Мультимодальные тематические модели социальных сетей

Костюк А. А.

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель: д.ф.-м.н. К. В. Воронцов

Группа 174, весна 2015

Цель исследования

Цель работы:

- Построить мультимодальную тематическую модель для коллекции текстов Живого Журнала;
- Показать, что такая модель будет более эффективна, чем модель LDA;
- С помощью этой модели выполнить информационный поиск и показать зависимости частоты употребления тем от метаданных (времени).

Основные понятия

Пусть C — коллеция документов, W — словарь терминов. Каждый документ $d \in C$ — последовательность треминов $w \in W$. Пусть существует конечное множество тем T.

Согласно гипотезе условной независимости,

$$p(w \mid d) = \sum_{t \in T} p(w \mid t) \cdot p(t \mid d)$$

Вероятностная модель описывает порождение коллекции C по известным вероятностям $p(w \mid t)$ и $p(t \mid d)$.

Постановка задачи

Построение тематической модели — обратная задача: по коллекции C требуется восстановить вероятности $p(w \mid t)$ и $p(t \mid d)$, её породившие.
Обычно задача сводится к поиску матрицы терминов тем Φ и матрицы тем документов Θ:

$$\Phi = (\phi_{wt})_{W \times T}; \qquad \phi_{wt} = p(w \mid t)$$

$$\Theta = (\theta_{td})_{T \times D}; \qquad \theta_{td} = p(t \mid d)$$

ARTM

В рамках аддитивной регуляризации тематических моделей (ARTM) выполняется:

$$\sum_{d \in D} \sum_{w \in d} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) \to \max$$

,
где $R(\Phi, \Theta) = \sum_{i=1}^{r} \tau_i R_i$.
Преимущества ARTM:

- Регуляризаторы могут не иметь вероятностной интерпретации;
- Можно регулировать важность каждого регуляризатора;
- Легко комбинировать существующие модели.

Экспериментальные данные

Для численного эксперимента была использована коллекция записей Живого Журнала, а также метаданные: имя автора каждой записи, дата и время её создания. После построения модели для оценки её качества сравним её с моделью LDA. В качестве показателей качества будем брать:

- перплексию $P = exp(-\frac{1}{n}L(\Phi, \Theta))$;
- разреженность матриц Φ и Θ.

