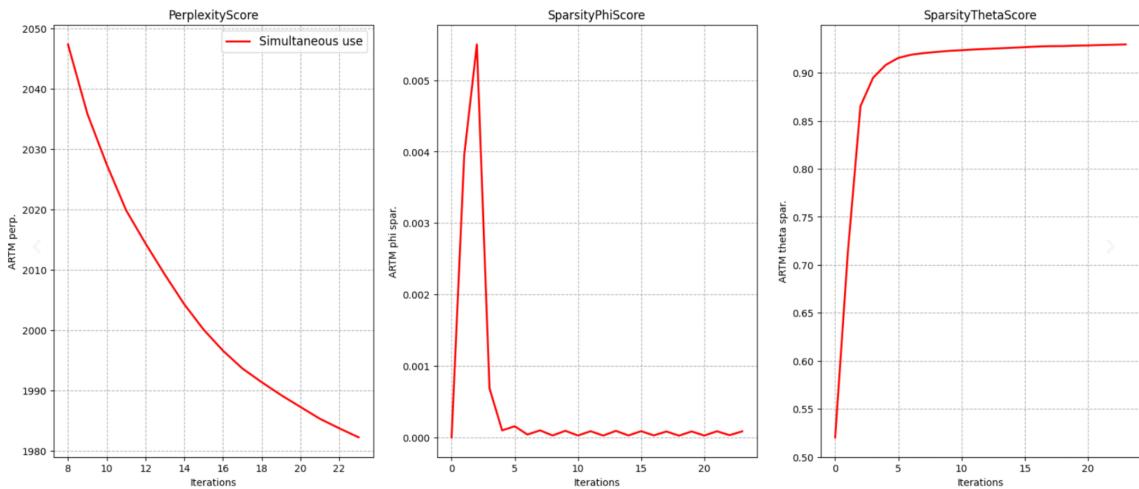


Рис. 9 Значения метрик при включении (2) с положительным параметром.

При обучении с малым значением параметра наблюдались сильные скачки перплексии, а при обучении с большим - критическое снижение матрицы Φ .

Более интересный результат был получен для регуляризатора отбора тем (перед его включением был отключен декоррелятор).

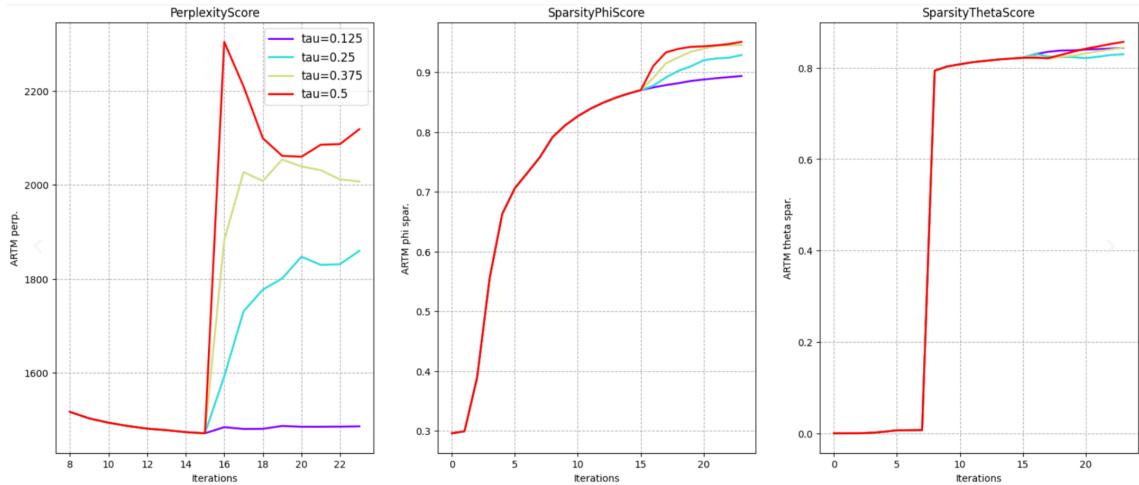


Рис. 10 Значения метрик при включении (4) третьим по счету

Можно видеть, что его поведение отличается от такового при включении в начале обучения. За счет увеличения перплексии происходит "корректировка" разреженности Φ и Θ .

Третьим регуляризатором будет выбран (4) с параметром $\tau = 0.25$.

Сравнение построенной методики с более простыми методами

Сравним построенную последовательность с двумя базовыми моделями: без регуляризаторов и со включением их одновременно.

