Отчет по серии экспериментов с онлайновым алгоритмом (06.08.13)

Потапенко Анна Александровна
МГУ имени М.В. Ломоносова, факультет ВМК, кафедра ММП, студент 4 курса

1. Выбор оптимального режима онлайнового алгоритма

Цель экспериментов. Исследовать зависимость достигаемого качества модели и времени обучения от значений параметров. Найти оптимально сочетание параметров, предоставляющее некоторый компромисс между затраченным временем и полученным качеством (хотим чудес – быстро и хорошо).

Условия проведения экспериментов. Эта серия экспериментов проводится с онлайновым PLSA (т.е. формулы без дигамм), без регуляризации Дирихле ($\alpha = 0$, $\beta = 0$). Настраиваются параметры:

— число итераций по документу
— число итераций по пачке документов
— размер пачки
— параметр забывания $\varkappa$

При настройке одного из параметров, все остальные фиксированы и принимают значения 10, 1, 50, 0.6 соответственно.

Эксперименты проводятся на 3 коллекциях: SmallLabelledCollection, NIPS, RuDis. Для всех коллекций вычисляется перплексия, для SmallLabelledCollection, кроме того, качество категоризации (F-мера).

1.1. Выбор числа итераций по документу

Переберем различные значения параметра, а также попробуем постепенно увеличивать число итераций: 1 раз пройдем всю коллекцию с числом итераций по документу 3, затем с 5, затем с 10, затем с 50.

Результаты на коллекции SLC:

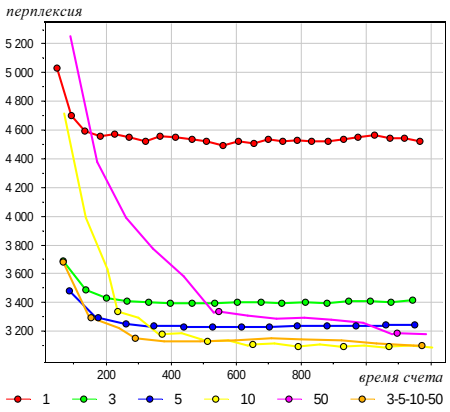

Рис. 1. При увеличении числа итераций алгоритм сходится за большее время, но к более высокому качеству. Компромисс – 5-10 итераций. Хорошо работает стратегия постепенного увеличения числа итераций.

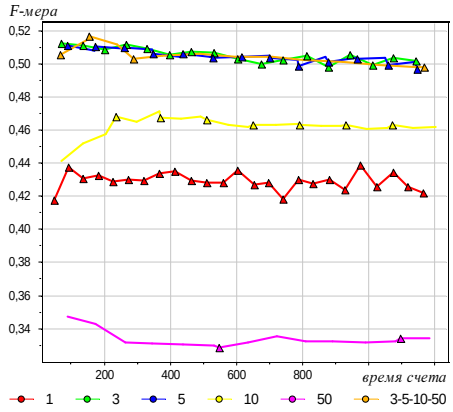

Рис. 2. Высокое качество при числе итераций 5-10, а также при стратегии постепенного увеличения числа итераций.

На других коллекциях результаты аналогичны.

1.2. Выбор числа итераций по пачке документов

Результаты на коллекции SLC:

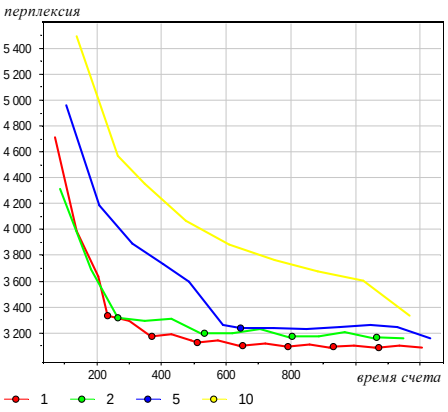

Рис. 3. Оптимальное число итераций – 1.

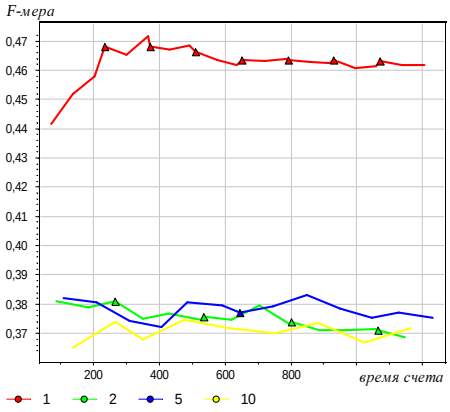

Рис. 4. Оптимальное число итераций – 1.

