Темпоральная тематическая модель коллекции пресс-релизов

Дмитрий Андреевич Персиянов

Московский физико-технический институт

Курс: Численные методы обучения по прецедентам (практика, В. В. Стрижов)/Группа YAD16, весна 2016

Цель исследования

Сравнить модель с модальностью времени с классической тематической моделью LDA. Провести анализ устойчивости модели со временем. Предложить способ отбора событийных тем.

Проблемы

В классической модели LDA темы не интерпретируемы и содержат фоновую лексику корпуса.
При разных начальных приближениях модель может давать различные темы.

Предположения

Использование дополнительных регуляризаторов и меток времени при построении модели позволит получить более интерпретируемые темы.

Литература

1. Воронцов К. В. Вероятностное тематическое моделирование, Москва, 2009.
2. Воронцов К. В. Аддитивная Регуляризация Тематических Моделей Коллекций Текстовых Документов, Доклады РАН, Т.455, №3. С.268-271, 2014.
3. David Hall, Daniel Jurafsky, Christopher D. Manning. Studying the History of Ideas Using Topic Models.
4. Xuerui Wang, Andrew McCallum. Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends.
5. Blei, David M and Ng, Andrew Y and Jordan, Michael I. Latent dirichlet allocation, the Journal of machine Learning research, 2003.

Постановка задачи тематического моделирования

W – словарь, из которого состоят документы D – коллекция документов $d = (w_1, ..., w_{n_d})$ Для каждого слова w известна его частота n_{dw} в данном документе **d**.

Предположения:

1. Появление каждого слова в каждом документе связано с некоторой латентной переменной t из некоторого множества тем T.
2. $D \times W \times T$ – дискретное вероятностное пространство, $|T| \ll |D|, |W|$.
3. d_i, w_i – просматриваемые, а темы t_i – латентные.
4. $p(w|d,t) = p(w|t)$ – гипотеза условной независимости.

Постановка задачи тематического моделирования

Используя формулу полной вероятности:

$$p(w|d) = \sum_{t \in T} p(t|d) p(w|t),$$

Оценка на распределение $p(w|d)$ известна: $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$.
Необходимо найти распределения $\phi_{wt} \equiv p(w|t)$ и $\theta_{td} \equiv p(t|d)$.
Оптимизационная задача:
$\log \mathscr{L}(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in W} n_{dw} \log \sum_{t \in T} \phi_{wt} \theta_{td} \longrightarrow \max_{\Phi, \Theta}$.

Постановка задачи тематического моделирования с модальностью времени

Каждому документу приписана метка времени $y \in Y$.
Гипотеза условной независимости: $p(y|d,t) = p(y|t)$.
Смесь распределений: $p(y|d) = \sum_{t \in T} p(t|d)p(y|t)$.
Оценка на распределение $p(y|d)$ известна, $\hat{p}(y|d) = [y = y_d]$.
Оптимизационная задача для матриц
$\Xi = (\xi_{yt})_{Y \times T}$ и $\Theta = (\theta_{td})_{T \times D}$:

$$\log \mathscr{L}(\Xi, \Theta) = \sum_{d \in D} \sum_{y \in Y} n_{dy} \log \sum_{t \in T} \xi_{yt} \theta_{td} \longrightarrow \max_{\Xi, \Theta}.$$

Полная оптимизационная задача:

$$\mathscr{L}(\Phi, \Theta, \Xi) = \mathscr{L}_1(\Phi, \Theta) + \tau \mathscr{L}_2(\Theta, \Xi),$$

$$\log \mathscr{L}(\Phi, \Theta, \Xi) \longrightarrow \max_{\Phi, \Theta, \Xi}.$$