Результаты построения модели

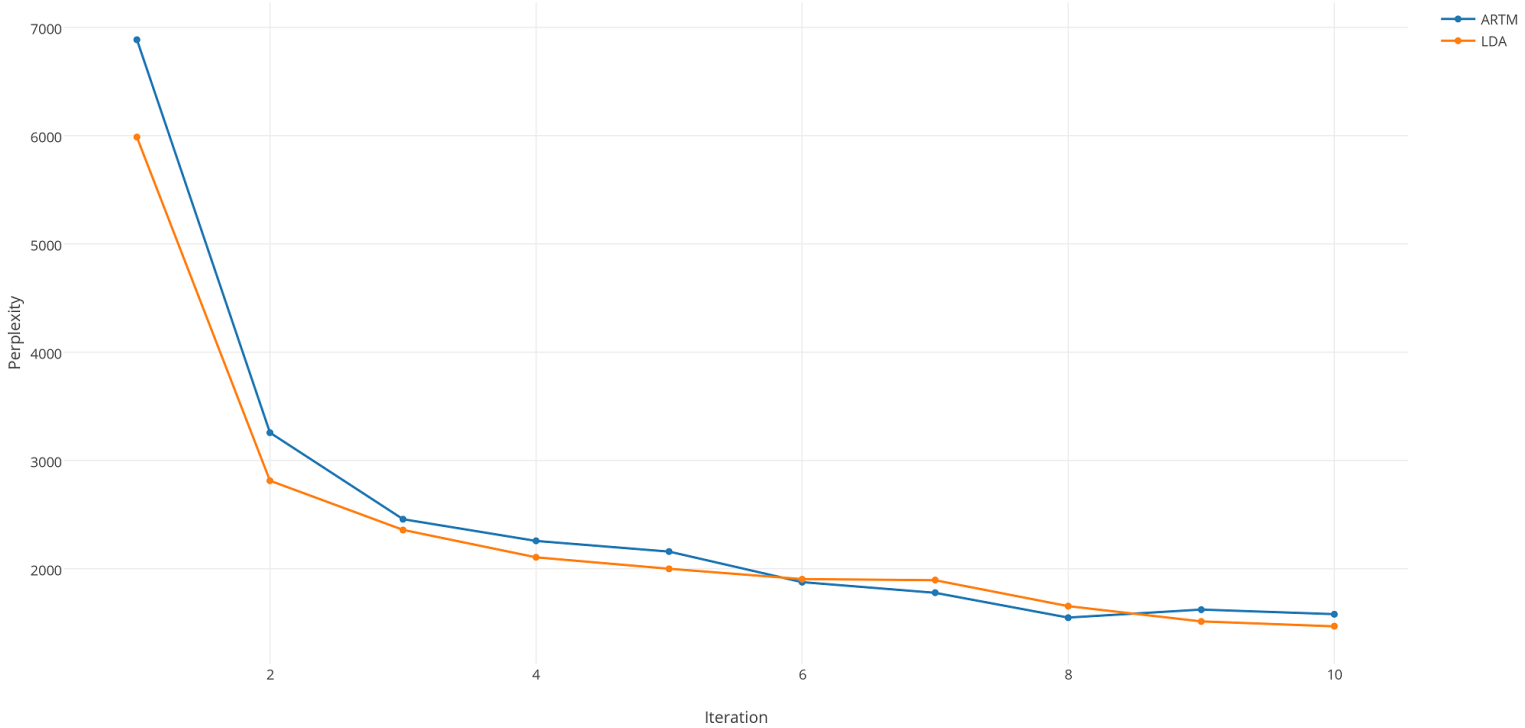

Рис.: Зависимость перплексии от числа тем (красный — LDA, синий — ARTM).

Результаты построения модели (2)