На других коллекциях результаты аналогичны.

1.3. Выбор размера пачки

Результаты на коллекции SLC (1000 документов в обучении):

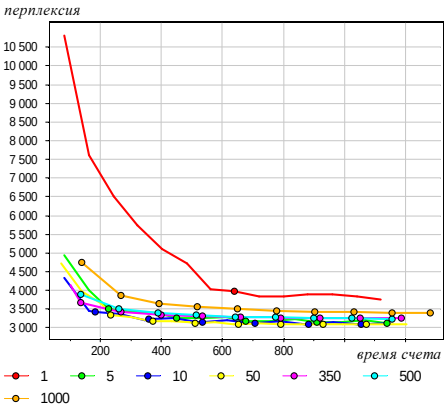

Рис. 5. Оптимум при размере пачки 10-50 документов.

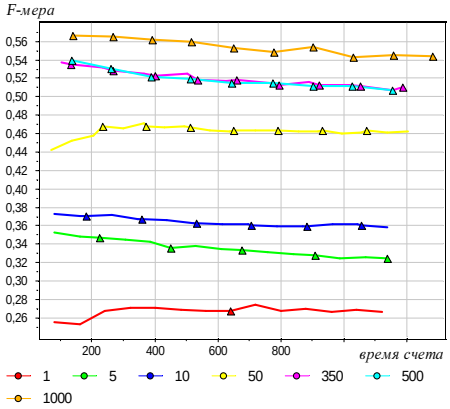

Рис. 6. Постоянное улучшение при увеличении пачки.

Результаты на коллекциях RuDis (2000 документов в обучении) и NIPS (1500 документов в обучении):

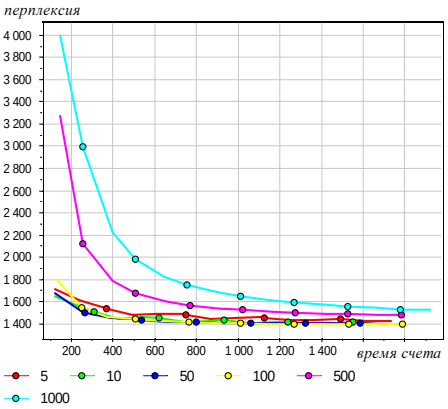

Рис. 7. RuDis: Оптимум при размере пачки 50-100 документов.

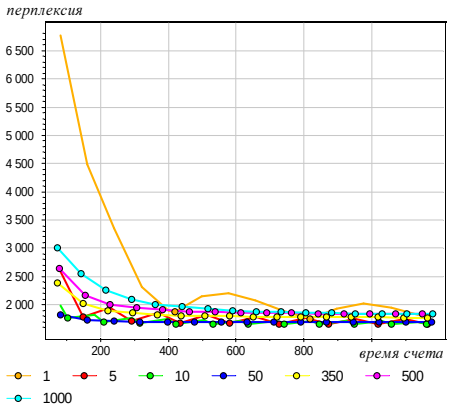

Рис. 8. NIPS: Оптимум при размере пачки 10 документов.

1.4. Выбор параметра забывания $\varkappa$

Для нескольких размеров пачек переберем различные значения параметра $\varkappa$ в формуле экспоненциального сглаживания для расчета λ после просмотра очередной пачки, а также формулу арифметического усреднения:

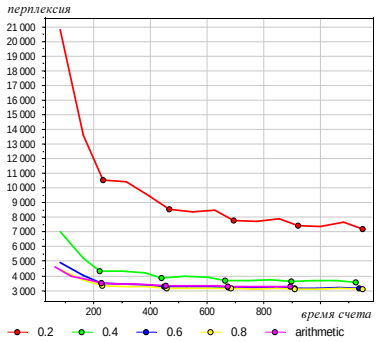

Рис. 9. Размер пачки 5: Оптимум при $\varkappa = 0.8$.

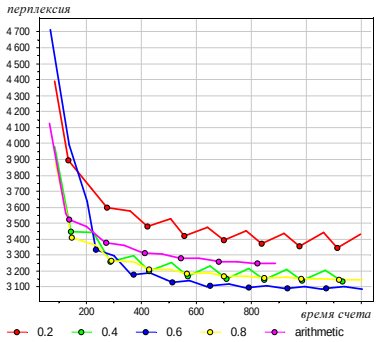

Рис. 10. Размер пачки 50: Оптимум при $\varkappa = 0.6$.