Аддитивная регуляризация тематических моделей

К задаче оптимизации добавить еще r функционалов $R_i(\Phi, \Theta),\ i = 1, \ldots, r$ называемых *регуляризаторами*, каждый со своим неотрицательным весом τ_i:

$$\begin{cases} R(\Phi, \Theta) = \sum_i \tau_i R_i(\Phi, \Theta), \qquad \log \mathscr{L}(\Phi, \Theta) + R(\Phi, \Theta) \longrightarrow \max\limits_{\Phi, \Theta}, \\ \phi_{wt} \geqslant 0, \quad \theta_{td} \geqslant 0, \quad \sum\limits_w \phi_{wt} = 1, \quad \sum\limits_t \theta_{td} = 1. \end{cases}$$

Метрики для оценки модели

1. Перплексия – позволяет отслеживать сходимость метода оптимизации. Чем значение меньше, тем лучше.

$$\text{Perplexity}(\Phi, \Theta) = \exp\left(-\frac{1}{n}\sum_{d\in D}\sum_{w\in W} n_{dw} \log p(w|d)\right).$$

2. Разреженность матриц Φ и Θ – доля нулевых элементов.
3. Чистота – суммарная вероятность слов ядра:

$$\text{Purity}(t) = \sum_{w\in W_t} p(w|t) = \sum_{w\in W_t} \phi_{wt},$$

где $W_t = \{w \in W \mid p(t|w) > \delta\}$.

4. Контрастность – средняя вероятность встретить слова ядра в конкретной теме:

$$\text{Contrast}(t) = \frac{1}{|W_t|}\sum_{w\in W_t} p(t|w).$$

Мера событийности Delta-AUC

Будем считать, что множество Y имеет вид отрезка $[0; M]$. Для всех $0 < \Delta \leq M$ найдем $0 \leq y_0 \leq M - \Delta$ такой, что интеграл

$$S = \int_{y_0}^{y_0+\Delta} p(y|t)dy$$

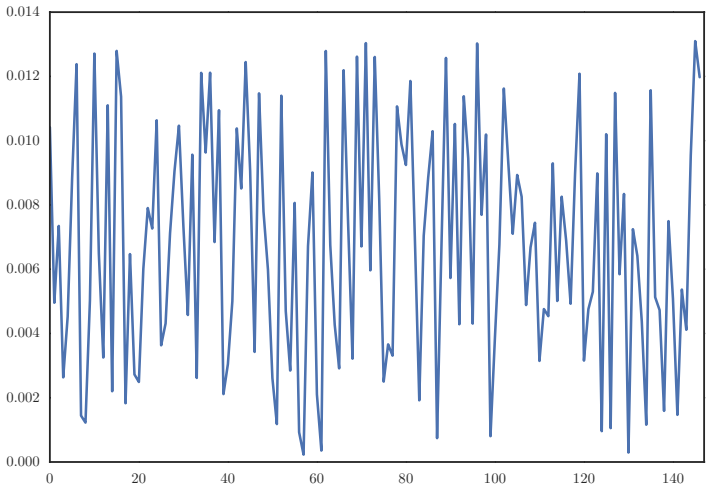
максимален. Построим график зависимости $S' = 1 - S$ от Δ. Для равномерного распределения он будет иметь вид убывающей прямой, а для событийных тем он будет сначала резко убывать вниз, затем плавно.

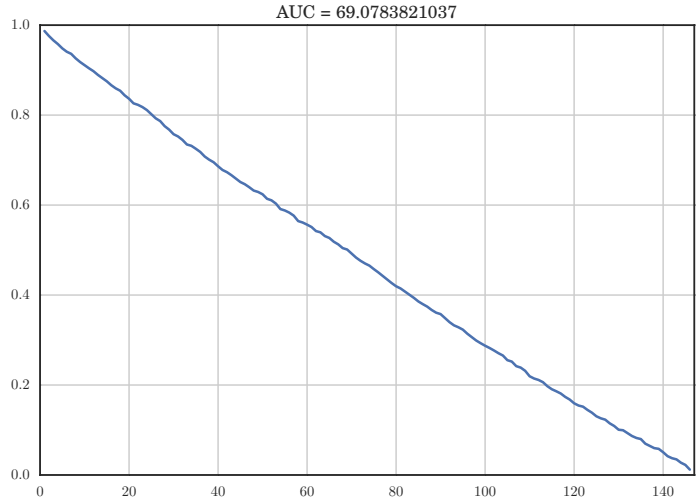
Определение 1

Назовем Delta-AUC мерой площадь под графиком $S'(\Delta)$.

