Темпоральная тематическая модель коллекции пресс-релизов*

Персиянов Д. А.[1]***, Дойков Н. В.***[2]***, Воронцов К. В.***[2]
persiyanov@phystech.edu

Данная работа посвящена методам анализа тематической структуры большой текстовой коллекции и её динамики во времени. В работе предлагается тематическая модель, учитывающая метки времени документов, и использующая подход аддитивной регуляризации ARTM. Разрабатываются критерии устойчивости и полноты для оценки качества модели. Модель обучена на коллекции пресс-релизов внешнеполитических ведомств ряда стран за 10 лет.

Ключевые слова: *тематическая модель, аддитивная регуляризация, LDA, критерий устойчивости, критерий полноты.*

1. Введение

Данная работа посвящена методам анализа тематической структуры большой текстовой коллекции и её динамики во времени.

Тематическая модель коллекции текстовых документов разбивает коллекцию на некоторое количество тем и определяет, к каким темам относятся документы, а также какие слова образуют каждую тему. Эта задача решается с помощью *вероятностного тематического моделирования* – пользователем фиксируется число тем, после чего модель находит распределения $\varphi_{wt} = p(w|t)$ слов по темам и $\theta_{td} = p(t|d)$ тем по документам.

Классический подход к решению задачи тематического моделирования – это латентное размещение Дирихле (LDA), описанный в работе [6]. Этот метод предполагает, что плотности φ_{wt} и θ_{td} имеют распределение Дирихле, которое является в данном случае байесовским регуляризатором модели, который предотвращает переобучение. Но эта модель не даёт возможности для добавления требований различности тем или разреженности распределений φ_{wt} и θ_{td} и внесения других ограничений на модель.

Другим подходом, устраняющим эти ограничения, является аддитивная регуляризация тематических моделей (ARTM, [3]). Он позволяет записать любое количество дополнительных требований к тематической модели в виде взвешенной суммы критериев, добавляемых к основному функционалу логарифмированного правдоподобия. В [3] показано, что функционалы правдоподобия многих известных тематических моделей, таких как LDA и PLSA, допускают такое представление, то есть фактически являются частными случаями регуляризации. При этом, в отличие от стандартных задач машинного обучения, таких как классификация и регрессия, в тематическом моделировании возникает огромное разнообразие регуляризаторов, направленных на учёт различной дополнительной информации о текстовой коллекции.

Темпоральные тематические модели учитывают дополнительно метки времени y_d, привязанные к каждому документу d. Помимо распределений φ_{wt} и θ_{td}, вводится распределение каждой темы во времени $\xi_{yt} = p(y|t)$, что позволяет рассмотреть динамику изменения тем во времени.

Научный руководитель: Стрижов В. В. Задачу поставил: Воронцов К. В. Консультант: Дойков Н. В.

Одним способом [7], [8] анализа тем во времени является разбиение исходной коллекции документов на пачки относящихся к одному временному интервалу и построение отдельной тематической модели для каждой пачки, с последующим анализом тем.

Способы явного включения времени в вероятностную модель чаще всего основаны на байесовском подходе: желаемые особенности модели добавляются с помощью указания априорных распределений на параметры. Можно выделить два направления: использование непрерывного априорного распределения $p(y|t)$ времени для каждой темы и модели с дискретным временем, основанные на Марковском свойстве. Относящаяся к первому классу модель TOT (Topics Over Time), [5] расширяет модель LDA, задавая априорное бета-распределение времени для каждой темы. Минусом этой модели является то, что в реальной жизни темы могут быть распределены совсем иначе и хочется найти их истиное распределение.

В данной работе с помощью подхода ARTM предлагается темпоральная тематическая модель, обученная на коллекции пресс-релизов внешнеполитических ведомств ряда стран за 10 лет, предлагается метрика для отбора событийных тем в модели, а также проводится анализ устойчивости модели с метками времени. Коллекция пресс-релизов собрана с вебсайтов внешнеполитических ведомств.

2. Постановка задачи

Пусть D – конечный набор текстов, называемый коллекцией, а W – набор слов, из которых состоят тексты (словарь). Для каждого слова w известно, сколько раз оно встречается в данном документе d, обозначим эту частоту как n_{dw}. Длину документа обозначим n_d. Предположим, что появление каждого слова в каждом документе связано с некоторой латентной переменной из некоторого множества тем T.

На множестве $D \times W \times T$ введём вероятностное пространство с плотностью $p(d, w, t)$. Примем *гипотезу условной независимости* – будем считать, что вероятность появления слова w, относящегося к теме t в документе d не зависит от документа и описывается общим для всей коллекции распределением:

$$p(w|d,t) \;=\; p(w|t). \tag{1}$$

Используя формулу полной вероятности и данную гипотезу, получаем:

$$p(w|d) = \sum_{t\in T} p(t|d)p(w|t). \tag{2}$$

Параметрами модели являются условные вероятности $\varphi_{wt} \equiv p(w|t)$ и $\theta_{td} \equiv p(t|d)$. Известными данными в данной задаче является матрица $F = (\hat{p}(w|d))_{W\times D}$ частотных оценок вероятностей $p(w|d)$:

$$\hat{p}(w|d) = \frac{n_{dw}}{n_d}.$$

Построить *тематическую модель* коллекции D значит найти множество тем T и стохастические матрицы $\Phi = (\varphi_{wt})_{W\times T}$ и $\Theta = (\theta_{td})_{T\times D}$, столбцы которых — распределения слов по темам и тем по документам. Поиск этих матриц производится методом максимизации логарифма правдоподобия коллекции:

$$\log \mathscr{L}(\Phi,\Theta) = \sum_{d\in D}\sum_{w\in W} n_{dw} \log \sum_{t\in T} \varphi_{wt}\theta_{td} \longrightarrow \max_{\Phi,\Theta}. \tag{3}$$

