

Тематическое моделирование

К. В. Воронцов
vokov@forecsys.ru

Этот курс доступен на странице вики-ресурса
<http://www.MachineLearning.ru/wiki>
«Машинное обучение (курс лекций, К.В.Воронцов)»

ШАД Яндекс • 29 октября 2019

- 1 Вероятностное тематическое моделирование**
 - Цели, приложения, постановка задачи
 - Аддитивная регуляризация тематических моделей
 - Классические модели: PLSA и LDA
- 2 Регуляризация тематических моделей**
 - Мультимодальные тематические модели
 - Классификация и регрессия на текстах
 - Предметные и фоновые темы. Иерархии тем
- 3 Оценивание качества и визуализация**
 - Внутренние (intrinsic) критерии качества
 - Внешние (extrinsic) критерии качества
 - Визуализация тематических моделей

Что такое «тема» в коллекции текстовых документов?

- *тема* — специальная терминология предметной области
- *тема* — набор часто совместно встречающихся терминов
- *тема* — семантически однородный кластер текстов

Более формально,

- *тема* — условное распределение на множестве терминов,
 $p(w|t)$ — вероятность термина w в теме t ;
- *тематический профиль* документа — условное распределение
 $p(t|d)$ — вероятность темы t в документе d .

Когда автор писал термин w в документе d , он думал о теме t , и мы хотели бы выявить, о какой именно.

Тематическая модель выявляет латентные темы по наблюдаемым распределениям слов $p(w|d)$ в документах.

Пример. Мультиязычная модель Википедии

216 175 русско-английских пар статей.

Первые 10 слов и их вероятности $p(w|t)$ в %:

| Тема №68 | | | | Тема №79 | | | |
|-------------|------|--------------|------|----------|------|-----------|------|
| research | 4.56 | институт | 6.03 | goals | 4.48 | матч | 6.02 |
| technology | 3.14 | университет | 3.35 | league | 3.99 | игрок | 5.56 |
| engineering | 2.63 | программа | 3.17 | club | 3.76 | сборная | 4.51 |
| institute | 2.37 | учебный | 2.75 | season | 3.49 | фк | 3.25 |
| science | 1.97 | технический | 2.70 | scored | 2.72 | против | 3.20 |
| program | 1.60 | технология | 2.30 | cup | 2.57 | клуб | 3.14 |
| education | 1.44 | научный | 1.76 | goal | 2.48 | футболист | 2.67 |
| campus | 1.43 | исследование | 1.67 | apps | 1.74 | гол | 2.65 |
| management | 1.38 | наука | 1.64 | debut | 1.69 | забивать | 2.53 |
| programs | 1.36 | образование | 1.47 | match | 1.67 | команда | 2.14 |

Ассессор оценил 396 тем из 400 как хорошо интерпретируемые.

Vorontsov, Frei, Apishev, Romov, Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Пример. Мультиязычная модель Википедии

216 175 русско-английских пар статей.

Первые 10 слов и их вероятности $p(w|t)$ в %:

| Тема №88 | | | | Тема №251 | | | |
|-------------|------|---------|------|------------|------|--------------|------|
| opera | 7.36 | опера | 7.82 | windows | 8.00 | windows | 6.05 |
| conductor | 1.69 | оперный | 3.13 | microsoft | 4.03 | microsoft | 3.76 |
| orchestra | 1.14 | дирижер | 2.82 | server | 2.93 | версия | 1.86 |
| wagner | 0.97 | певец | 1.65 | software | 1.38 | приложение | 1.86 |
| soprano | 0.78 | певица | 1.51 | user | 1.03 | сервер | 1.63 |
| performance | 0.78 | театр | 1.14 | security | 0.92 | server | 1.54 |
| mozart | 0.74 | партия | 1.05 | mitchell | 0.82 | программный | 1.08 |
| sang | 0.70 | сопрано | 0.97 | oracle | 0.82 | пользователь | 1.04 |
| singing | 0.69 | вагнер | 0.90 | enterprise | 0.78 | обеспечение | 1.02 |
| operas | 0.68 | оркестр | 0.82 | users | 0.78 | система | 0.96 |

Ассессор оценил 396 тем из 400 как хорошо интерпретируемые.