Мера событийности Delta-AUC

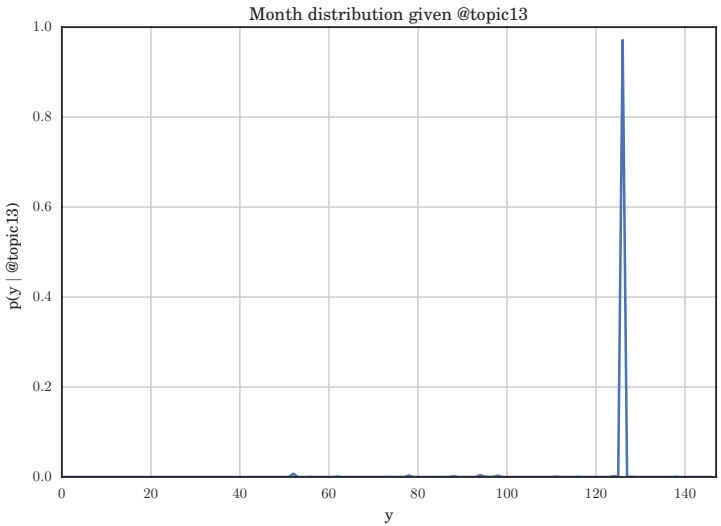
Равномерное распределение:

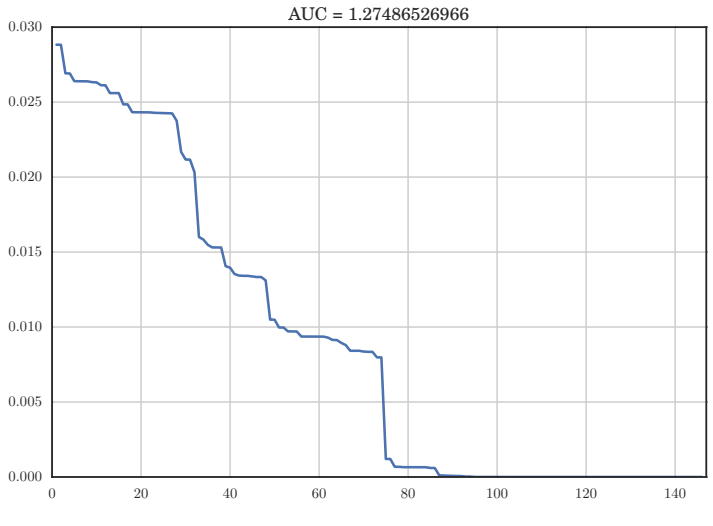




Мера событийности Delta-AUC

Распределение, характерное для событийной темы:





Цели экспериментов

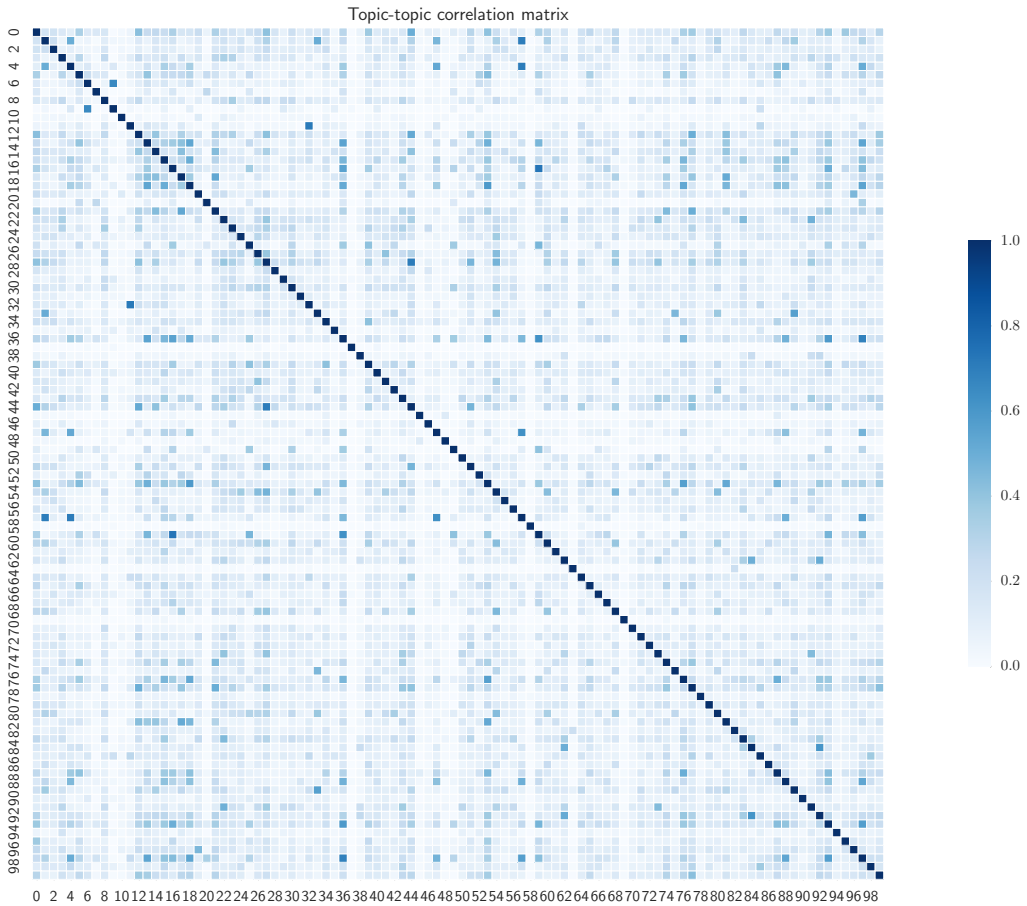
1. Сравнить модели LDA и ARTM с модальностью времени.
2. Проверить работоспособность метрики Delta-AUC.
3. Проанализировать устойчивость модели ARTM со временем.

Базовый эксперимент

Количество тем: 100.

Ключевые слова темы
state, foreign, secretary, president, relationship, very, unite, security, meet, thank, issue, together, minister, here, today
question, department, state, information, release, site, office, view, subject, u.s., internet, email, answer, page, should
state, question, designate, missile, under, russia, designation, act, order, entity, sanction, department, europe, decision, council
thank, remark, society, clinton, civil, secretary, here, today, welcome, president, very, minister, foreign, meet, madam
very, thank, much, state, here, many, inaudible, remark, country, together, great, clinton, visit, more, important
secretary, clinton, president, defense, question, unite, state, government, both, nato, administration, gate, missile, thank, very