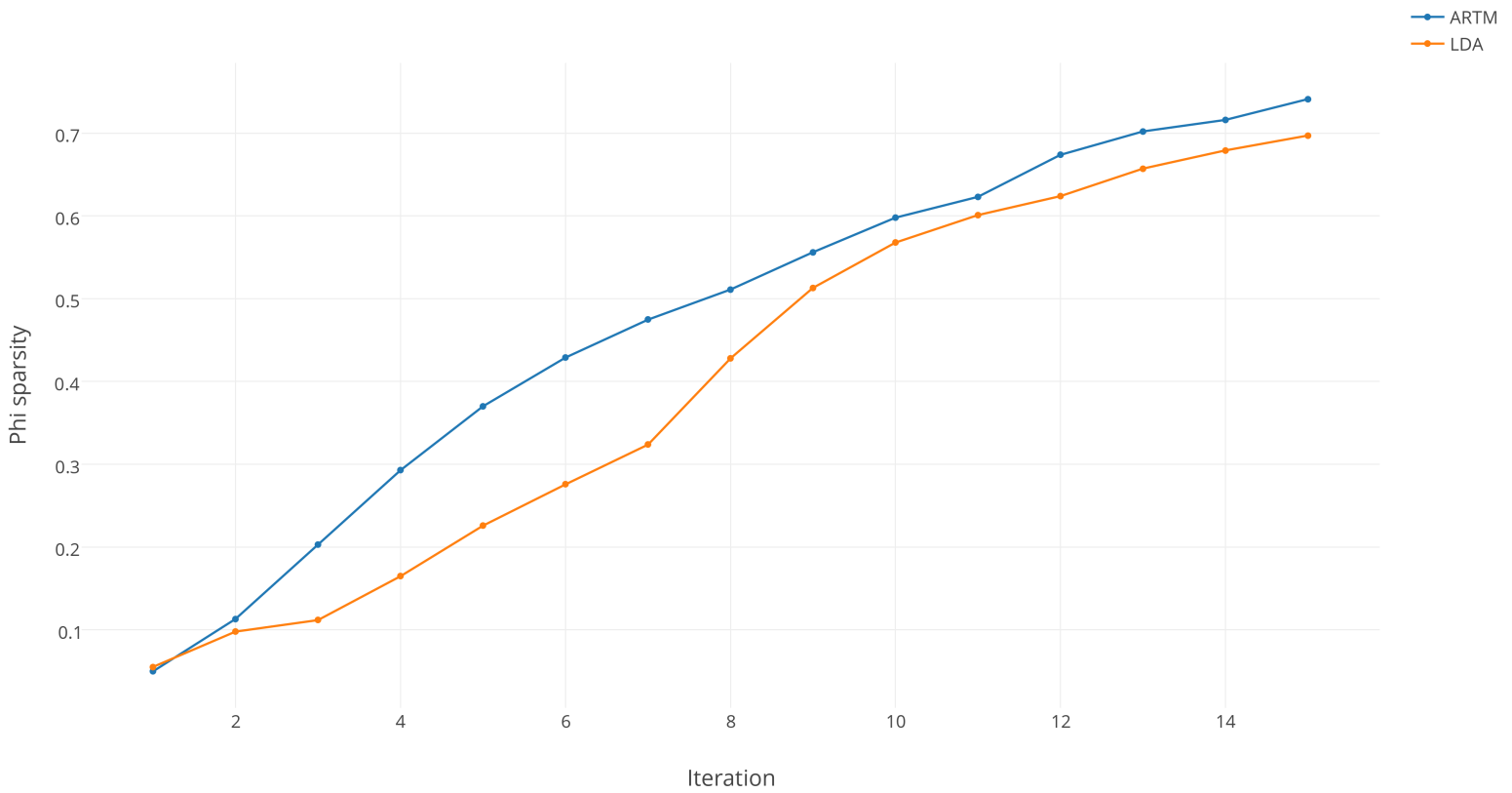

Рис.: Зависимость разреженности Φ от числа тем (красный — LDA, синий — ARTM).

Результаты построения модели (3)

