Аддитивная регуляризация тематических моделей в задаче анализа этносоциального дискурса

Мурат Апишев
great-mel@yandex.ru

МГУ им. Ломоносова, Яндекс

April 14, 2016

Тематическое моделирование

Тематическое моделирование — приложение машинного обучения к статистическому анализу текстов.

Тема — терминология предметной области, набор терминов (униграм или $n-$грам) часто встречающихся вместе в документах.

Тематическая модель исследует скрытую тематическую структуру коллекции текстов:

- *тема* t — это вероятностное распределение $p(w|t)$ над терминами w
- *документ* d — это вероятностное распределение $p(t|d)$ над темами t

Приложения — информационный поиск по длинным текстовым запросам, классификация, категоризация и суммаризация текстов.

Задача тематического моделирования

Дано: W — словарь терминов (униграм или n–биграм),
D — коллекция текстовых документов $d \subset W$,
n_{dw} — счётчик частоты появления слова w в документе d.

Найти: модель $p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$ с параметрами $\underset{W \times T}{\Phi}$ и $\underset{T \times D}{\Theta}$:
$\phi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t,
$\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d.

Критерий максимизация логарифма правдоподобия:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} \;\to\; \max_{\phi,\theta};$$

$$\phi_{wt} \geqslant 0; \quad \textstyle\sum_w \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \textstyle\sum_t \theta_{td} = 1.$$

Проблема: задача стохастического матричного разложения некорректно поставленная: $\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$.

PLSA и EM-алгоритм

Максимизация логарифма правдоподобия:

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

EM-алгоритм: метод простых итерация для решения системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right) \\ \text{M-шаг:} & \phi_{wt} = \underset{w\in W}{\text{norm}}\left(n_{wt}\right), \quad n_{wt} = \sum\limits_{d\in D} n_{dw}p_{tdw} \\ & \theta_{td} = \underset{t\in T}{\text{norm}}\left(n_{td}\right), \quad n_{td} = \sum\limits_{w\in d} n_{dw}p_{tdw} \end{cases}$$

где $\underset{i\in I}{\text{norm}}\, x_i = \dfrac{\max\{x_i,0\}}{\sum\limits_{j\in I}\max\{x_j,0\}}$

ARTM и регуляризованный EM-алгоритм

Максимизация логарифма правдоподобия с дополнительными аддитивными регуляризаторами R:

$$\sum_{d\in D}\sum_{w\in W} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

EM-алгоритм: метод простых итераций для системы уравнений

$$\text{E-шаг:} \quad p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w\in W}{\text{norm}}\left(n_{wt} + \phi_{wt}\dfrac{\partial R}{\partial \phi_{wt}} \right), & n_{wt} = \sum\limits_{d\in D} n_{dw} p_{tdw} \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(n_{td} + \theta_{td}\dfrac{\partial R}{\partial \theta_{td}} \right), & n_{td} = \sum\limits_{w\in d} n_{dw} p_{tdw} \end{cases}$$