Базовый эксперимент



Эксперимент с модальностью времени

Таблица: Примеры событийных тем и значения Delta-AUC для них

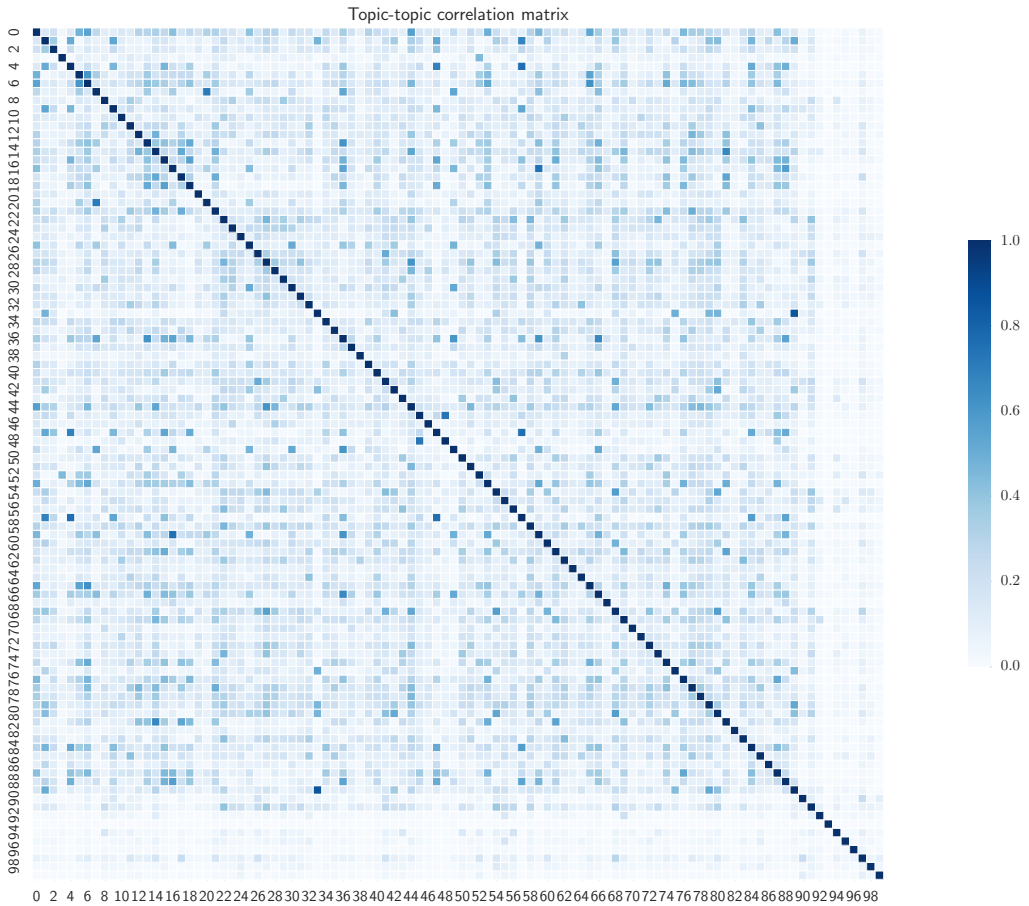
Ключевые слова темы	Delta-AUC
ambassador, question, report, state, material, right, facility, party, return, concern, fuel, government, answer, venezuela, reactor	0.0010
secretary, those, number, process, very, question, under, country, document, here, assistant, important, forward, brief, congress	0.0086
turkey, state, japan, investment, economic, unite, economy, trade, turkish, apec, vietnam, company, business, japanese, country	0.0123
very, minister, people, inaudible, president, here, important, prime, government, madame, opportunity, forward, help, future, secretary	0.0132