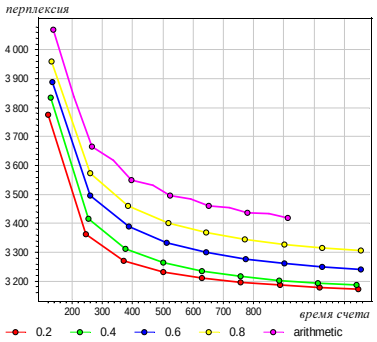

Рис. 11. Размер пачки 500: Оптимум при $\varkappa = 0.2$.

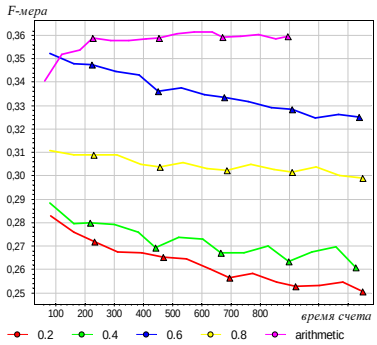

Рис. 12. Размер пачки 5: Арифметическое сглаживание лучше экспоненциального.

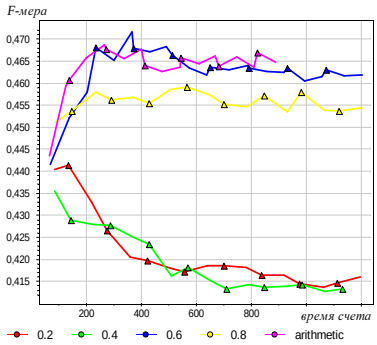

Рис. 13. Размер пачки 50: Арифметическое сглаживание лучше экспоненциального.

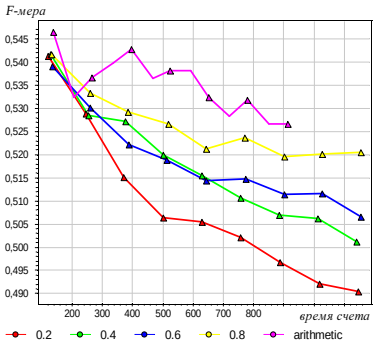

Рис. 14. Размер пачки 500: Арифметическое сглаживание лучше экспоненциального.

Теперь вернемся к вопросу выбора оптимального размера пачки, но при каждом размере будем использовать свое оптимальное $\varkappa$:

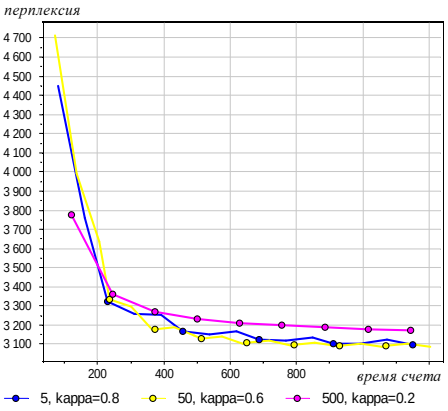

Рис. 15. Оптимум по-прежнему при размере пачки около 50 документов.

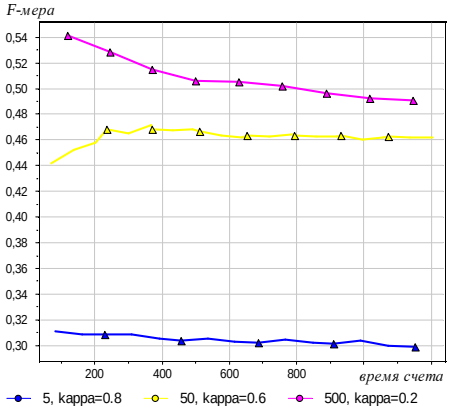

Рис. 16. По-прежнему постоянное улучшение при увеличении пачки.

Эти графики получены на SLC, на других коллекциях результаты аналогичны.

Выводы по экспериментам. По результатам работы алгоритма на рассматриваемых коллекциях, рекомендуется выбирать следующие параметры: число итераций по одному документу – около 10, число итераций по пачке – 1, размер пачки – около 50, параметр $\varkappa = 0.6$.

