Выявление этничных тем в LiveJournal

Мурат Апишев
great-mel@yandex.ru
MelLain@github.com

МГУ им М.В. Ломоносова

6 ноября 2015

Коллекция и задача

Параметры коллекции:

- 1.58 млн. документов в виде «мешка слов»;
- 860 тыс. слов словаре;
- коллекция прошла лемматизацию.

Особенности:

- много слов с ошибками;
- коллекция русскоязычная, но присутствую термины на английском, украинском;
- много жаргонных слов и терминов специфических областей — сложно понимать и интерпретировать темы!

Парсинг и предобработка

Парсим в формат Vowpal Wabbit, подходящий для BigARTM.

Сохраним только те слова, которые:

1. содержат только символы киррилицы и дефис;
2. содержат не более одного дефиса (есть слова вроде --, ----);
3. имеют длину не менее 3-х символов (есть слова вроде 'а', 'ж');
4. встречаются в коллекции не менее 20 раз;

Хотелось бы отрезать по большему порогу, но есть риск выкинуть этнонимы.

Объём итогового словаря: 90 тыс слов.

Составление словаря этнонимов

Описание проблемы:

- Имеется словарь из нескольких сотен этнонимов.
- Слова собраны в списки (например [абхаз, абхазец, абхазка])
- Пересечение слов этого словаря со словарём LJ непустое, но многие слова теряются.
- Нужно составить аналогичный словарь, специфичный для LJ.

Можно сделать вручную:

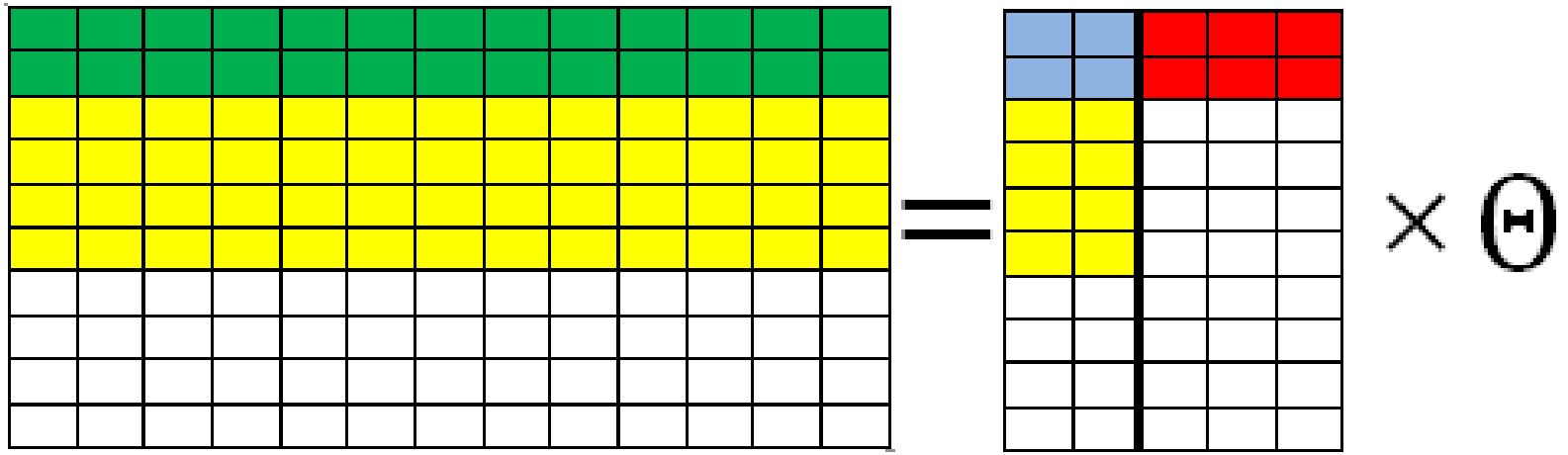
1. преобразовать списки всех слов в один линейный список;
2. пройтись по этому списку и для каждого слова найти все максимально похожие на него;
3. выбрать вручную в получившемся множестве все наиболее этничные слова, по 1-2 на каждый этноним исходного списка.