Эксперимент с модальностью времени

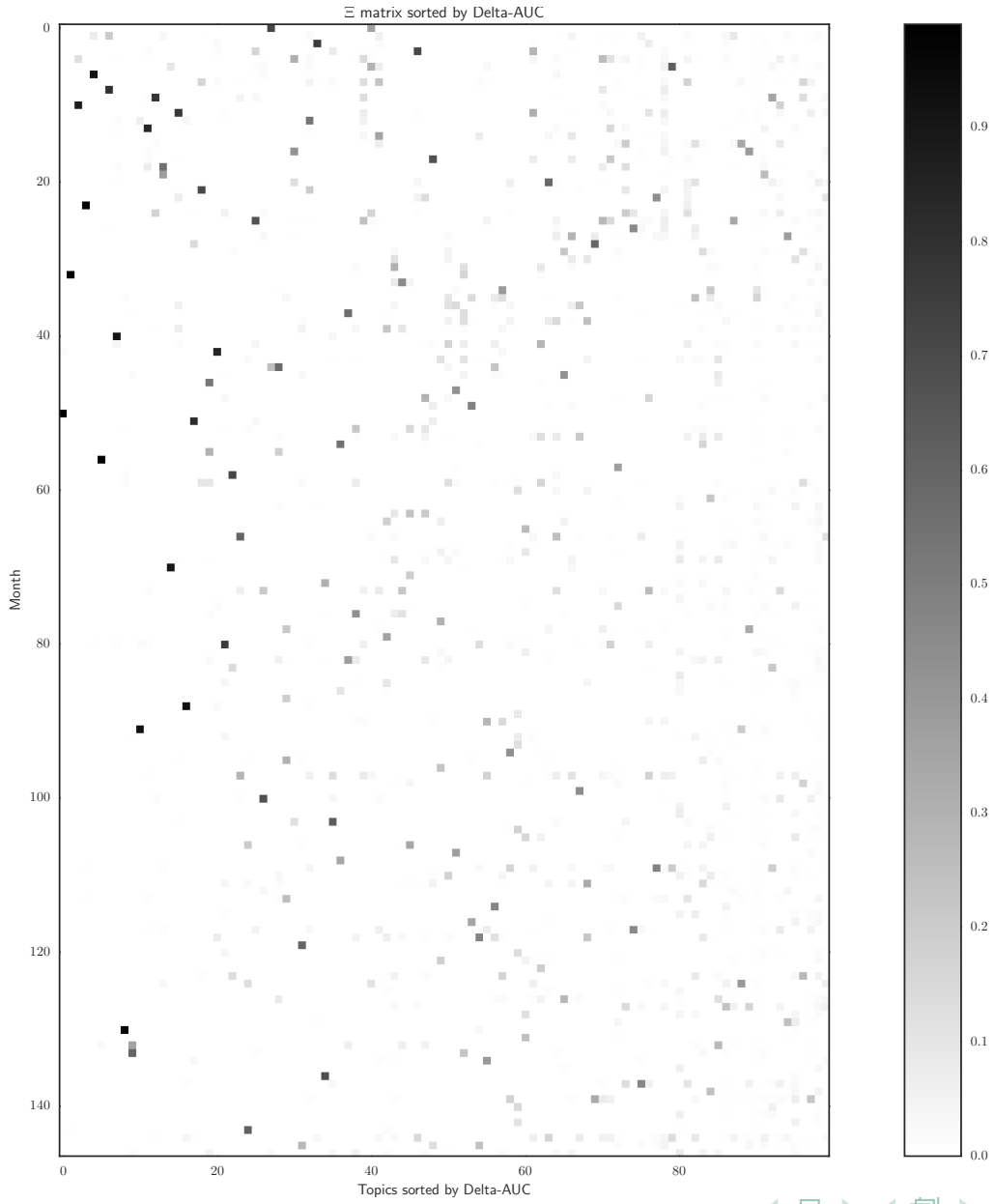
Таблица: Примеры несобытийных тем и значения Delta-AUC для них

Ключевые слова темы	Delta-AUC
u.s., designate, flood, water, designation, provide, company, conservation, sea, organization, entity, under, million, state, include	0.6989
mexico, development, fund, mexican, u.s., sector, initiative, support, country, law, group, train, private, corporation, bank	0.7119
aid, food, development, more, water, usaid, country, program, percent, need, million, people, resource, investment, administrator	0.7288
court, criminal, arrest, crime, tribunal, justice, former, war, rwanda, trial, sentence, charge, genocide, yugoslavia, conviction	0.7308
visa, entry, country, applicant, department, submit, program, application, may, select, receive, state	0.7321