При добавлении модальности времени вводится аналогичная гипотеза условной независимости:

$$p(y|d,t) \;=\; p(y|t), \tag{4}$$

где y – момент времени. По формуле полной вероятности выражаем распределение моментов времен по документам через смесь распределений:

$$p(y|d) = \sum_{t\in T} p(t|d)p(y|t). \tag{5}$$

Так как в задаче каждому документу d приписана метка времени $y_d \in Y$, есть эмпирическое распределение:

$$\hat{p}(y|d) = [y = y_d].$$

Аналогично ставя задачу оптимизации для матриц

$$\Xi = (\xi_{yt})_{Y\times T} \text{ и } \Theta = (\theta_{td})_{T\times D}$$

и взвешенно суммируя две функции правдоподобия, получаем общую задачу оптимизации для темпоральной тематической модели:

$$\mathscr{L}(\Phi,\Theta,\Xi) = \mathscr{L}_1(\Phi,\Theta) + \tau\mathscr{L}_2(\Theta,\Xi), \tag{6}$$

$$\log\mathscr{L}(\Phi,\Theta,\Xi) \longrightarrow \max_{\Phi,\Theta,\Xi} \tag{7}$$

Задача максимизации правдоподобия имеет бесконечно много локальных максимумов, что влечёт за собой неустойчивость модели.

В подходе ARTM [3] авторы предлагают в оптимизационной задаче (3) добавить к логарифму правдоподобия еще r функционалов: $R_i(\Phi,\Theta),\ i = 1,\dots,r$ называемых *регуляризаторами*, каждый со своим неотрицательным весом τ_i:

$$\begin{cases} R(\Phi,\Theta) = \sum_i \tau_i R_i(\Phi,\Theta), \qquad \log\mathscr{L}(\Phi,\Theta) + R(\Phi,\Theta) \longrightarrow \max\limits_{\Phi,\Theta}, \\ \varphi_{wt} \geqslant 0, \quad \theta_{td} \geqslant 0, \quad \sum\limits_w \varphi_{wt} = 1, \quad \sum\limits_t \theta_{td} = 1. \end{cases} \tag{8}$$

В данной работе используются следующие регуляризаторы из предложенных [3] авторами:

1. Регуляризатор сглаживания
2. Регуляризатор разреживания
3. Регуляризатор декорреляции

2. Метрики для оценки модели

От распределений φ_{wt} и θ_{td}, полученных в ходе построения тематической модели, требуется обладание многими полезными свойствами: разреженностью — большим числом нулей, отсутствием фоновых слов в предметных темах, различностью предметных тем друг от друга, плавностью изменения тем во времени, а главное — интерпретируемостью.

Будем следить за набором дополнительных метрик, позволяющих наблюдать за процессом сходимости и определять, обладают ли искомые распределения φ_{wt} и θ_{td} перечисленными свойствами.

1. **Перплексия** — величина, выражающаяся через правдоподобие выборки и позволяющая отслеживать сходимость метода оптимизации:

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

$$p(w|d) \;=\; \sum_{t\in T}\varphi_{wt}\theta_{td}, \quad n \;\equiv\; \sum_{d\in D}\sum_{w\in W} n_{dw}.$$

   Численное значение перплексии не имеет интерпретации и позволяет лишь сравнивать алгоритмы между собой. Значения чем меньше, тем лучше.
2. **Разреженность матриц Φ и Θ** — доля нулевых элементов.
   В предметных темах разреженность достигает $90-95\%$, поэтому, для хорошей тематической модели разреженность необходима.

Лексическим *ядром темы* будем называть множество слов, отличающих данную тему от остальных:

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

Параметр $\delta = 0.25$, подбирается с тем расчетом, чтобы *размер ядра* $|W_t|$ был от 20 до 200 слов.

На основе ядра темы строятся следующие две оценки:

3. **Чистота** — суммарная вероятность слов ядра:

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

   показывает насколько хорошо тема описывается своим ядром. Чем выше, тем лучше.
4. **Контрастность** — средняя вероятность встретить слова ядра в конкретной теме:

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

   При большой контрастности тема однозначно угадывается по своему ядру, при малой – тема размывается, становится нечеткой.

Для исследования моделей со временем, где каждому документу d привязана метка времени $y_d \in Y$ из множества временных отчетов, добавим еще три характеристики:

5. **Разреженность распределений $p(t|y)$** — доля нулевых элементов среди всех распределений тем во времени. Позволяет оценивать воздействие на модель регуляризатора *разреживания* $p(t|y)$.
6. **Колебание темы во времени:**

$$\text{Variation}(t) \;=\; \sum_{y \in Y} \Big|\, \sqrt{p(y|t)} \;-\; \sqrt{p(y-1|t)} \,\Big|.$$

   Меньшие значения соответствуют более плавному изменению темы во времени.
7. **Событийность темы.**
   Большое число тем для потоков текстовых документов можно разбить на два класса: постоянные темы и темы-события.
   Постоянные темы присутствуют на протежении всего промежутка времени, распределение $p(y|t)$ для такой темы близко к равномерному.
   Тема-событие характеризуется внезапным появлением и постепенным затуханием во времени, её распределение $p(y|t)$ обладает большим числом нулей.

   В данной работе предлагается ряд метрик, которыми можно мерить событийность темы.

Меры событийности тем

Будем считать, что множество Y имеет вид отрезка $[0; M]$.

— **Доля нулей слева и справа.** Для $p(y|t)$ найдем $y_l = \min\{y \in Y \mid p(y|t) > \varepsilon\}$ и $y_r = \max\{y \in Y \mid p(y|t) > \varepsilon\}$. Введем меру как суммарная доля нулей распределения слева и справа. Константа ε определяет какие значения в $p(y|t)$ мы полагаем равными нулю и служит для устранения шума.

$$\text{ZerosLeftRight}(t) = 1 - \frac{y_r - y_l}{M};$$

Чем метрика больше, тем тема событийнее.

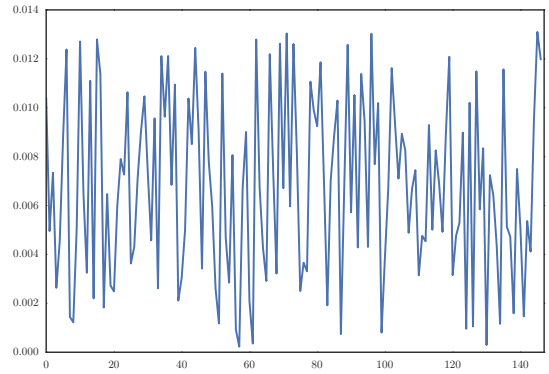
— **Дисперсия** $p(y|t)$**.** Чем дисперсия меньше, тем тема событийнее.