Примеры регуляризаторов

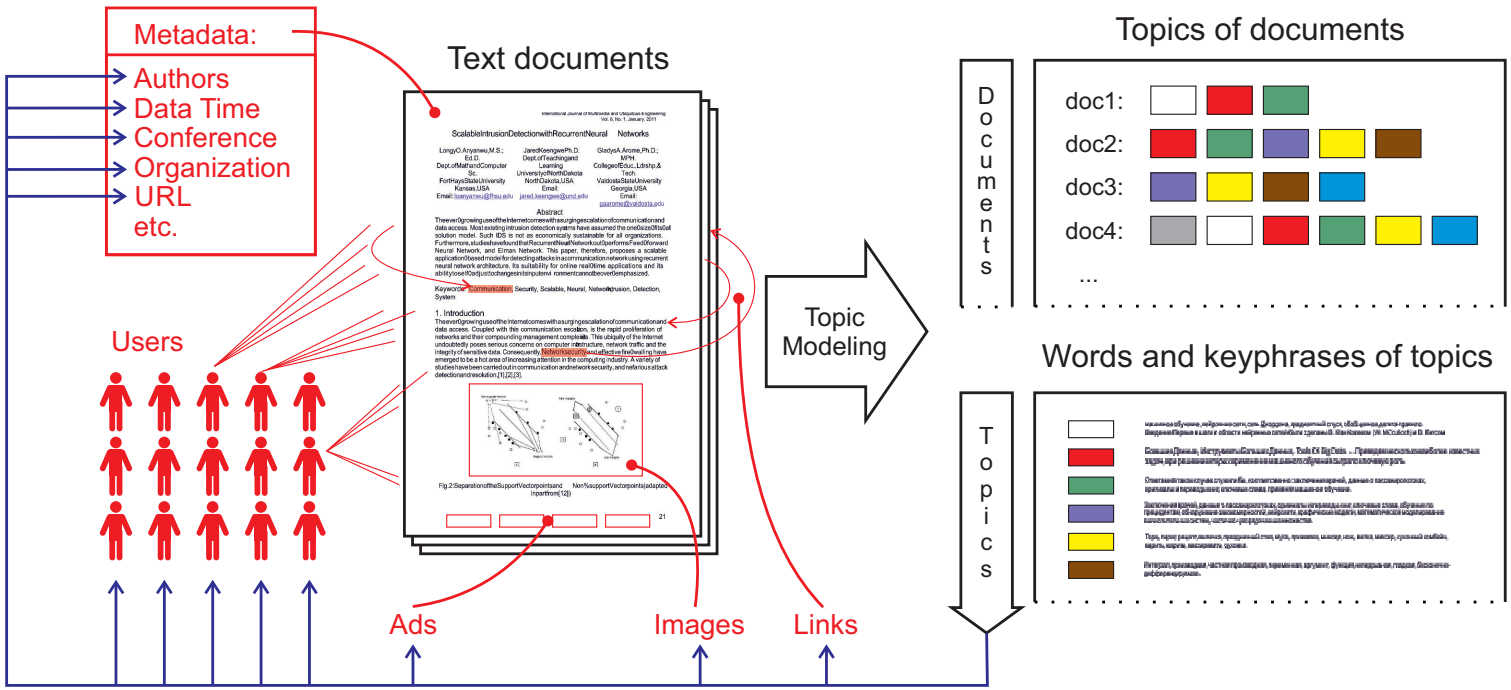
Многие байесовские модели могут быть интерпретированы в терминах ARTM.

Примеры регуляризаторов:

1. Сглаживание Φ / Θ (приводит к известной модели LDA)
2. Разреживание Φ / Θ
3. Декорреляция тем в Φ
4. Частичное обучение
5. Максимизация когерентности тем
6. Отбор тем
7. ...

Мультимодальная тематическая модель

Мультимодальная тематическая модель распределения тем на терминах $p(w|t)$, авторах $p(a|t)$, метках времени $p(y|t)$, изображениях $p(o|t)$, связанных документвх $p(d'|t)$, рекламных баннерах $p(b|t)$, пользователях $p(u|t)$, и объединяет все эти модальности в одно тематическую модель.



M-ARTM и мультимодальный регулризованный EM-алгоритм

W^m — словарь терминов m-й модальности, $m \in M$,
$W = W^1 \sqcup W^m$ как объединение словарей всех модальностей.

Максимизация логарифма мультимодального правдоподобия с аддитивными регуляризаторами R:

$$\sum_{m \in M} \lambda_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} + R(\Phi, \Theta) \;\to\; \max_{\Phi, \Theta}$$

EM-алгоритм: метод простых итерация для системы уравнений

$$\text{E-шаг:} \quad p_{tdw} = \underset{t \in T}{\text{norm}}\left(\phi_{wt} \theta_{td}\right)$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w \in W^m}{\text{norm}}\left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right), \quad n_{wt} = \sum_{d \in D} \lambda_{m(w)} n_{dw} p_{tdw} \\ \theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right), \quad n_{td} = \sum_{w \in d} \lambda_{m(w)} n_{dw} p_{tdw} \end{cases}$$

Проект BigARTM

Особенности BigARTM:

- Быстрая[1] параллельная и онлайновая обработка данных;
- Поддержка мультимодальных регуляризованных тематических моделей;
- Встроенная расширяемая библиотека регуляризаторов и метрик качества;