Vorontsov, Frei, Apishev, Romov, Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Некоторые приложения тематического моделирования

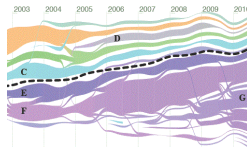
разведочный поиск в
электронных библиотеках



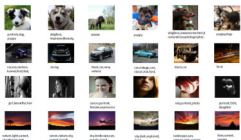
поиск тематического
контента в соцсетях



детектирование и трекинг
новостных сюжетов



мультимодальный поиск
текстов и изображений



анализ банковских
транзакционных данных



управление диалогом в
разговорном интеллекте



Пусть

- W — конечное множество слов (терминов, токенов)
- D — конечное множество текстовых документов
- T — конечное множество тем
- каждое слово w в документе d связано с некоторой темой t
- $D \times W \times T$ — дискретное вероятностное пространство
- **порядок слов в документе не важен (bag of words)**
- порядок документов в коллекции не важен
- коллекция — это i.i.d. выборка $(d_i, w_i, t_i)_{i=1}^n \sim p(d, w, t)$
- d_i, w_i — наблюдаемые, темы t_i — скрытые
- **гипотеза условной независимости: $p(w|d, t) = p(w|t)$**

Тематическая модель, по формуле полной вероятности:

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

Прямая задача — порождение коллекции по $p(w|t)$ и $p(t|d)$

Вероятностная тематическая модель описывает каждый документ $p(w|d)$ как вероятностную смесь тем $p(w|t)$:

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



w_1, \dots, w_{n_d} :

Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

Обратная задача — восстановление $p(w|t)$ и $p(t|d)$ по коллекции

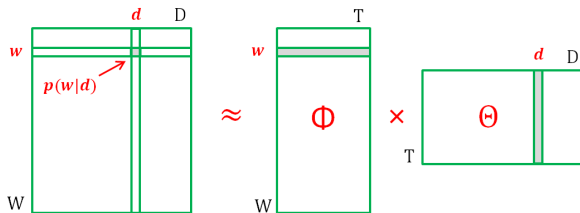
Дано: коллекция текстовых документов

- n_{dw} — частоты терминов в документах, $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$

Найти: параметры тематической модели $p(w|d) = \sum_{t \in T} \phi_{wt} \theta_{td}$

- $\phi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t
- $\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Это задача стохастического матричного разложения:



Принцип максимума правдоподобия

Правдоподобие — плотность распределения выборки $(d_i, w_i)_{i=1}^n$:

$$\prod_{i=1}^n p(d_i, w_i) = \prod_{d \in D} \prod_{w \in d} p(d, w)^{n_{dw}}$$

Максимизация логарифма правдоподобия

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d) \xrightarrow[\text{const}]{p(d)} \max_{\Phi, \Theta}$$

приводит к задаче математического программирования:

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

при ограничениях неотрицательности и нормировки

$$\phi_{wt} \geq 0; \quad \sum_{w \in W} \phi_{wt} = 1; \quad \theta_{td} \geq 0; \quad \sum_{t \in T} \theta_{td} = 1.$$

Задачи, некорректно поставленные по Адамару

Задача *корректно поставлена*,
если её решение

- существует,
- единственно,
- устойчиво.



Жак Саломон Адамар
(1865–1963)

Наша задача матричного разложения *некорректно поставлена*:
если Φ, Θ — решение, то стохастические Φ', Θ' — тоже решения

- $\Phi'\Theta' = (\Phi S)(S^{-1}\Theta)$, $\text{rank } S = |T|$
- $\mathcal{L}(\Phi', \Theta') = \mathcal{L}(\Phi, \Theta)$
- $\mathcal{L}(\Phi', \Theta') \leq \mathcal{L}(\Phi, \Theta) + \varepsilon$ — приближённые решения

Регуляризация — стандартный приём доопределения решения
с помощью дополнительных критериев.