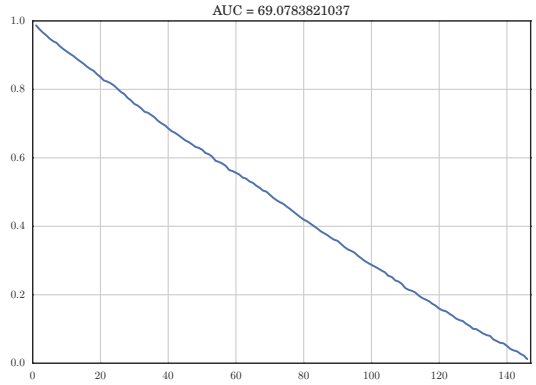
— **Delta-AUC.** Для всех $0 < \Delta \leqslant M$ найдем $0 \leqslant y_0 \leqslant M - \Delta$ такой, что интеграл

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

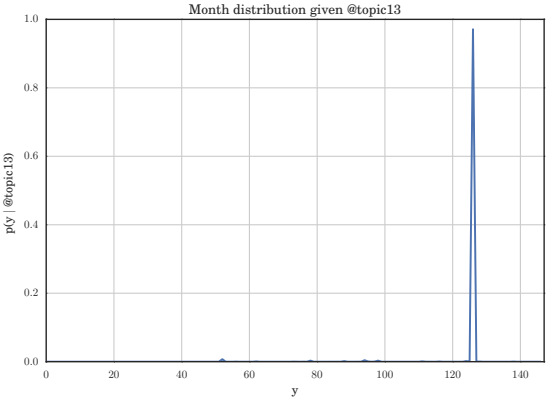
максимален. Построим график зависимости $1 - S$ от Δ. Для равномерного распределения он будет иметь вид убывающей прямой, а для событийных тем он будет сначала резко убывать вниз, затем плавно. Назовем Delta-AUC метрикой площадь под этим графиком. Можно понять, что для событийных тем она будет меньше. На Рис. 1 изображены примеры распределений и Delta-ROC кривых для них.

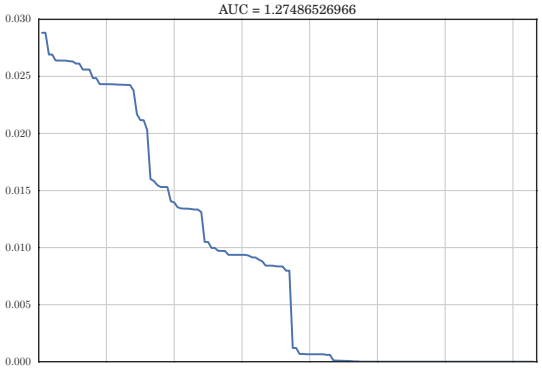
В метрике ZerosLeftiRight есть недостатки. Можно придумать два распределения, равномерное и событийное, для которых доли нулей слева и справа будут совпадать. Аналогично, можно придумать пример для дисперсии.



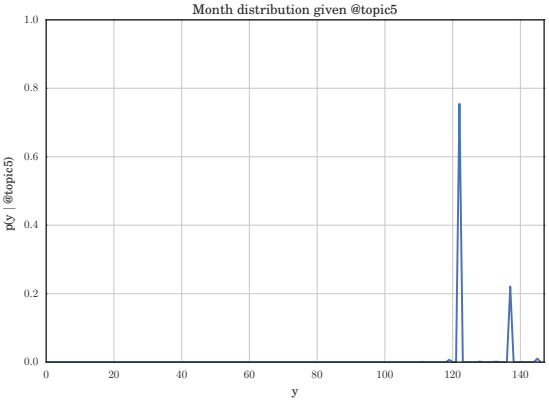


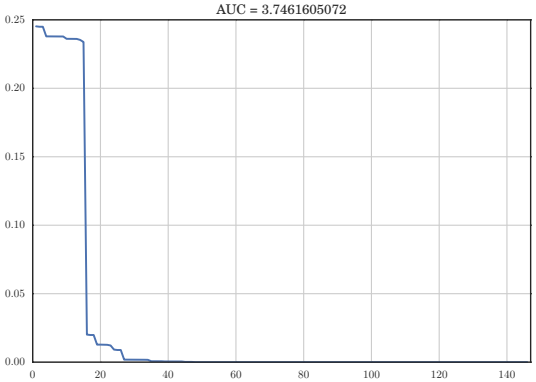
(a) Равномерное распределение





(b) Распределение с одним пиком





(c) Распределение с двумя пиками

Рис. 1. Примеры распределений и Delta-ROC кривых для них.

3. Базовый эксперимент

В рамках первичного эксперимента была построена тематическая модель без модальности времени с регуляризатором сглаживания. Такая регуляризация является аналогом известной тематической модели LDA [6].

Данными является коллекция официальных пресс-релизов внешнеполитических ведомств ряда стран, на английском языке. Более 20 тыс. сообщений за 10 лет.

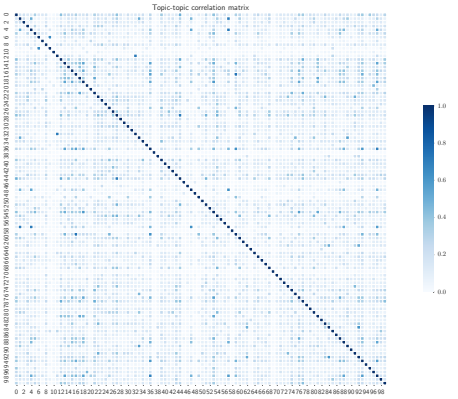
Недостатки модели LDA в том, что темы часто получаются скоррелированными между собой и содержат слова из общей лексики. На Рис. 2 показаны основные метрики модели. Для построения использовалась библиотека BigARTM. Количество тем 100. На все темы применялся регуляризатор сглаживания. Разреженность матрицы Φ – 0.0%, матрицы Θ – 0.4%. Средняя контрастность ядра по темам – 0.410. Средняя чистота ядра – 0.114. Средняя корреляция между темами 0.109.