2. Формулы PLSA, регуляризованного PLSA и LDA-VB

Цель экспериментов. Сравнить и выбрать формулы, дающее лучшее качество.

Условия проведения экспериментов. Параметры работы онлайнового алгоритма фиксированы в соответствии с рекомендациями из предыдущего пункта. Рассматриваются коллекции данных SLC, RuDis и NIPS.

Результаты на SLC:

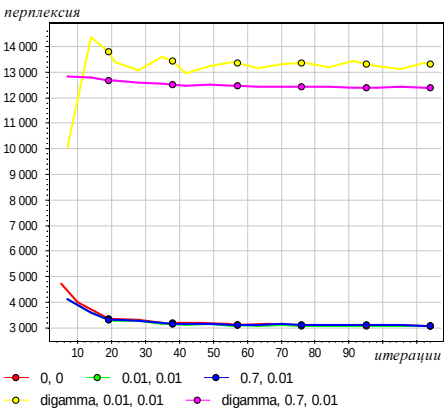

Рис. 17. Формулы LDA-VB приводят к плохому качеству, более того, процесс не сходится по итерациям. Наличие регуляризации сказывается несильно, значения гиперпараметров $\alpha = 0.01$, $\beta = 0.01$ лучшие из рассмотренных.

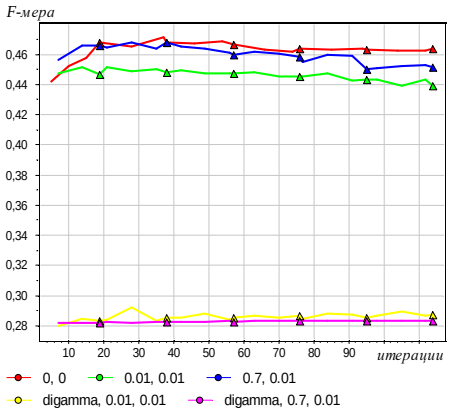

Рис. 18. Формулы LDA-VB приводят к плохому качеству, более того, процесс не сходится по итерациям. Наличие регуляризации сказывается несильно, значения гиперпараметров $\alpha = 0$, $\beta = 0$ лучшие из рассмотренных.

Результаты на RuDis и NIPS:

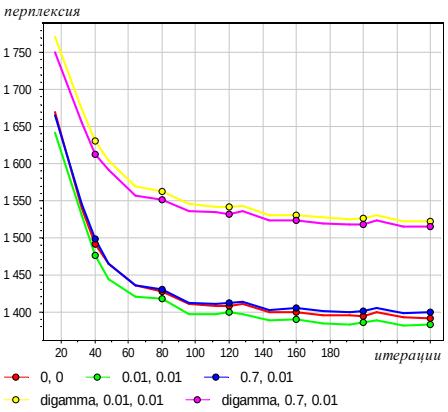

Рис. 19. RuDis: Формулы LDA-VB приводят к плохому качеству, процесс по итерациям сходится. Наличие регуляризации сказывается несильно, значения гиперпараметров $\alpha = 0.01$, $\beta = 0.01$ лучшие из рассмотренных.

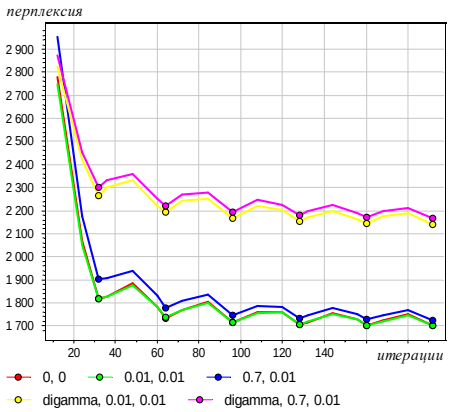

Рис. 20. NIPS: Формулы LDA-VB приводят к плохому качеству, процесс по итерациям сходится. Наличие регуляризации сказывается несильно, значения гиперпараметров $\alpha = 0.01$, $\beta = 0.01$ лучшие из рассмотренных.

Выводы по экспериментам. Формулы LDA-VB либо неправильно реализованы, либо существенно проигрывают формулам PLSA по перплексии. Регуляризация не играет существенной роли.

3. Постепенное разреживание в онлайновом алгоритме

Цель экспериментов. Подобрать стратегию и параметры разреживания.