Объём итогового словаря этнонимов: 250 слов.

Примеры этнонимов

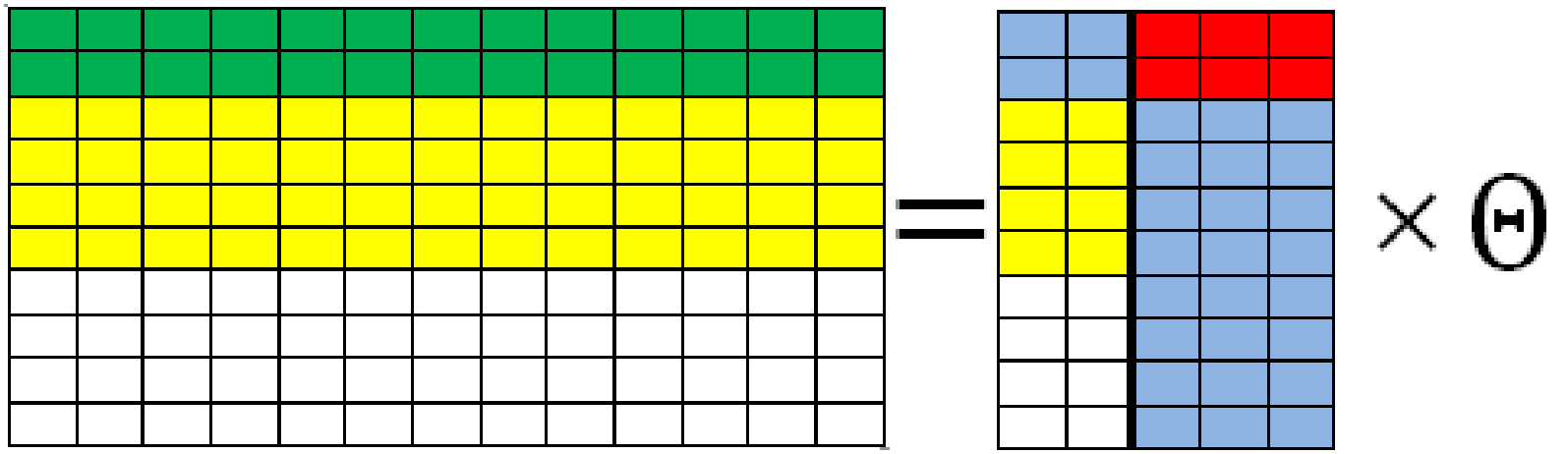
османский	русич
восточноевропейский	сингапурец
эвенк	перуанский
швейцарская	словенский
аланский	вепсский
саамский	ниггер
латыш	адыги
литовец	сомалиец
цыганка	абхаз
ханты-мансийский	темнокожий
карачаевский	нигериец
кубинка	лягушатник
гагаузский	камбоджиец

