Инструкция по проведению эксперимента с регуляризацией тематических моделей в библиотеке BigARTM с использованием скрипта experiment02\_artm.

Скрипт предоставляет возможность построения тематической модели с тремя регуляризаторами (разреживание \Phi, разреживание \Theta и декорреляция тем в \Phi). Также он позволяет контролировать процесс построения модели с помощью функционалов качества, таких как перплексия на отложенной выборке, разреженности матриц \Phi и \Theta, средние характеристики ядер тем.

[Следует учесть, что в целях сокращения влияния времени пересчёта перплексии на работу алгоритма, этот функционал считается один раз в 20 итераций. Это число можно изменить в перменной ‘test\_every’.] **--- […] сноска или выделено отдельно**

Предполагается, что у вас уже установлена BigARTM в директории ‘your\_home\_folder’. Последовательность действий следующая:

1. Скачать коллекцию в виде батчей BigARTM. Наш эксперимент производился на коллекции английской Википедии, которая доступна по ссылкам
   1. <https://s3-eu-west-1.amazonaws.com/artm/enwiki-20141208_1k.7z>
   2. <https://s3-eu-west-1.amazonaws.com/artm/enwiki-20141208_10k.7z>

В первом варианте размер батча равен 1000, во втором — 10000 документов. Мы использовали 10000. Разархивированную папку с коллекцией следует поместить в ‘your\_home\_folder’. Туда же из папки с батчами следует переложить файл словаря ‘dictionary’.

1. Батч, который вы хотите использовать для подсчёта тестовой перплексии, нужно положить в ‘your\_home\_folder’. В нашем эксперименте использовался батч ‘243af5b8-beab-4332-bb42-61892df5b044.batch’
2. Подготовка скрипта. Откройте текст скрипта и найдите объявление (-я) переменной (-ых)
   1. ‘home\_folder’ (8 строка) и присвойте ей адрес ‘your\_home\_folder’;
   2. ‘batch\_size’ (28 строка) и присвойте ей выбранный размер батча;
   3. ‘batches\_disk\_path’ (36 строка) и замените строку 'wiki\_10k' на название вашей директории с батчами;
   4. ‘test\_batch\_name’ (43 строка) и заменить строку с непосредственным именем батча на имя выбранного вами тестового батча;
   5. ‘tau\_decor’, ‘tau\_phi’ и ‘tau\_theta’ (57-59 строки) и подставьте те коэффициенты регуляризации, которые считаете нужными.
3. В том случае, если вы хотите отдельно посчитать финальную перплексию на другой отложенной выборке произвольного размера, разместите выбранные тестовые батчи в папке ‘test’ (в директории ‘your\_home\_folder’), затем найдите в тексте скрипта объявление ‘save\_and\_test\_model’ (строка 30) и присвойте ей ‘True’.
4. Запустите скрипт. Текущие значения функционалов будут печататься в консоль, при этом, если на данной синхронизации не считалась контрольная перплексия, вместо неё будет выводиться строка ‘NO’. Результаты синхронизаций, на которых перплексия подсчитывалась, помимо консоли будут записываться в соответствующие файлы в директории ‘results’. Формат всех файлов одинаков — строки вида “(‘кумулятивное число обработанных документов’, ‘значение функционала’)”.

При желании вы можете менять значения любых переменных в тексте скрипта, смысл каждой из них ясен из комментариев. Если вы использовали все параметры и данные, соответствующие нашему эксперименту, то должны получить результаты, близкие к этим

[Компьютер, на котором мы проводили эксперименты — **amazon machine**]

Таблица (**можно взять из файла experiments.tex**)

Графики (**приложены .pdf**)

Здесь показаны сравнительные результаты ARTM и LDA. Чтобы провести эксперимент с LDA вместо ARTM, достаточно в тексте скрипта значения переменных ‘tau\_decor’, ‘tau\_phi’ и ‘tau\_theta’ заменить на 0, 1/topics\_count и 1/topics\_count соответственно и повторить эксперимент (не забудьте сохранить куда-нибудь файлы из папки ‘results’, они будут перезаписаны!).

*Такие графики с вашими результатами можно быстро построить, используя файлы ‘plot\_perplexity\_sparsity.tex’, ‘plot\_kernel.tex’ и результирующие файлы. Откройте tex-файлы и вставляйте целиком содержимое файлов из директории ‘results’ в те места, которые помечены соответствующими комментариями (следите, чтобы после вставки не оставалась пустых строк!). Затем скомпилируйте tex-файлы.*

*[ВАЖНО: Построение графиков описано для нашего эксперимента! В любом другом эксперименте корректность отображения графиков не гарантируется, но вы можете построить их самостоятельно, используя файлы с результатами, либо отредактировав необходимые переменные в текстах ‘plot\_perplexity\_sparsity.tex’ и ‘plot\_kernel.tex’]* --- **я могу это описать подробно, только имеет ли смысл?**

**Другой вариант вместо выделенного курсивом**

Для построения графиков следует использовать файлы со значениями функционалов из папки ‘results’. Графики нашего эксперимента можно получить компиляцией файлов ‘plot\_perplexity\_sparsity.tex’ и ‘plot\_kernel.tex’.