Условия проведения экспериментов. Параметры работы онлайнового алгоритма фиксированы в соответствии с рекомендациями из предыдущего пункта. Рассматриваются две коллекции данных – SLC и NIPS. Разреживание производится с помощью постепенного обнуления наименьших значений в профилях тем и профилях документов. Обнуление профилей тем включается после обработки A пачек, обнуляются наименьшие значения, дающие в сумме B, при этом их число не превосходит долю C от размерности профиля. Обнуление профилей тем включается после обработки A пачек, начиная с D-ой итерации по документу, обнуляются наименьшие значения, дающие в сумме E, при этом их число не превосходит долю F от размерности профиля. Кроме того, обнуление не производится, если превышена верхняя граница разреженности: для профилей тем это некая константа (0.97), для профилей документов – текущая средняя разреженность профилей тем. Параметры A-B-C-D-E-F настраиваются в ходе серии экспериментов.

Результаты на SLC:

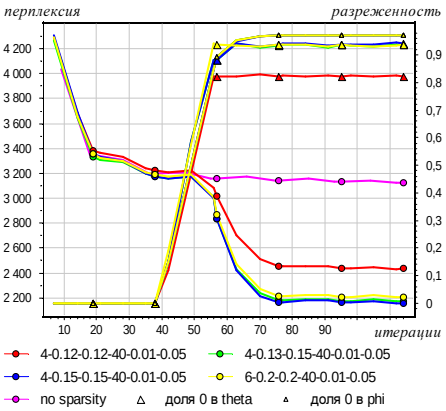

Рис. 21. Разреженность профилей более 95 процентов, перплексия при этом улучшается.

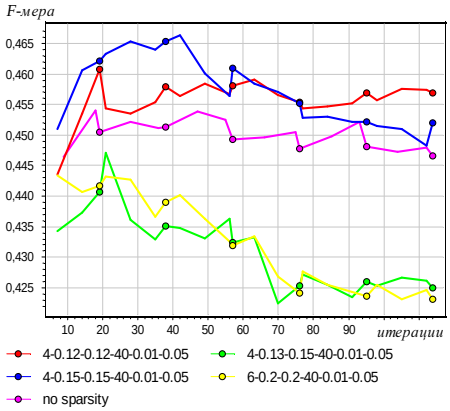

Рис. 22. При некоторых параметрах разреживание улучшает F-меру.

Проанализируем разреженность более подробно для одного из набора параметров (4-0.15-0.15-40-0.01-0.05).

Гистограммы разреженности матриц Φ и Θ:

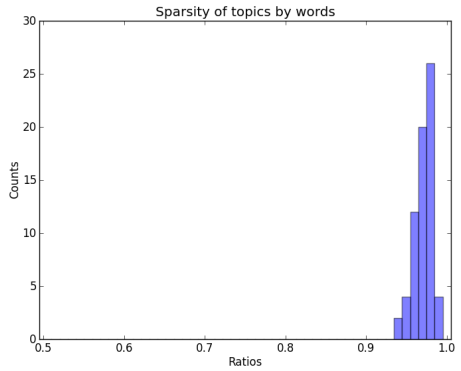

Рис. 23. Число тем с различной разреженностью соответствующих столбцов матрицы Φ.

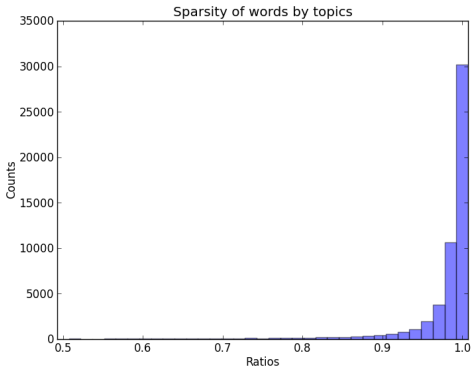

Рис. 24. Число слов с различной разреженностью соответствующих строк матрицы Φ.

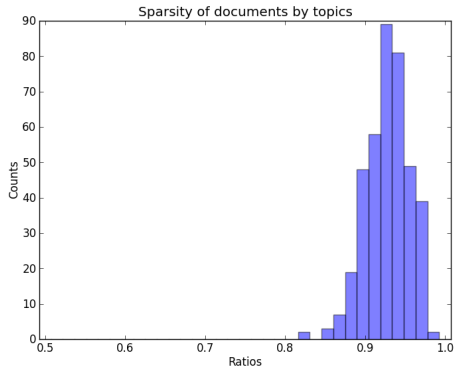

Рис. 25. Число документов с различной разреженностью соответствующих столбцов матрицы Θ.