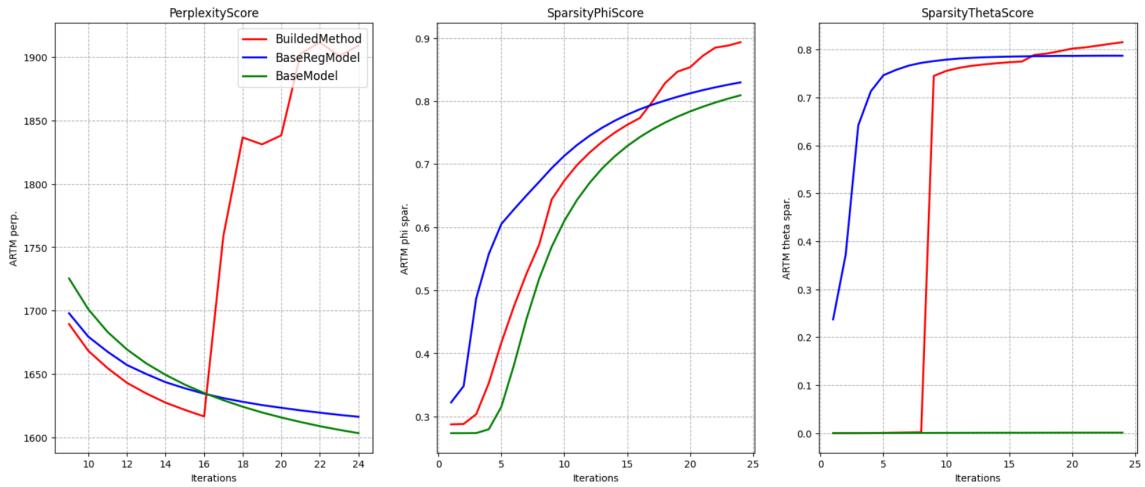


Рис. 11 Значения метрик при включении регуляризатора отбора тем

В данном случае наблюдается проигрыш в перплексии, так как регуляризатор отбора тем направлен на улучшение интерпретируемости тем и на уменьшение их количества. Тем не менее видно, что построенный в ходе исследования метод работает лучше, чем одновременное включение тех же регуляризаторов.

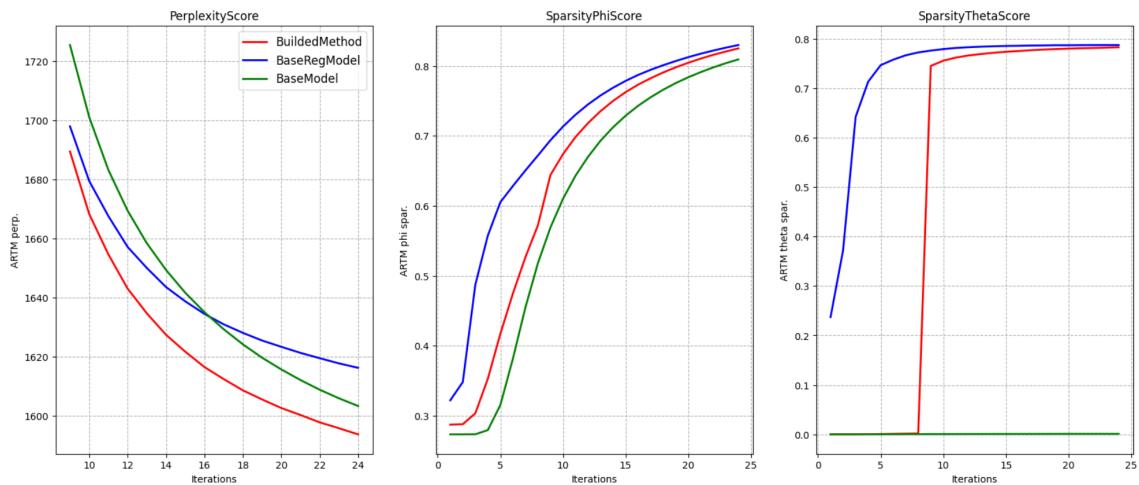


Рис. 12 Значения метрик без включения регуляризатора отбора тем

При отключении регуляризатора отбора тем мы видим преимущество полученного метода во всех метриках, но небольшое отставание от модели, обученной при одновременном включении регуляризаторов, в разреженности распределений слов в темах и тем в документах. Что можно интерпретировать как увеличение этих характеристик во вред качеству модели.

Конечно, наиболее сильно результат заметен в интерпретируемости тем. После применения регуляризаторов модель стала способна выявлять куда более узкие тематические группы, например:

topic9: ['energy', 'water', 'sleep', 'carbon', 'temperature', 'wind', 'solar', 'electricity', 'power', 'renewable', 'environmental', 'fuel'] - альтернативные способы добычи энергии

topic13: ['course', 'learn', 'yoga', 'practice', 'logo', 'free', 'fitness', 'skill', 'passive', 'body', 'comfort', 'class'] - фитнес

topic32: ['person', 'self', 'feel', 'emotional', 'relationship', 'others', 'love', 'care', 'someone', 'feeling', 'partner', 'negative'] - отношения

topic37: ['install', 'notebook', 'python', 'conda', 'installed', 'jupyter', 'environment', 'command', 'cell', 'package', 'select', 'shell'] - jupyter notebook

Выводы

В результате исследования получилось построить последовательность регуляризаторов и подобрать набор параметров, позволивших улучшить одновременно несколько метрик. Но результат все же отличался от показанного на лекции, так как не существовало оптимума для параметра по всем метрикам одновременно.

Полученный в данной работе результат подтверждает гипотезу о том, что регуляризаторы влияют положительно на общее качество модели, а не на ее характеристики в отдельности. Более того, наблюдая сразу несколько метрик можно эффективно подбирать оптимальные параметры для построения модели.

Подводя итог, стоит выдвинуть предположения, почему не был достигнут идеальный результат и определить направления для дальнейшего исследования:

- В процессе выполнения работы малое внимание было уделено подбору количества тем, так как в совокупности с подбором параметров это заняло бы слишком много времени. Тем не менее, возможно, существует оптимальное число тем, при котором реально получить улучшение всех трех метрик.
- Для проведения эксперимента была выбрана униграмная модель. Более сложные версии помогают улучшить качество тематического моделирования и, возможно, для них результат измениться.

Литература

Список литературы

- [1] Воронцов К. В. Вероятностное тематическое моделирование
<http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf>
- [2] Воронцов К. В. Аддитивная Регуляризация Тематических Моделей Коллекций Текстовых Документов. - Доклады РАН, Т.455, №3. С.268-271, 2014