ARTM: аддитивная регуляризация тематических моделей

Максимизация логарифма правдоподобия с регуляризатором:

$$\sum_{d,w} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}; \quad R(\Phi, \Theta) = \sum_i \tau_i R_i(\Phi, \Theta)$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdw} \equiv p(t|d, w) = \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td}) \\ \text{M-шаг:} & \begin{cases} \phi_{wt} = \operatorname{norm}_{w \in W} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right), & n_{wt} = \sum_{d \in D} n_{dw} p_{tdw} \\ \theta_{td} = \operatorname{norm}_{t \in T} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right), & n_{td} = \sum_{w \in W} n_{dw} p_{tdw} \end{cases} \end{cases}$$

где $\operatorname{norm}_{t \in T}(x_t) = \frac{\max\{x_t, 0\}}{\sum_{s \in T} \max\{x_s, 0\}}$ — операция нормировки вектора.

Условия вырожденности модели для тем и документов

Решение может быть вырожденным для некоторых тем (столбцов матриц Φ) и документов (столбцов матрицы Θ).

Тема t вырождена, если для всех терминов $w \in W$

$$n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \leq 0.$$

Если тема t вырождена, то $p(w|t) = \phi_{wt} \equiv 0$; это означает, что тема исключается из модели (происходит отбор тем).

Документ d вырожден, если для всех тем $t \in T$

$$n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \leq 0.$$

Если документ d вырожден, то $p(t|d) = \theta_{td} \equiv 0$; это означает, что модель не в состоянии описать данный документ.

Напоминания. Условия Каруша–Куна–Таккера

Задача математического программирования:

$$\begin{cases} f(x) \rightarrow \min_x; \\ g_i(x) \leq 0, & i = 1, \dots, m; \\ h_j(x) = 0, & j = 1, \dots, k. \end{cases}$$

Необходимые условия. Если x — точка локального минимума, то существуют множители $\mu_i, i = 1, \dots, m, \lambda_j, j = 1, \dots, k$:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial x} = 0, & \mathcal{L}(x; \mu, \lambda) = f(x) + \sum_{i=1}^m \mu_i g_i(x) + \sum_{j=1}^k \lambda_j h_j(x); \\ g_i(x) \leq 0; h_j(x) = 0; & \text{(исходные ограничения)} \\ \mu_i \geq 0; & \text{(двойственные ограничения)} \\ \mu_i g_i(x) = 0; & \text{(условие дополняющей нежёсткости)} \end{cases}$$

Вывод системы уравнений из условий Каруша–Куна–Таккера

1. Условия ККТ для ϕ_{wt} (для θ_{td} всё аналогично):

$$\sum_d n_{dw} \frac{\theta_{td}}{p(w|d)} + \frac{\partial R}{\partial \phi_{wt}} = \lambda_t - \mu_{wt}; \quad \mu_{wt} \geq 0; \quad \mu_{wt} \phi_{wt} = 0.$$

2. Умножим обе части равенства на ϕ_{wt} и выделим p_{tdw} :

$$\phi_{wt} \lambda_t = \sum_d n_{dw} \frac{\phi_{wt} \theta_{td}}{p(w|d)} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} = n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}}.$$

3. Если $\lambda_t \leq 0$, то тема t вырождена, $\phi_{wt} \equiv 0$ для всех w .

4. Если $\lambda_t > 0$, то либо $\phi_{wt} = 0$, либо $n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} > 0$:

$$\phi_{wt} \lambda_t = \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)_+.$$

5. Суммируем обе части равенства по $w \in W$:

$$\lambda_t = \sum_{w \in W} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)_+.$$

6. Подставим λ_t из (5) в (4), получим требуемое. ■

Элементарная интерпретация EM-алгоритма

EM-алгоритм — это чередование E и M шагов до сходимости.