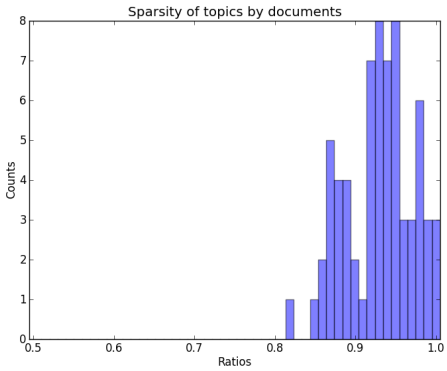

Рис. 26. Число тем с различной разреженностью соответствующих строк матрицы Θ.

Гистограммы долей шума в документах:

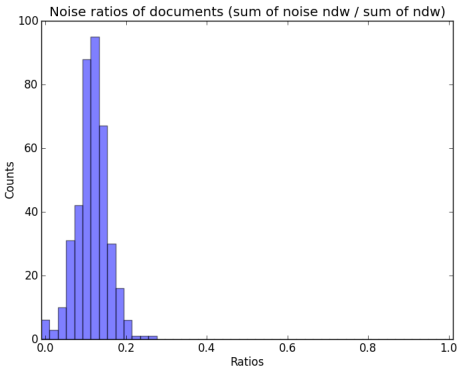

Рис. 27. Число документов с различной долей шума. Шум считается как (сумма шумовых n_{dw}) / (длина документа).

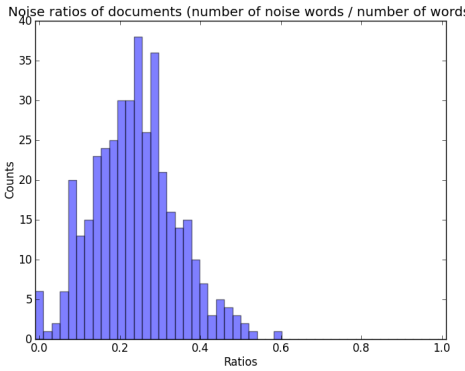

Рис. 28. Число документов с различной долей шума. Шум считается как (число шумовых слов) / (число различных слов в документе).

При таком сильном разреживании очень много слов уходит в шум, т.е. $p(t|d,w) = 0, \forall t \in T$. Более половины слов словаря имеют нулевую вероятность в любой теме, значит, каждое вхождение таких слов – шум. Если посмотреть на эти слова глазами, кажется, что это просто редко встречаемые слова, которые вполне могут быть как шумовыми, так и тематическими.

Рассмотрим более «мягкие» параметры разреживания. Гистограммы при этом существенно меняются (представлены для набора параметров 6-0.1-0.1-40-0.01-0.01).

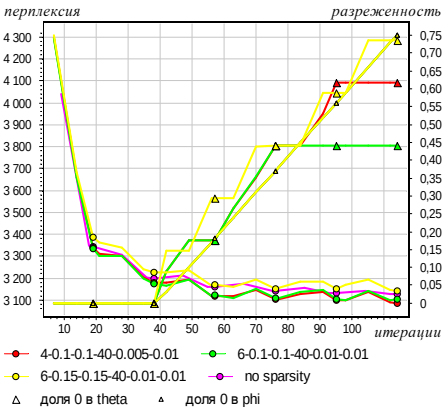

Рис. 29. Разреженность профилей около 60 процентов, перплексия отличается от исходной не сильно.

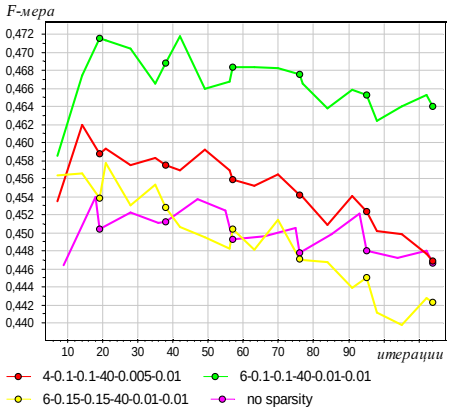

Рис. 30. При некоторых параметрах разреживание улучшает F-меру.