В табл. 1 представлены примеры тем для модели.

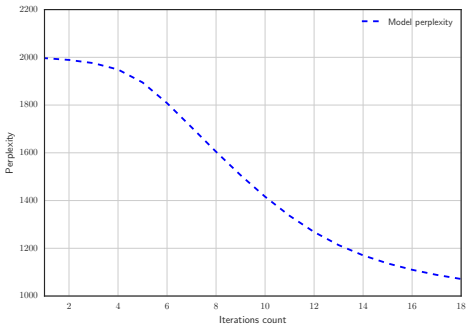
Ключевые слова темы
state, foreign, secretary, president, relationship, very, unite, security, meet, thank, issue, together, minister, here, today
question, department, state, information, release, site, office, view, subject, u.s., internet, email, answer, page, should
state, question, designate, missile, under, russia, designation, act, order, entity, sanction, department, europe, decision, council
woman, society, support, opportunity, world, more, tunisia, violence, gender, civil, leader, community, business, international, help
question, inaudible, mean, talk, don, thing, more, secretary, kind, term, very, ask, try, actually, obviously
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

Таблица 1. Ключевые слова тем в модели LDA

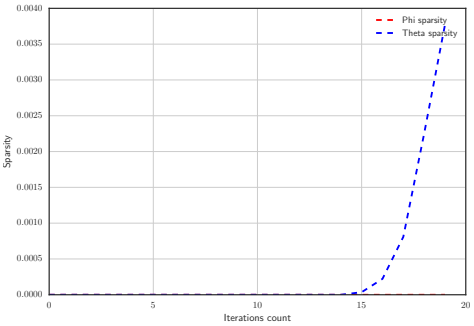
Видно, что в темах содержится много слов, которые не характеризуют ядро темы, а являются фоновой лексикой корпуса.



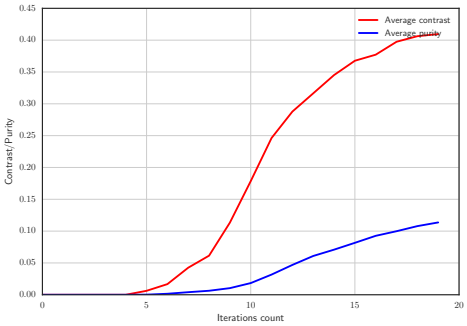
(a) Матрица корреляций между темами



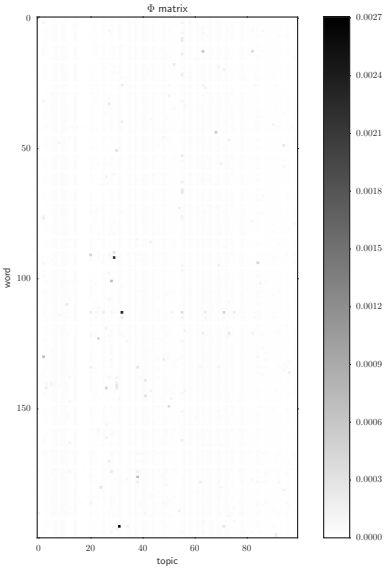
(b) График перплексии модели по итерациям



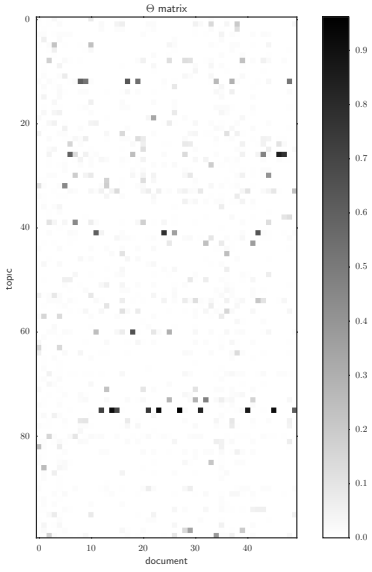
(c) График разреженности матриц Φ и Θ по итерациям



(d) График чистоты и контрастности модели по итерациям



(e) Срез матрицы Φ для первых 200 слов



(f) Срез матрицы Θ для первых 50 документов

Рис. 2. Графики для модели с регуляризатором сглаживания фоновых тем

4. Эксперимент с модальностью времени

В рамках основного эксперимента к предыдущей модели были добавлены регуляризаторы декорреляции, разреживания матриц Φ и Θ, а также были учтены метки времени документов. Меткой времени в данном случае являлся месяц публикации статьи. Все месяцы были пронумерованы числами от 0 до 146.

На Рис. 5 приведены аналогичные Рис. 2 графики. В новой модели матрицы стали разреженнее, темы более интепретируемы и декоррелированы. Разреженность матрицы Φ достигла 85.3%, матрицы Θ – 73.8%. Средняя контрастность ядра по темам – 0.540. Средняя чистота ядра – 0.183. Средняя корреляция между темами 0.151.

В таблице 2 приведены примеры наиболее событийных тем с точки зрения метрики Delta-AUC по 15 топ-слов в распределении.

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
applause, woman, thank, family, child, life, here, help, many, more, world, today, every, honor, service	0.0170
egypt, egyptian, reform, democratic, more, secretary, election, process, support, president, security, need, those, term, very	0.0221
foreign, state, very, relationship, president, secretary, unite, thank, minister, meet, issue, security, much, together, discuss	0.0247
question, thank, next, issue, operator, line, open, meet, term, process, again, both, please, comment, country	0.0287
north, korea, korean, question, program, weapon, president, state, south, international, rice, september, talk, intelligence, security	0.0341
secretary, assistant, deputy, state, issue, affair, negroponte, hill, john, ambassador, rice, october, question, armitage, condoleezza	0.0641

Таблица 2. Примеры событийных тем и значения Delta-AUC для них

На рис. 3 показаны распределения для некоторых событийных тем.

В табл. 3 приведены примеры не событийных тем с точки зрения метрики Delta-AUC по 15 топ-слов в распределении.

На рис. 4 показаны распределения некоторых несобытийных тем.