E-шаг: условные вероятности тем $p(t|d, w)$ для всех t, d, w вычисляются через ϕ_{wt}, θ_{td} по формуле Байеса:

$$p(t|d, w) = \frac{p(w, t|d)}{p(w|d)} = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}.$$

M-шаг: при $R = 0$ частотные оценки условных вероятностей вычисляются суммированием счётчика $n_{tdw} = n_{dw}p(t|d, w)$:

$$\begin{aligned} \phi_{wt} &= \frac{n_{wt}}{n_t}, & n_{wt} &= \sum_{d \in D} n_{tdw}, & n_t &= \sum_{w \in W} n_{wt}; \\ \theta_{td} &= \frac{n_{td}}{n_d}, & n_{td} &= \sum_{w \in D} n_{tdw}, & n_d &= \sum_{t \in T} n_{td}. \end{aligned}$$

Два частных случая: модели PLSA и LDA

PLSA: probabilistic latent semantic analysis [Hofmann, 1999]
(вероятностный латентный семантический анализ):

$$R(\Phi, \Theta) = 0.$$

M-шаг — частотные оценки условных вероятностей:

$$\phi_{wt} = \underset{w}{\text{norm}}(n_{wt}), \quad \theta_{td} = \underset{t}{\text{norm}}(n_{td}).$$

LDA: latent Dirichlet allocation (латентное размещение Дирихле):

$$R(\Phi, \Theta) = \sum_{t,w} (\beta_w - 1) \ln \phi_{wt} + \sum_{d,t} (\alpha_t - 1) \ln \theta_{td}.$$

M-шаг — сглаженные частотные оценки с параметрами β_w, α_t :

$$\phi_{wt} = \underset{w}{\text{norm}}(n_{wt} + \beta_w - 1), \quad \theta_{td} = \underset{t}{\text{norm}}(n_{td} + \alpha_t - 1).$$

Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999.

Blei D., Ng A., Jordan M. Latent Dirichlet allocation. 2003.

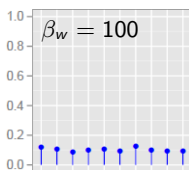
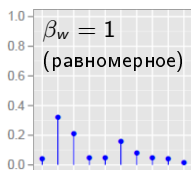
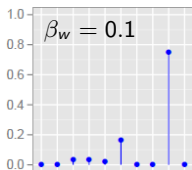
Распределение Дирихле

Гипотеза. Вектор-столбцы $\phi_t = (\phi_{wt})$ и $\theta_d = (\theta_{td})$ порождаются распределениями Дирихле, $\alpha \in \mathbb{R}^{|T|}$, $\beta \in \mathbb{R}^{|W|}$:

$$\text{Dir}(\phi_t | \beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1}, \quad \phi_{wt} > 0; \quad \beta_0 = \sum_w \beta_w, \quad \beta_t > 0;$$

$$\text{Dir}(\theta_d | \alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \quad \theta_{td} > 0; \quad \alpha_0 = \sum_t \alpha_t, \quad \alpha_t > 0;$$

Пример. Распределение $\text{Dir}(\phi | \beta)$ при $\phi, \beta \in \mathbb{R}^{10}$:



Максимизация апостериорной вероятности для модели LDA

Совместное правдоподобие данных и модели:

$$\ln \prod_{d \in D} \prod_{w \in d} p(w, d | \Phi, \Theta)^{n_{dw}} \prod_{t \in T} \text{Dir}(\phi_t | \beta) \prod_{d \in D} \text{Dir}(\theta_d | \alpha) \rightarrow \max_{\Phi, \Theta}$$

Регуляризатор — логарифм априорного распределения:

$$R(\Phi, \Theta) = \sum_{t, w} (\beta_w - 1) \ln \phi_{wt} + \sum_{d, t} (\alpha_t - 1) \ln \theta_{td}$$

M-шаг — сглаженные или разреженные частотные оценки:

$$\phi_{wt} = \text{norm}_w(n_{wt} + \beta_w - 1), \quad \theta_{td} = \text{norm}_t(n_{td} + \alpha_t - 1).$$

при $\beta_w > 1$, $\alpha_t > 1$ — сглаживание,

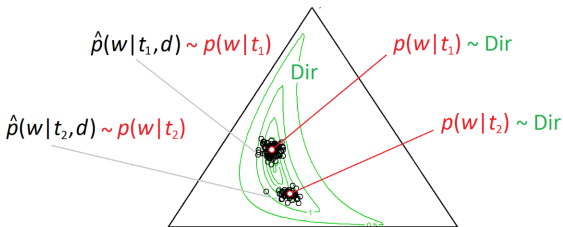
при $0 < \beta_w < 1$, $0 < \alpha_t < 1$ — слабое разреживание,

при $\beta_w = 1$, $\alpha_t = 1$ априорное распределение равномерно, PLSA.