Сообщество BigARTM:

- Открытый репозиторий `https://github.com/bigartm`
- Описание и документация `http://bigartm.org`

Лицензия BigARTM и программные особенности:

- Бесплатное коммерческое использование (BSD 3-Clause license)
- Кроссплатформенная — Windows, Linux, Mac OS X (32 bit, 64 bit)
- Программные API: command line, C++, Python

[1] Vorontsov K., Frei O., Apishev M., Romov P., Dudarenko M. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections Analysis of Images, Social Networks and Texts. 2015

Поиск этнорелевантного контена в блогосфере

Создание концепции и методологии для мониторинга остояния межэтничных отношений по данным социальных медиа.

Задачи тематчиеского моделирования в проекте:

1. Извлечение этно-релевантных тем из данных социальных медиа
2. Распознавание событийных и непрерывных во времени тем
3. Сантимент-анализ этничного дискурса

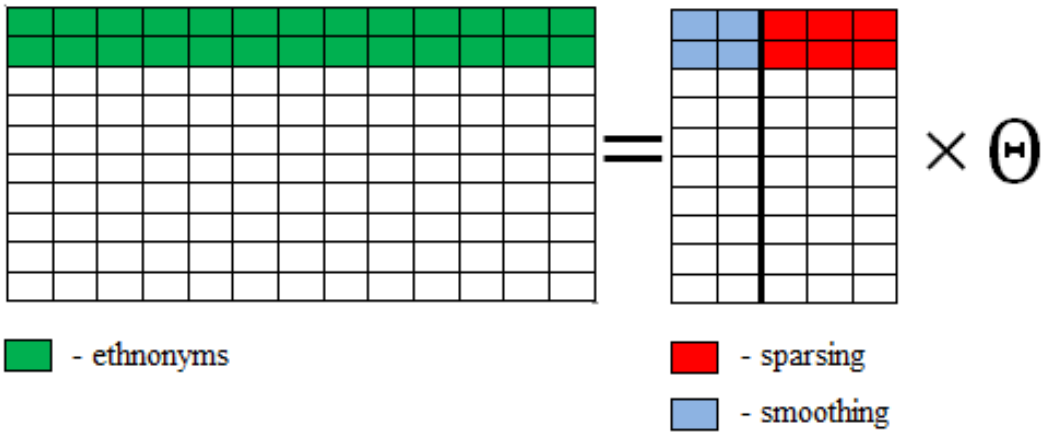
Грант Российского Научного Фонда 15-18-00091 (2015–2017)
(Высшая Школа Экономики, С.-Петербургская Школа социальных и общественных наук, Лаборатория интрнет-исследований ЛИНИС)

Примеры этнонимов для частичного обучения

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

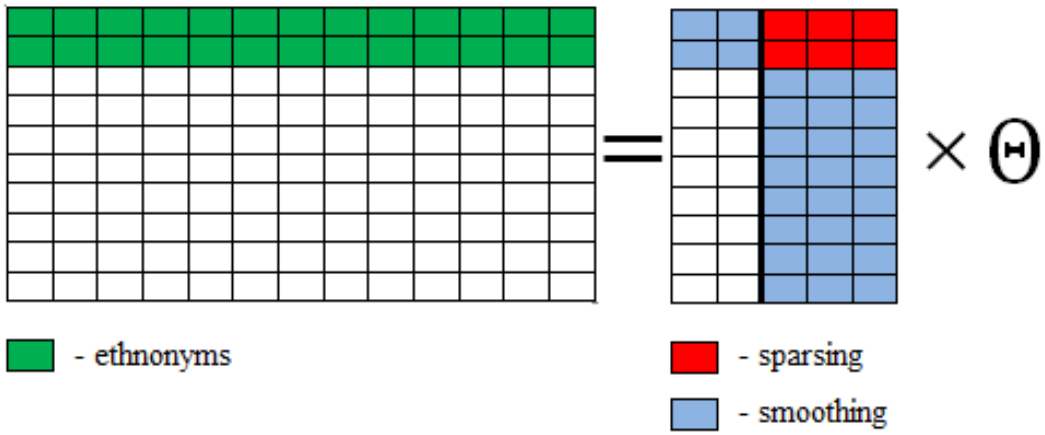
Регуляризация для поиска этничных тем

- сглаживание этнонимов в этничных темах
- разреживание этнонимов в общих темах
-
-
-