Сглаживание/разреживание этнонимов

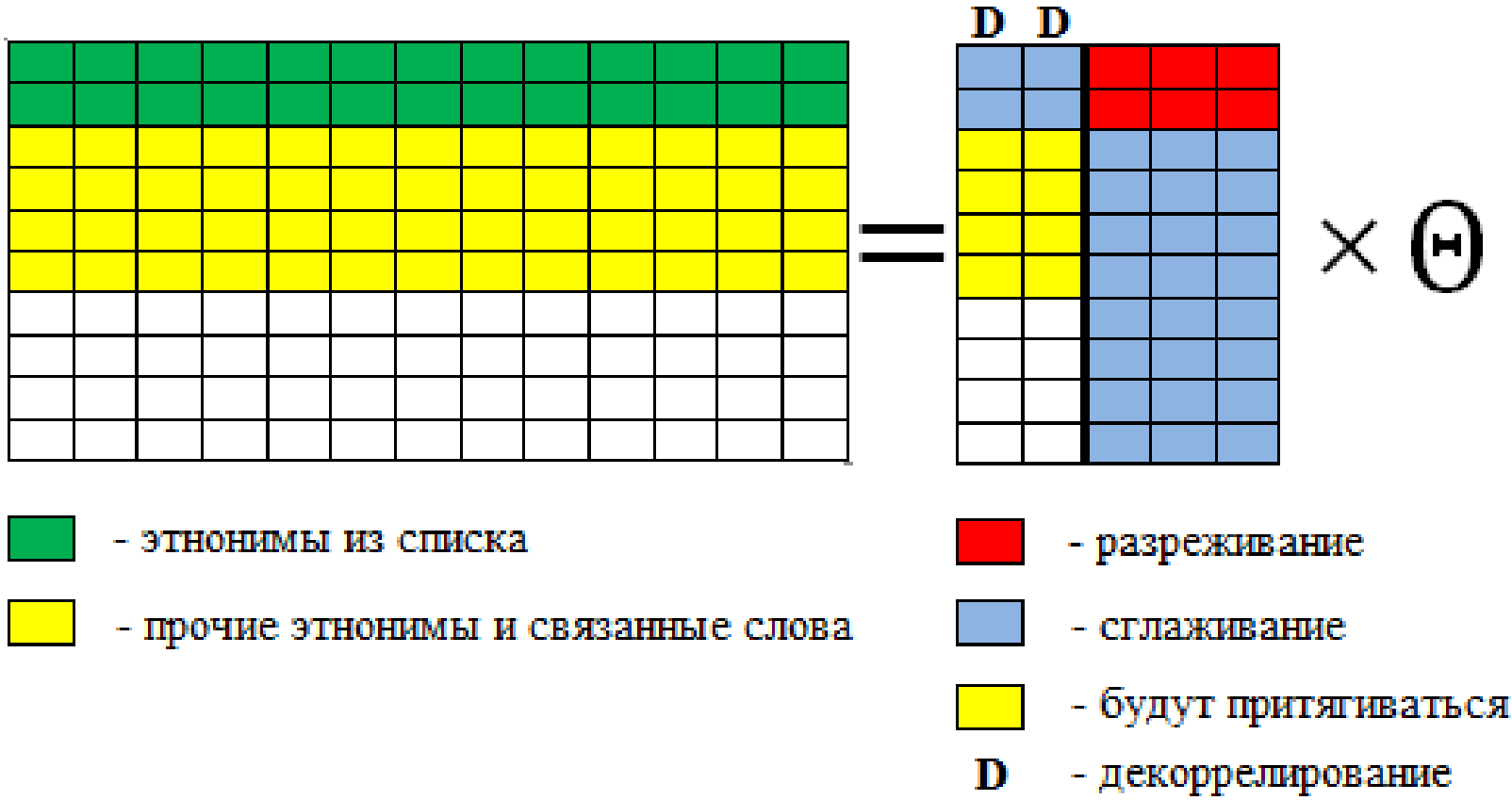


- этнонимы из списка
- прочие этнонимы и связанные слова
- разреживание
- сглаживание
- будут притягиваться