Почему именно распределение Дирихле?

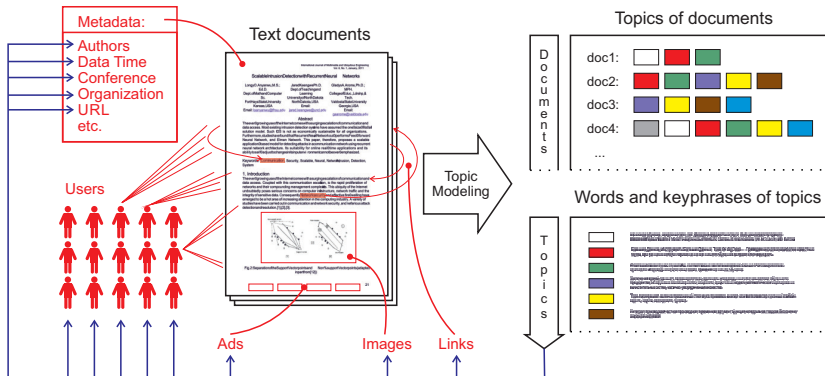
- оно способно порождать разреженные векторы,
- имеет параметры, управляющие степенью разреженности,
- описывает кластерные структуры на симплексе (см. рис.),
- математически удобно для байесовского вывода

Распределение $\text{Dir}(\phi|\alpha)$ порождает векторы тем $\phi_t = p(w|t)$, которые порождают мультиномиальные распределения $\hat{p}(w|t, d)$.



Задачи мультимодального тематического моделирования

Темы определяют распределения не только терминов $p(w|t)$, но и других модальностей: $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{ссылка}|t)$, $p(n\text{-грамма}|t)$, $p(w_{\text{язык}}|t)$, $p(\text{пользователь}|t)$, $p(\text{баннер}|t), \dots$



Мультимодальная ARTM

W^m — словарь токенов m -й модальности, $m \in M$

Максимизация суммы \log правдоподобий с регуляризацией:

$$\sum_{m \in M} \tau_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{aligned} \text{E-шаг:} & \left\{ p_{tdw} = \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td}) \right. \\ \text{M-шаг:} & \left\{ \begin{aligned} \phi_{wt} &= \operatorname{norm}_{w \in W^m} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right), & n_{wt} &= \sum_{d \in D} \tau_m(w) n_{dw} p_{tdw} \\ \theta_{td} &= \operatorname{norm}_{t \in T} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right), & n_{td} &= \sum_{w \in d} \tau_m(w) n_{dw} p_{tdw} \end{aligned} \right. \end{aligned}$$

Модальность биграмм улучшает интерпретируемость тем

Коллекция 850 статей конференций ММРО, ИОИ на русском

| распознавание образов в биоинформатике | | теория вычислительной сложности | |
|--|-------------------------|---------------------------------|----------------------|
| unigrams | bigrams | unigrams | bigrams |
| объект | задача распознавания | задача | разделять множества |
| задача | множество мотивов | множество | конечное множество |
| множество | система масок | подмножество | условие задачи |
| мотив | вторичная структура | условие | задача о покрытии |
| разрешимость | структура белка | класс | покрытие множества |
| выборка | распознавание вторичной | решение | сильный смысл |
| маска | состояние объекта | конечный | разделяющий комитет |
| распознавание | обучающая выборка | число | минимальный аффинный |
| информативность | оценка информативности | аффинный | аффинный комитет |
| состояние | множество объектов | случай | аффинный разделяющий |
| закономерность | разрешимость задачи | покрытие | общее положение |
| система | критерий разрешимости | общий | множество точек |
| структура | информативность мотива | пространство | случай задачи |
| значение | первичная структура | схема | общий случай |
| регулярность | тупиговое множество | комитет | задача MASC |

Стенин С. С. Мультиграммные аддитивно регуляризованные тематические модели. Магистерская диссертация, МФТИ, 2015.