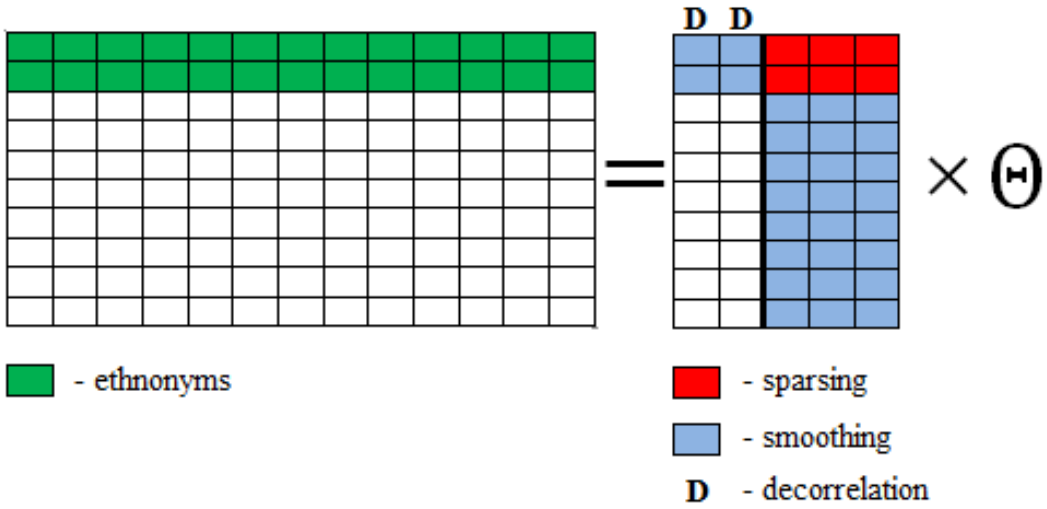
Регуляризация для поиска этничных тем

- сглаживание этнонимов в этничных темах
- разреживание этнонимов в общих темах
- сглаживание не-этнонимов в общих темах
-
-



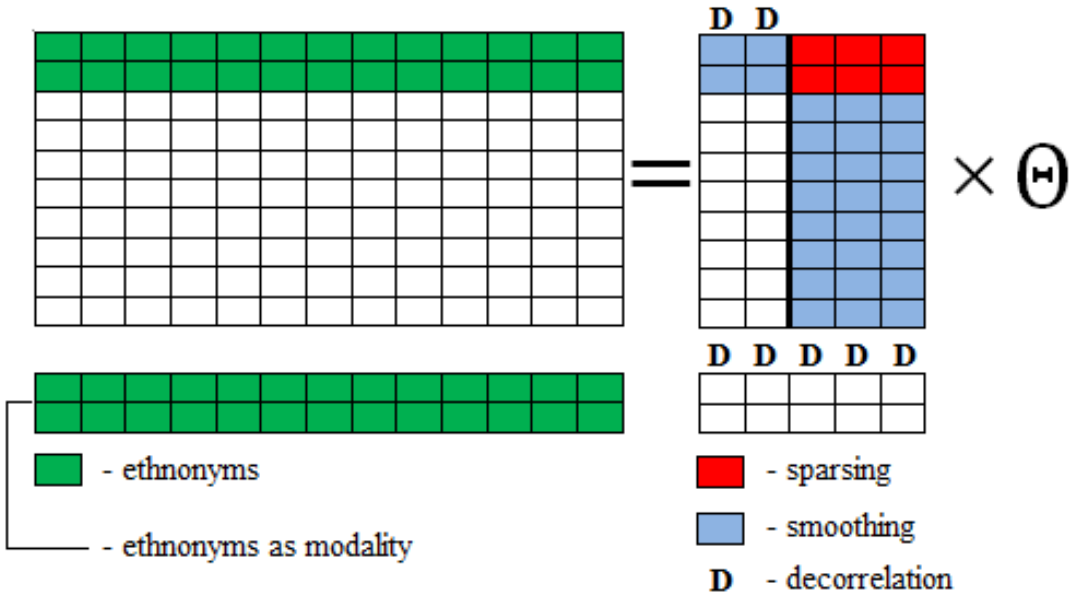
Регуляризация для поиска этничных тем

- сглаживание этнонимов в этничных темах
- разреживание этнонимов в общих темах
- сглаживание не-этнонимов в общих темах
- декоррелирование этничных тем
-