На Рис. 6 показана матрица Ξ, столбцы которой отсортированы по введённой выше метрике событийности Delta-AUC. Видно, что в левой части матрицы, где Delta-AUC меньше, распределения столбцов сконцентрированы вокруг нескольких точек. В правой же части матрицы распределения размазаны. Это соответствует нашему представлению о событийности темы. Разреженность матрицы достигла 79.9%.

Ключевые слова темы	Delta-AUC
cote, ricewell, hungary, diamond, brimmer, hungarian, ricei, questionand, kimberley, tribe, allen, esther, questioni, questionmadame, gbagbo	0.6810
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, process, service, must	0.7321
support, president, unite, commission, u.s., national, election, january, country, include, state, council, group, december, nation	0.7588
health, medical, care, treatment, cancer, mali, hiv, prevention, center, los, pepfar, child, drug, expo, pavilion	0.8122
press, state, department, medium, brief, identification, u.s., secretary, a.m., card, event, photo, p.m., issue, street	0.8314
award, corporate, anniversary, embassy, winner, ceremony, excellence, compound, company, present, employee, whale, outstanding, honor, build	0.8315

Таблица 3. Примеры не событийных тем и значения Delta-AUC для них

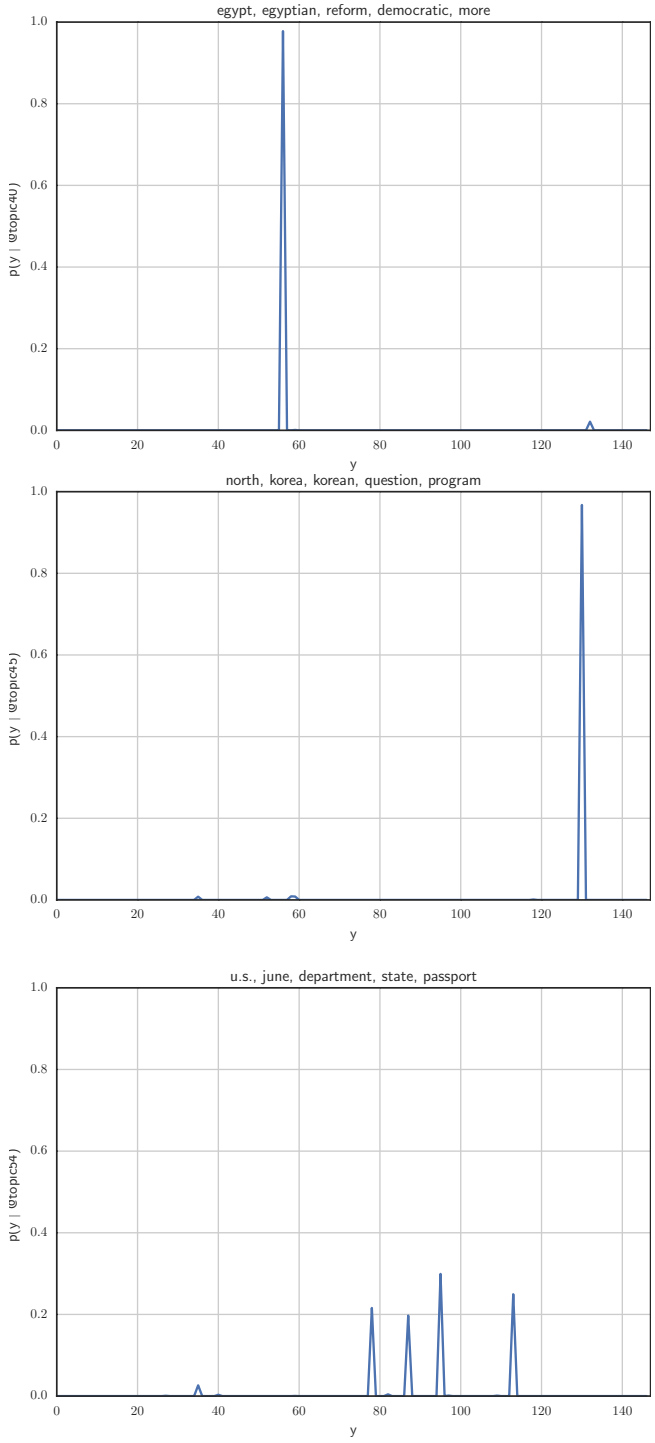


Рис. 3. Распределения некоторых событийных тем

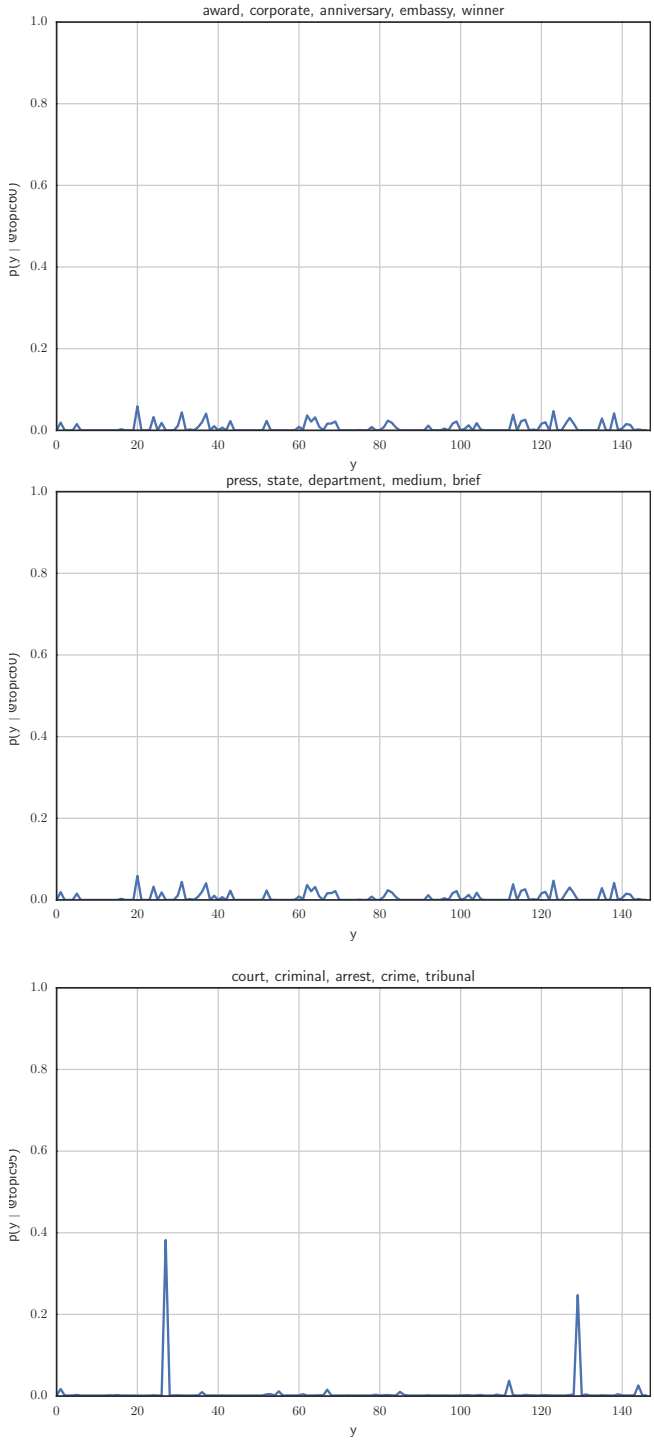
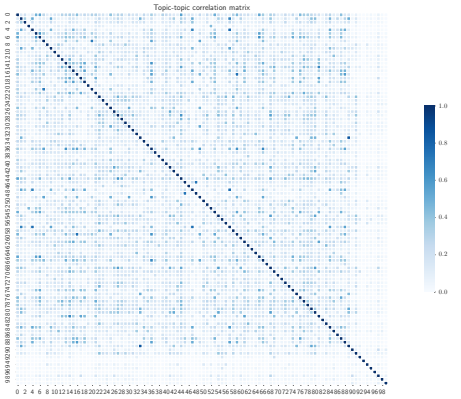
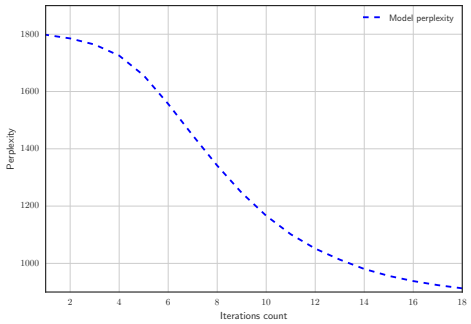


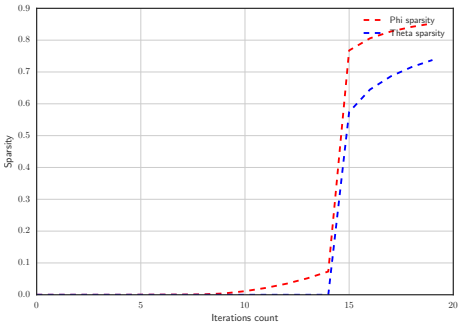
Рис. 4. Распределения некоторых событийных тем



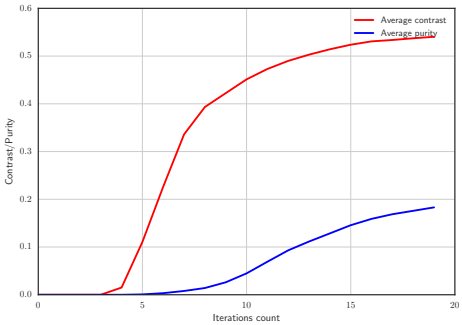
(a) Матрица корреляций между темами



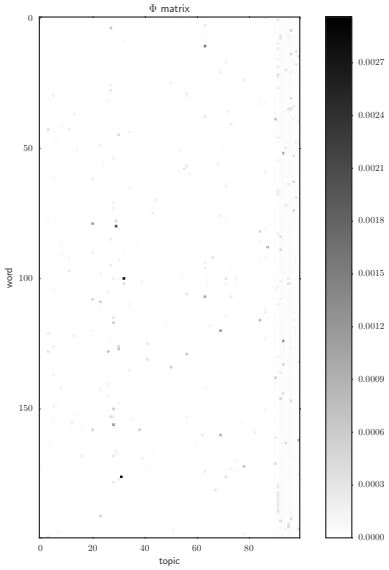
(b) График перплексии модели по итерациям



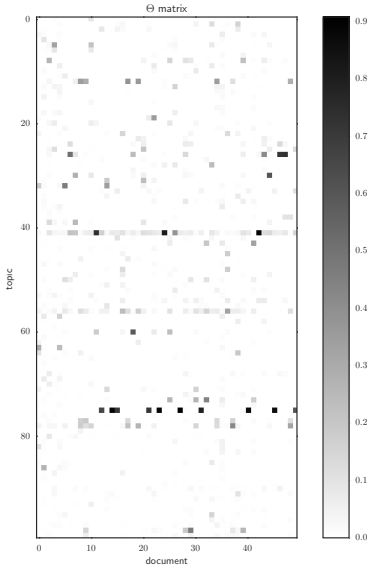
(c) График разреженности матриц Φ и Θ по итерациям