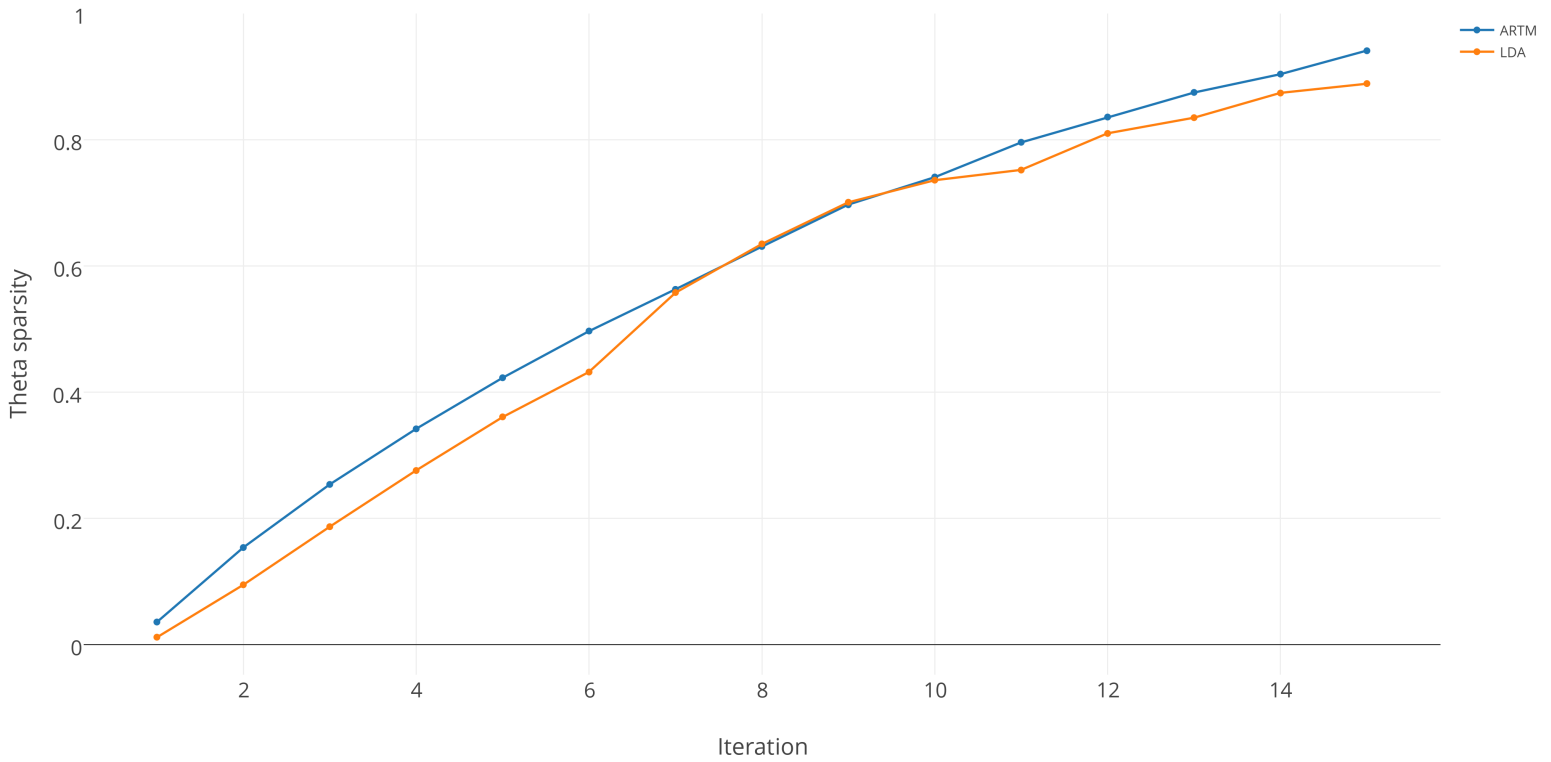

Рис.: Зависимость разреженности Θ от числа тем (красный — LDA, синий — ARTM).

Результаты анализа частоты тем

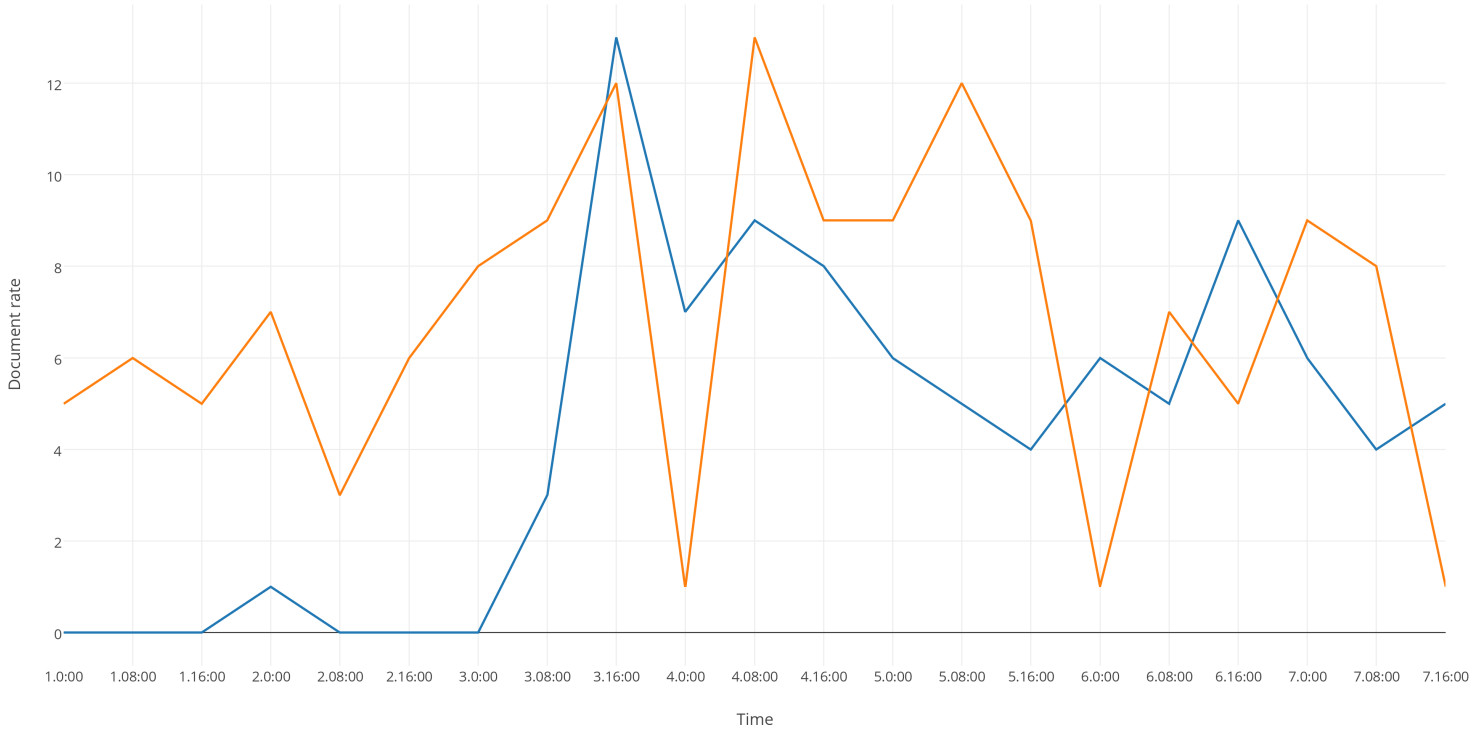

Рис.: Зависимость частоты употребления темы от времени. Ключевое слово — «американец»