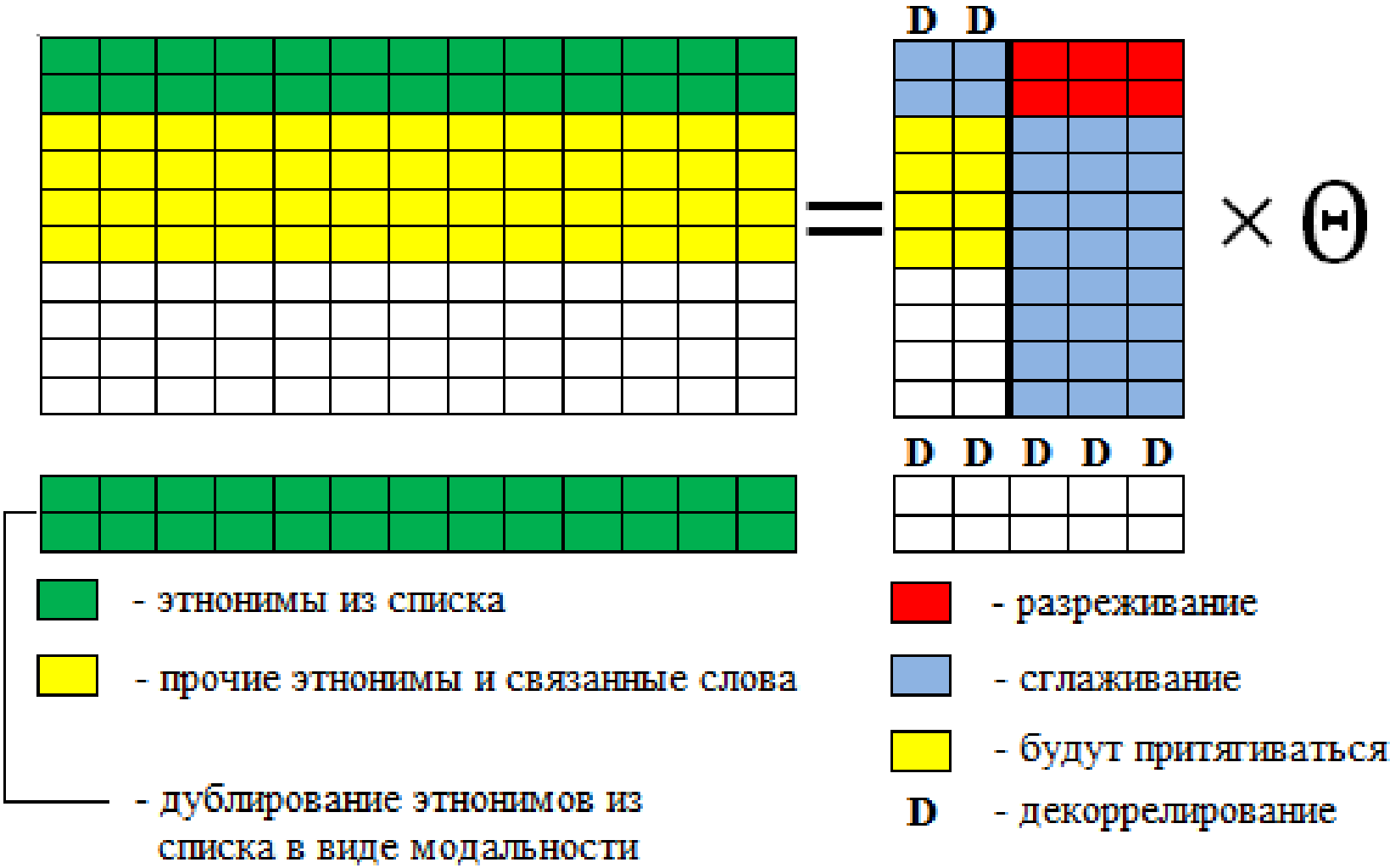
+ сглаживание обычных слов



+ декорреляция этничных тем



+ модальность этнонимов



Данные и код

Содержимое архива:

- директория `lj_batches` — содержит батчи и словари.
- `lj_experiment.ipynb` — ноутбук с кодом экспериментов.

- `lj_full.vw.txt` — «мешок слов» коллекции LJ (после описанной предобработки).

- `ethnic_list_lj.txt` — список этнонимов LiveJournal.

- `ethnic_list.txt` — список этнонимов (список).

- `ethnic_lj.txt` — список этнонимов (исходник).
- `global_dict.txt` — список всех слов коллекции LJ с абсолютными частотами.
- `lj_parser.ipynb` — ноутбук с парсером.

Подготовленные словари BigARTM

Следующие файлы со словарями идут «из коробки»:

1. `dictionary` - все слова имеют вес, равный своей частоте в коллекции;
2. `dictionary_ethnic_non_weighted` - этничные слова имеют вес 1, прочие 0;
3. `dictionary_ethnic_weighted` - этничные слова имеют вес, равный своей частоте в коллекции, прочие 0;
4. `dictionary_ethnic_weighted_modal` - этничные слова имеют вес, равный своей частоте в коллекции, прочие 0, и этничные продублированы как отдельная модальность.
5. `dictionary_ethnic_non_weighted_modal` - этничные слова имеют вес 1, прочие 0, и этничные продублированы как отдельная модальность, в которой они имеют вес, равный своей частоте в коллекции.