Эксперимент с модальностью времени



Сортировка тем по событийности

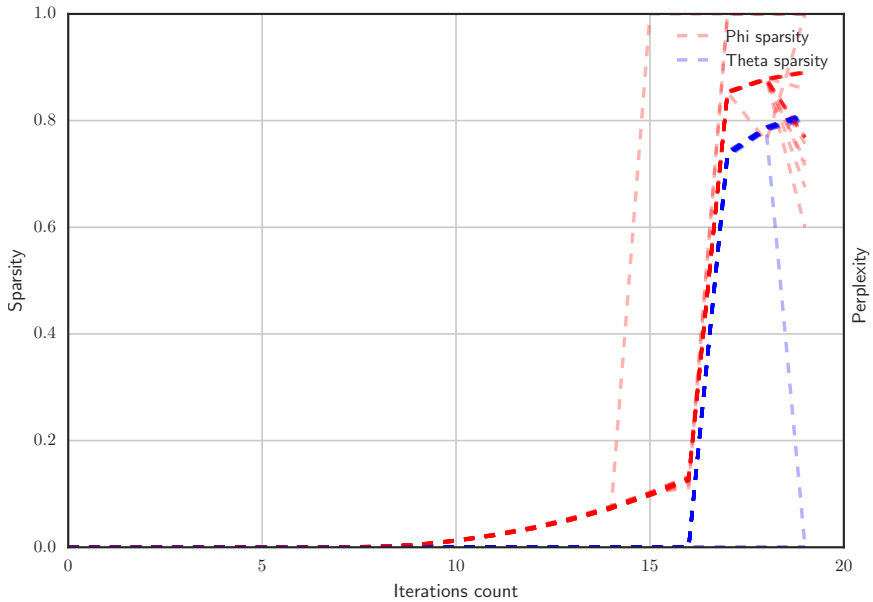


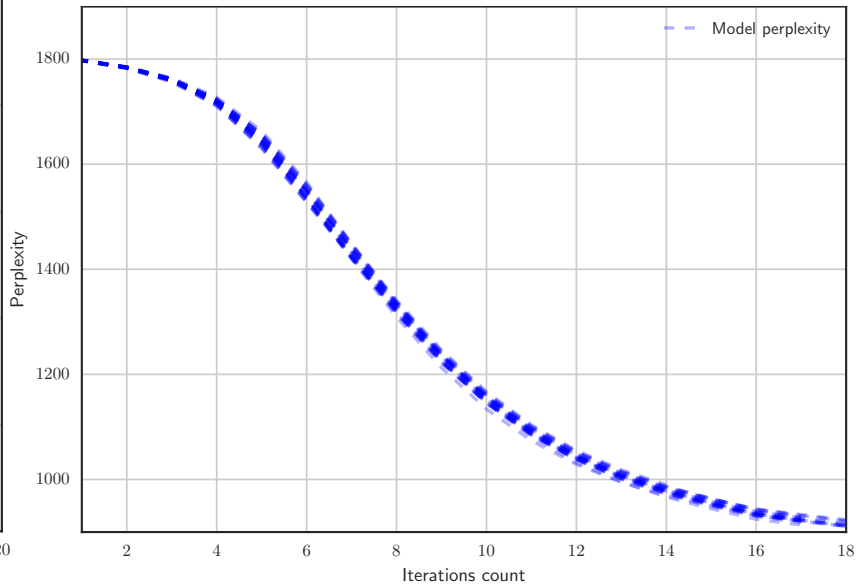
Сравнение полученных моделей

	LDA	ARTM + time
Sparsity Φ	0.0%	85.3%
Sparsity Θ	0.4%	73.8%
Purity	0.114	0.183
Contrast	0.410	0.540
Avg. correlation	0.109	0.133