(d) График чистоты и контрастности модели по итерациям



(e) Срез матрицы Φ для первых 200 слов



(f) Срез матрицы Θ для первых 50 документов

Рис. 5. Графики для модели с модальностью времени

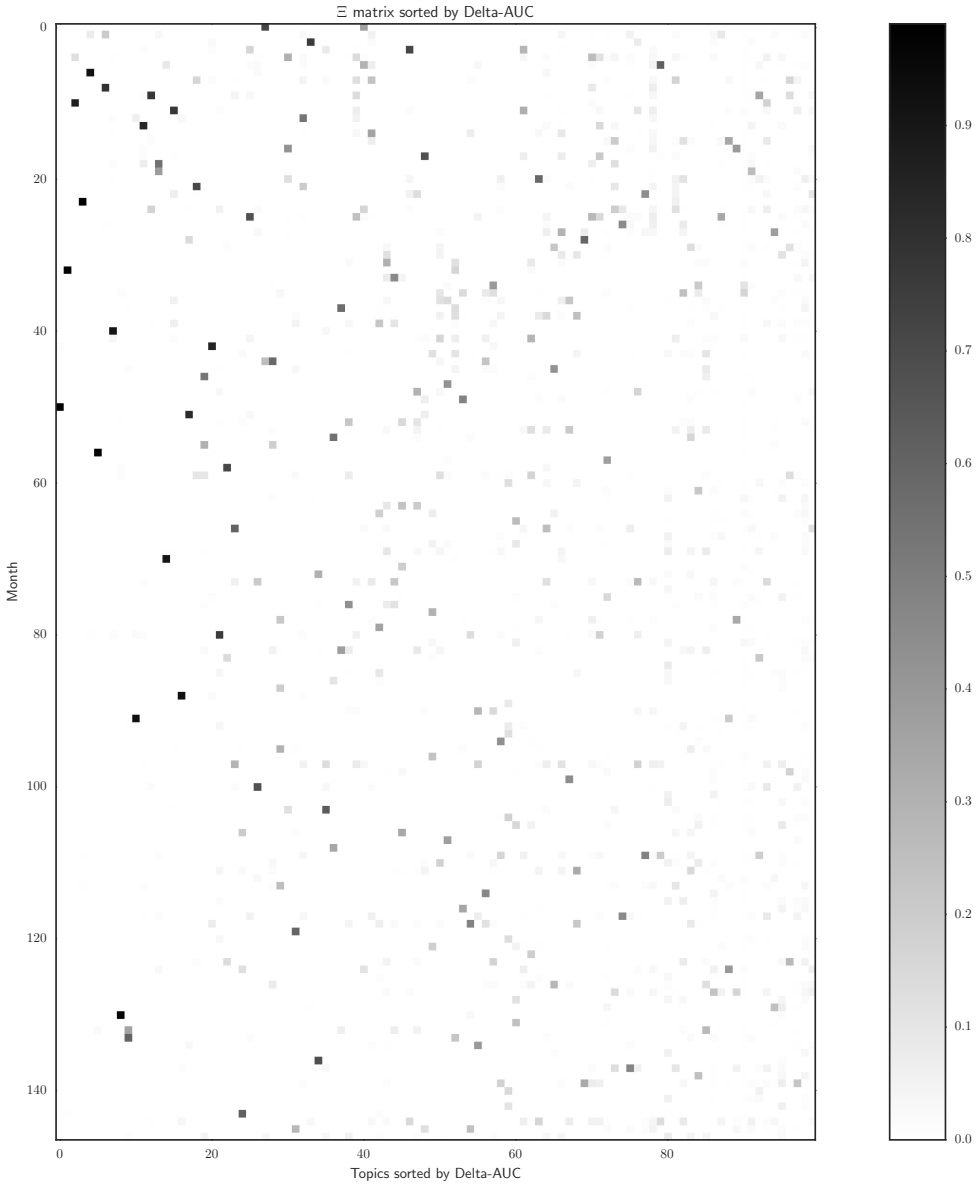
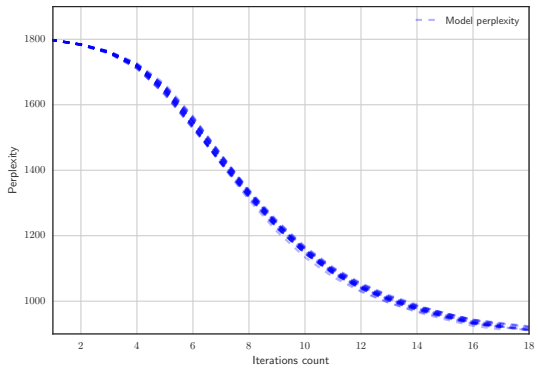


Рис. 6. Матрица Ξ

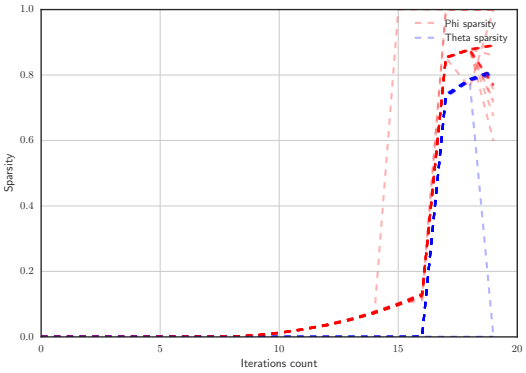
5. Эксперимент с устойчивостью модели

Модель со временем была обучена на 25 различных начальных приближениях матриц Θ, Φ, Ξ с целью понять, какие свойства модели устойчивы, а какие нет.

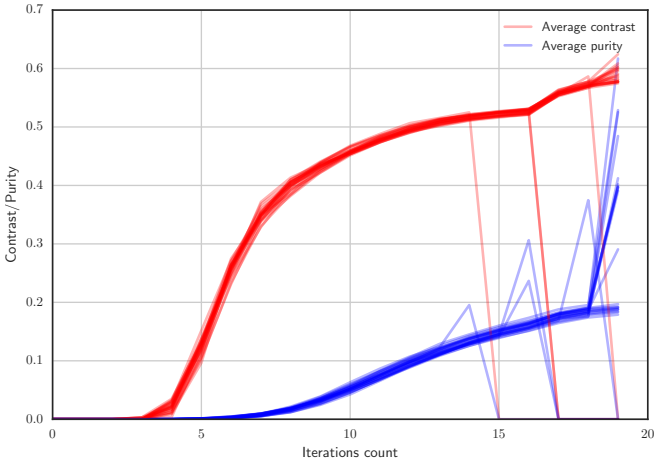
Модель из второго эксперимента оказалась неустойчивой. Из-за больших по модулю коэффициентов регуляризации при регуляризаторах ($\tau_\Phi = \tau_\Theta = -0.2$) разреживания матриц Φ и Θ в большинстве случаев матрицы становились вырожденными. Уменьшив регуляризаторы до значения -0.1, и включая их только на три последние итерации алгоритма, удалось получить более устойчивую модель, для которой на Рис. 7 приведены результаты.