Настройка модели

Можно производить настройку следующих величин:

- Число тем: `num_ethnic_topics`, `num_general_topics`
- Используемый словарь: нужное имя файла со словарём в вызове `model.load_dictionary()`
- Коэффициент сглаживания этничных слов в этничных темах: `tau` в `SmoothPhiEthnic`
- Коэффициент разреживания этничных слов в общих темах: `tau` в `SparsePhiGeneral`
- Коэффициент сглаживания общих слов в общих темах: `tau` в `SmoothPhiGeneral`
- Коэффициент декоррелирования этничных тем: `tau` в `DecorrelatorPhiEthnic`

Настройка модели

- Коэффициент разреживания этничных тем в Θ: `tau` в `SparseThetaEthnic`
- Коэффициент сглаживания общих тем в Θ: `tau` в `SmoothThetaGeneral`
- Число итераций по документу: `num_document_passes` в вызове `model.fit_online()`
- Частота обновления Φ (раз в заданное число батчей): `update_every` в вызове `model.fit_online()`
- Число проходов по коллекции: количество вызовов `model.fit_online()`

Оценивание качества

Грубое оценивание качества модели можно произвести «на глаз».

Функции `print_scores(model)` и `save_scores(model)` печатают и сохраняют в файл топ-слова каждой темы.

- Тема считается приемлимой, если в ней собрались родственные этнонимы.
- Тема считается хорошей, если она притянула родственные этнонимы и общие слова.
- Тема плохая, если она не этничная, либо представляет собой просто набор несвязанных этнонимов.

Примеры тем

Приемлимая тема:

кавказ, чеченец, дагестан, кавказский, кавказец, алиса, ингушетия, республика, аварец, осетия, дагестанец, северный, северн, ингуш, дагестанский, россиянин, горсада, чечня, чеченский, аут, ставрополье, азербайджанец, национальность, гкб, кабардино-балкария

Хорошая тема:

латвия, эстония, эстонский, латвийский, латыш, русич, рита, латышский, прибавлять, рижский, талант, самозванец, прибалтийский, боярский, игнатово, тамплиер, грамота, опричник, игнор, таллинн, россиянин, ливонский, сакко, царын, курбанов

Ещё примеры тем

А вот так выглядят плохие темы:

философия, субъект, дискурс, философский, философ, субъективный, бытие, лин, авиакатастрофа, парадигма, метафаера, сущий, россиянин, диалектик, русич, вебер, гегель, химкинский, фетисов, метафизический, субъектный, лосев, разбиваться, аварец, постмодернизм

Большинство же тем примерно такие:

зеркало, брюшко, люстра, голландский, рейес, отражение, эйр, проезжий, гильтон, климовск, караваево, хокон, фейков, захарченко, зеркальный, антверпен, энгр, монография, малинин, французский, кавакоса, многоженство, кудри, линия, мемлюков

Текущие результаты

Модель	Лучших тем	Хороших тем	Удовл. тем	Всего
PLSA (300)	9	11	18	38
PLSA (400)	12	15	17	44
Сглаж. + разреж. + декор. (200 + 100)	18	33	20	71
Сглаж. + разреж. + декор. (250 + 150)	21	27	20	68
Сглаж. + разреж. + декор. + модальность (300 + 100)	28	23	23	74
Сглаж. + разреж. + декор. + модальность (250 + 150)	22	25	33	80
Сглаж. + разреж. + декор. + модальность (250 + 150) (настр.)	32	42	40	104

Примеры лучших тем

Тема про таджиков и узбеков:

мигрант, страна, россия, миграция, азия, нелегальный, миграционный, таджикистан, гастарбайтер, гражданка, трудовой, рабочий, фмс, коренево, среднее, узбекистан, таджик, проблема, русский, население

Тема про канадцев:

команда, игра, игрок, канадский, сезон, хоккей, амур, сборная, играть, болельщик, победа, кубок, клюв, счет, забирать, хоккейный, выигрывать, хоккеист, чемпионат, шайба

Тема про азербайджанцев:

русский, азербайджан, азербайджанец, россия, азербайджанский, таксист, диаспора, анапа, народ, бан, местами, москва, страна, армянин, горин, время, слово, рынок, старое, группа