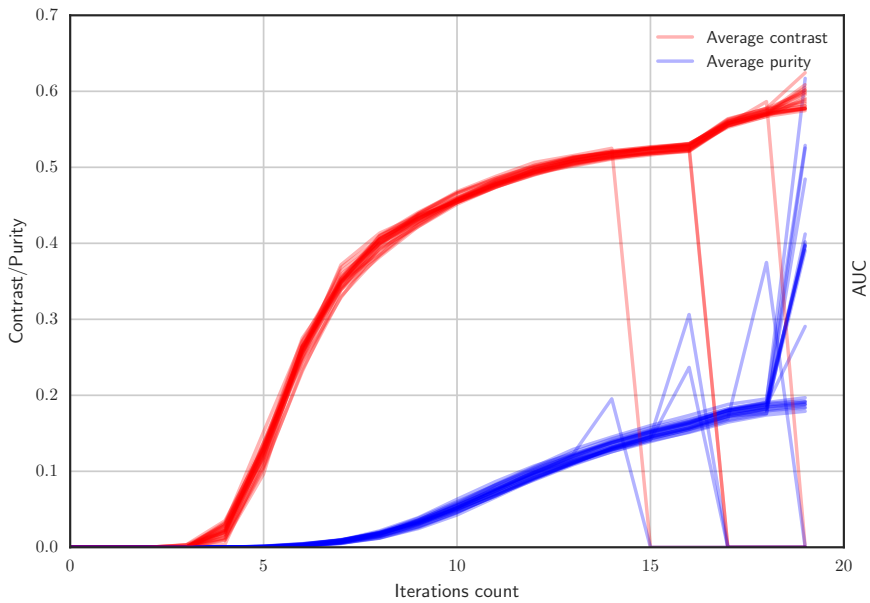
Анализ устойчивости модели ARTM + time

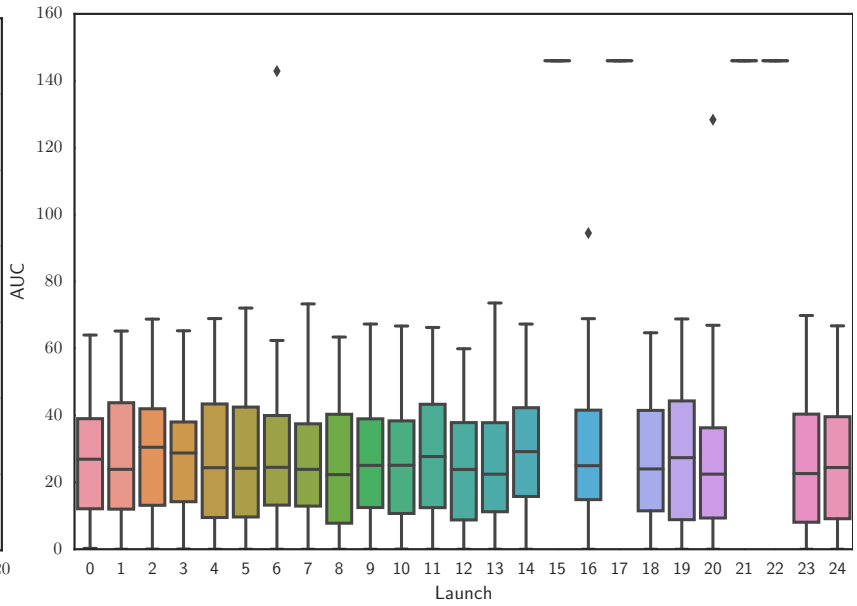
Модель обучена из 25 начальных приближений матриц Φ, Θ. В 4 случаях модель выродилась, в остальных случаях ведет себя устойчиво.





Анализ устойчивости модели ARTM + time





Заключение

- Построена тематическая модель в библиотеке BigARTM с модальностью времени.
- Произведено сравнение модели LDA и ARTM с модальностью времени.
- Предложена метрика для отбора событийных тем и произведен ее анализ.
- Проанализирована устойчивость модели ARTM с модальностью времени.

Дальнейшее исследование

Сравнение модели с современными байесовскими моделями со временем. Разработка критериев качества модели: устойчивости, полноты, интепретируемости.