Регуляризация для поиска этничных тем

- сглаживание этнонимов в этничных темах
- разреживание этнонимов в общих темах
- сглаживание не-этнонимов в общих темах
- декоррелирование этничных тем
- дублирование этнонимов в качестве новой модальности и декоррелирование тем в ней



Experiment

- Коллекция Живого Журнала: 1.58M документов
- 860K слов в словаре после лемматизации
- 90K слов после фильтрации
  - коротких слов с длиной $\leqslant 2$,
  - редких слов со встречаемости в коллекции $n_w < 20$,
  - нерусских слов
- 250 этнонимов в словаре

ARTM с частичным обучением для поиска этнорелевантных тем

Число и качество этничных тем, найденных моделью:

модель	ethnic $\|S\|$	background $\|B\|$	$++$	$+-$	$-+$	coh_{20} [2]	$tfidf_{20}$
PLSA		400	12	15	17	-1447	-1012
LDA		400	12	15	17	-1540	-1121
ARTM-4	250	150	21	27	20	-1651	-1296
ARTM-5	250	150	38	42	30	-1342	-908

- ARTM-4:
  - этничные темы: декорреляция, сглаживание этнонимов
  - фоновые темы: сглаживание, разреживание этнонимов
- ARTM-5:
  - ARTM-4 + декоррелируемая модальность этнонимов

[2] Когерентность и TF-IDF когереность — метрики, коррелирующие с человеческими оценками интерпретируемости. Тема тем лучше, чем выше её когерентность.

Ethnic topics examples

(русские): русский, князь, россия, татарин, великий, царить, царь, иван, император, империя, грозить, государь, век, московская, екатерина, москва,
(русские): акция, организация, митинг, движение, активный, мероприятие, совет, русский, участник, москва, оппозиция, россия, пикет, протест, проведение, националист, поддержка, общественный, проводить, участие,
(славяне, византийцы): славянский, святослав, жрец, древние, письменность, рюрик, летопись, византия, мефодий, хазарский, русский, азбука,
(сирийцы): сирийский, асад, боевик, район, террорист, уничтожать, группировка, дамаск, оружие, алесио, оппозиция, операция, селение, сша, нусра, турция,
(турки): турция, турецкий, курдский, эрдоган, стамбул, страна, кавказ, горин, полиция, премьер-министр, регион, курдистан, ататюрк, партия,
(иранцы): иран, иранский, сша, россия, ядерный, президент, тегеран, сирия, оон, израиль, переговоры, обама, санкция, исламский,
(палестинцы): террорист, израиль, терять, палестинский, палестинец, террористический, палестина, взрыв, территория, страна, государство, безопасность, арабский, организация, иерусалим, военный, полиция, газ,
(ливанцы): ливанский, боевик, район, ливан, армия, террорист, али, военный, хизбалла, раненый, уничтожать, сирия, подразделение, квартал, армейский,
(ливийцы): ливан, демократия, страна, ливийский, каддафи, государство, алжир, война, правительство, сша, арабский, али, муаммар, сирия,
(евреи): израиль, израильский, страна, израил, война, нетаньяху, тель-авив, время, сша, сирия, египет, случай, самолет, еврейский, военный, ближний,

Заключение

- BigARTM — открытая библиотека с поддержкой ARTM и мультимодальных моделей.
- Комбинация восьми регуляризаторов в задаче поиска этнорелевантных тем показала свое превосходство над LDA.
- Дальнейшие исследование предполагают работу с короткими текстами коллекции Вконтакте, использование биграм этнонимов и модальностей меток времени и геотегов для мониторинга изменения тем во времени и по регионам.

Contacts: bigartm.org, great-mel@yandex.ru