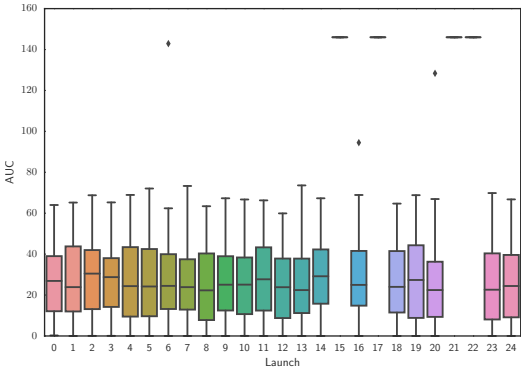
(a) График перплексии модели по итерациям



(b) График разреженности матриц Φ и Θ по итерациям



(c) График чистоты и контрастности модели по итерациям



(d) Боксплот распределений метрики Delta-AUC в зависимости от номера запуска

Рис. 7. Графики для эксперимента с устойчивостью модели с регуляризатором сглаживания фоновых тем

По-прежнему в нескольких случаях модель выродилась, но в остальных случаях метрики стабильны.

Была исследована способность модели находить одни и те же темы при разных начальных приближениях матриц.

Опишем сначала алгоритм для двух матриц Φ_1 и Φ_2. Обозначим $\Phi(:, i)$ – распределение i-ой темы в матрице Φ. Введём матрицу $T = (t_{ij})$ расстояний между распределенями.

Конкретно, $t_{ij} = H\big(\Phi_1(:,i), \Phi_2(:,j)\big)$, где $H(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{k}(\sqrt{p_i} - \sqrt{q_i})^2}$ – расстояние Хеллингера между дискретными распределениями.

Применим Венгерский алгоритм для матрицы T. Венгерский алгоритм решает задачу о назначениях. В нашем случае, алгоритм попытается найти каждую тему в другой модели, то есть сгенерировать перестановку на матрице Φ_2.

Для запусков из различных начальных приближений будем считать матрицу Φ_0, получившуюся на первой итерации, эталонной. Далее, для остальных итераций применим Венгерский алгоритм к Φ_0 и Φ_i и посчитаем среднюю стоимость – среднее расстояние Хеллингера для перестановки тем.

В четырех случаях это среднее было равно 0.707, так как матрицы выродились. В остальных случаях среднее значение колебалось около 0.038 со стандартным отклонением 0.001.

В Таблице 4 показаны примеры сопоставленных друг другу тем. Оказалось, что взяв расстояние Хеллингера и применив Венгерский алгоритм, не удается получить адекватное сопоставление тем друг другу. Однако устойчивость модели к восстановлению одних и тех же матриц при разных начальных приближениях проверить удалось.

Ключевые слова тем в матрице Φ_0	Ключевые слова тем в матрице Φ_0	Расстояние Хеллингера
musical, customer, ensue, walt, nongovernment, release, witness, ban	unite, nation, people, state, world, global, international, national	0.036
musical, customer, ensue, walt, nongovernment, release, witness, ban	country, development, help, partner, support, sector, need, more	0.028
customer, walt, release, witness, ban, consolidate, representative, solve	assistant, secretary, state, government, meet, unite, liberia, congo	0.104
musical, customer, ensue, walt, nongovernment, release, witness, ban	attack, state, unite, terrorist, terrorism, government, release, information	0.044
musical, customer, walt, release, witness, ban, consolidate, representative	defense, missile, nato, question, poland, threat, unite, state	0.028

Таблица 4. Сопоставленные друг другу темы для матриц Φ_0 и Φ_1

На Рис. 8 показаны распределения метрики Delta-AUC для запусков из разных начальных приближений.

5. Заключение

Произведено сравнение моделей LDA и ARTM с модальностью времени. Показано, что в модели ARTM темы получаются более интепретируемые, фоновая лексика уходит в фоновые темы. Матрицы Φ и Θ разреженнее.

Предложена метрика для отбора событийных тем Delta-AUC, основной результат которой можно видеть на Рис. 6.

Был поставлен эксперимент с устойчивостью модели со временем.

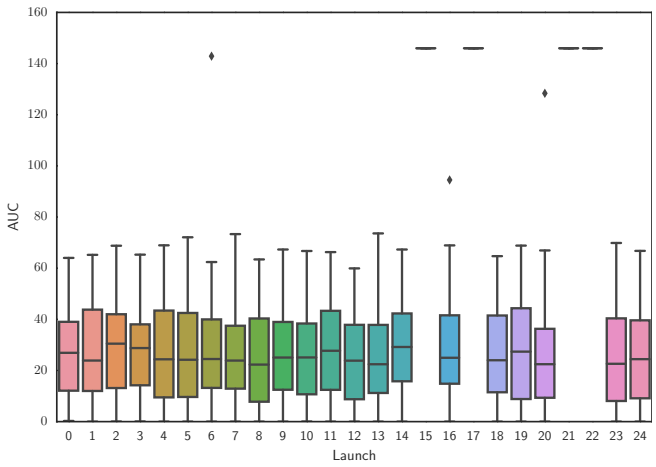


Рис. 8. Распределения Delta-AUC значений по топикам для разных начальных приближений

Литература

[1] *Дойков Н. В.* Адаптивная регуляризация вероятностных тематических моделей. 2015 http://www.machinelearning.ru/wiki/images/9/9f/2015_417_DoykovNV.pdf

[2] *Воронцов К. В.* Вероятностное тематическое моделирование. Москва, 2009. http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf

[3] *Воронцов К. В.* Аддитивная Регуляризация Тематических Моделей Коллекций Текстовых Документов *Доклады РАН, Т.455, №3. С.268-271* 2014

[4] *David Hall, Daniel Jurafsky, Christopher D. Manning.* Studying the History of Ideas Using Topic Models. https://web.stanford.edu/~jurafsky/hallemnlp08.pdf

[5] *Xuerui Wang, Andrew McCallum* Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends https://people.cs.umass.edu/~mccallum/papers/tot-kdd06.pdf

[6] *Blei, David M and Ng, Andrew Y and Jordan, Michael I* Latent dirichlet allocation *the Journal of machine Learning research* 2003

[7] *Griffiths T. L., Steyvers M.* Finding scientific topics *Proceedings of the National Academy of Sciences.* 2004

[8] *Hall D., Jurafsky D., Manning C. D.* Studying the history of ideas using topic models *Proceedings of the conference on empirical methods in natural language processing / Association for Computational Linguistics.* 2008