Результаты анализа частоты тем(2)

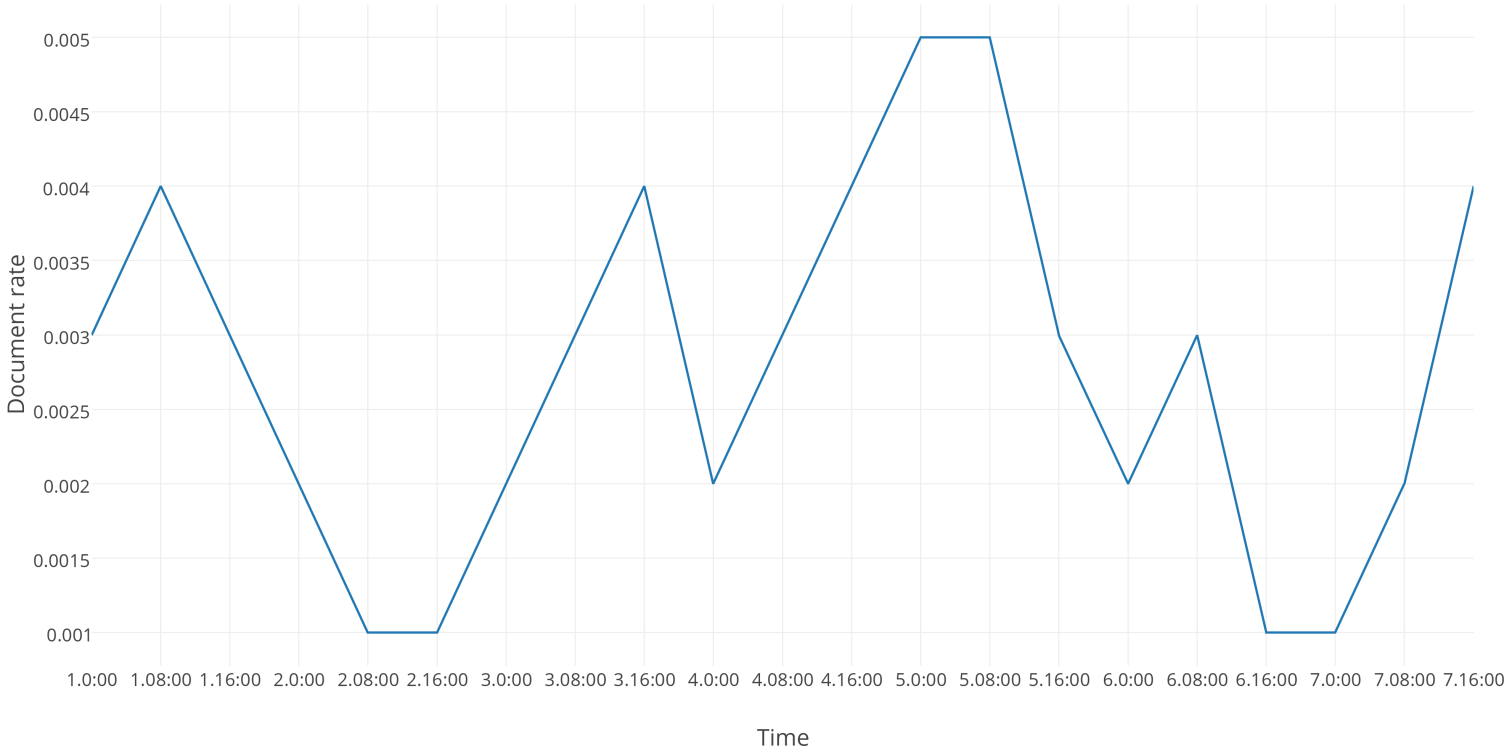

Рис.: Зависимость частоты употребления темы от времени. Ключевое слово — «еврей»

Выводы

- Построенная модель работает лучше, чем LDA;
- Модель может быть использована для поиска записей, связанных с обсуждением этничности.