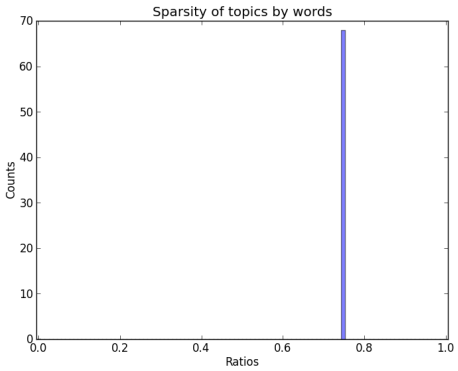

Рис. 31. Число тем с различной разреженностью соответствующих столбцов матрицы Φ.

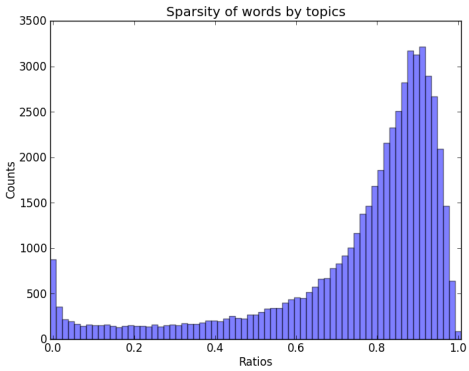

Рис. 32. Число слов с различной разреженностью соответствующих строк матрицы Φ.

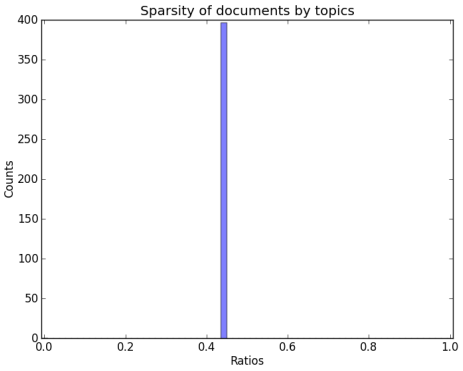

Рис. 33. Число документов с различной разреженностью соответствующих столбцов матрицы Θ.

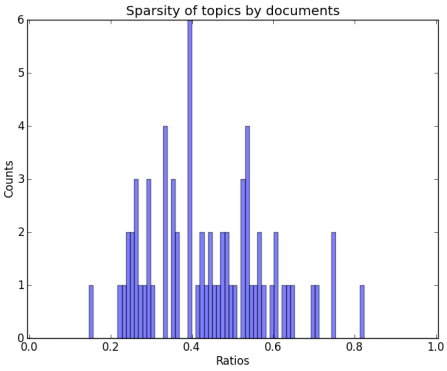

Рис. 34. Число тем с различной разреженностью соответствующих строк матрицы Θ.

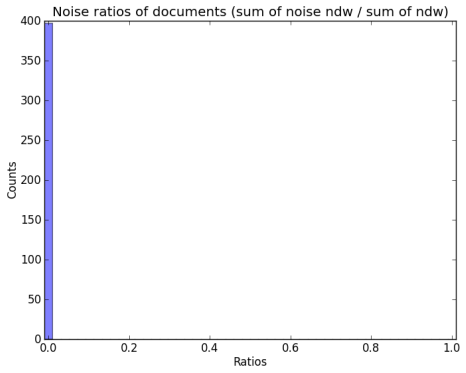

Рис. 35. Число документов с различной долей шума. Шум считается как (сумма шумовых n_{dw}) / (длина документа).

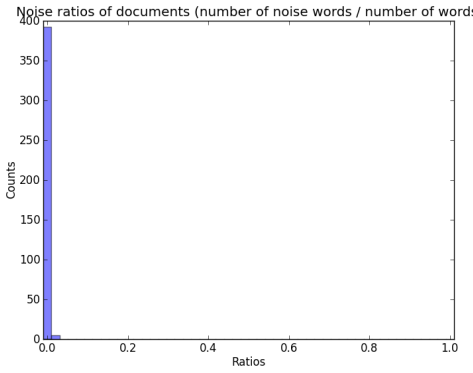

Рис. 36. Число документов с различной долей шума. Шум считается как (число шумовых слов) / (число различных слов в документе).

Выводы по экспериментам.

Постепенное обнуление элементов в профилях тем и документов может привести к любой доле разреженности, при этом перплексия улучшится. Для более грамотных экспериментов нужен критерий останова разреживания, а также подсчет большого числа метрик качества модели помимо перплексии.