Регуляризатор для классификации и категоризации текстов

Y — множество классов;

n_{dy} = [документ d относится к классу y] — обучающие данные;

$p(y|d) = \sum_{t \in T} \phi_{yt} \theta_{td}$ — линейная модель классификации.

Регуляризатор — правдоподобие модальности классов:

$$R(\Phi, \Theta) = \tau \sum_{d \in D} \sum_{y \in Y} n_{dy} \ln \sum_{t \in T} \phi_{yt} \theta_{td} \rightarrow \max,$$

это тематическая модель с двумя модальностями, W и Y .

TM превосходит SVM в случае несбалансированных классов.

Rubin T. N., Chambers A., Smyth P., Steyvers M. Statistical topic models for multi-label document classification. Machine Learning, 2012.

Vorontsov, Frei, Apishev, Romov, Suvorova, Yanina. Non-Bayesian additive regularization for multimodal topic modeling of large collections. CIKM-2015 WTM.

Регуляризатор для задач регрессии

$y_d \in \mathbb{R}$ для всех документов d — обучающие данные.

$E(y|d) = \sum_{t \in T} v_t \theta_{td}$ — линейная модель регрессии, $v \in \mathbb{R}^{|T|}$.

Регуляризатор — среднеквадратичная ошибка (МНК):

$$R(\Theta, v) = -\tau \sum_{d \in D} \left(y_d - \sum_{t \in T} v_t \theta_{td} \right)^2 \rightarrow \max$$

Подставляем, получаем формулы М-шага:

$$\theta_{td} = \operatorname{norm}_{t \in T} \left(n_{td} + \tau v_t \theta_{td} \left(y_d - \sum_{t \in T} v_t \theta_{td} \right) \right);$$
$$v = (\Theta \Theta^T)^{-1} \Theta y.$$

Sokolov E., Bogolubsky L. Topic Models Regularization and Initialization for Regression Problems // CIKM-2015 Workshop on Topic Models. ACM, pp. 21–27.

Примеры задач регрессии на текстах

MovieReview [Pang, Lee, 2005]

d — текст отзыва на фильм

y_d — рейтинг фильма (1..5), поставленный автором отзыва

Salary (kaggle.com: *Adzuna Job Salary Prediction*)

d — описание вакансии, предлагаемой работодателем

y_d — годовая зарплата

Yelp (kaggle.com: *Yelp Recruiting Competition*)

d — отзыв (на ресторан, отель, сервис и т.п.)

y_d — число голосов «useful», которые получит отзыв

Прогнозирование скачков цен на финансовых рынках

d — текст новости

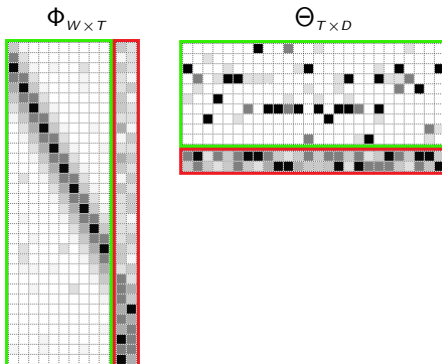
y_d — изменение цены в последующие 10–60 минут

B. Pang, L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales // ACL, 2005.

Разделение тем на предметные и фоновые

Предметные темы S содержат термины предметной области, $p(w|t)$, $p(t|d)$, $t \in S$ — разреженные, существенно различные

Фоновые темы B содержат слова общей лексики, $p(w|t)$, $p(t|d)$, $t \in B$ — существенно отличные от нуля



Регуляризаторы сглаживания и разреживания

Сглаживание фоновых тем $B \subset T$:

Распределения ϕ_{wt} близки к заданному распределению β_w

Распределения θ_{td} близки к заданному распределению α_t

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in B} \sum_{w \in W} \beta_w \ln \phi_{wt} + \alpha_0 \sum_{d \in D} \sum_{t \in B} \alpha_t \ln \theta_{td} \rightarrow \max,$$

где β_0, α_0 — коэффициенты регуляризации

Разреживание предметных тем $S = T \setminus B$:

Распределения ϕ_{wt} **далеки** от заданного распределения β_w

Распределения θ_{td} **далеки** от заданного распределения α_t

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in S} \sum_{w \in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d \in D} \sum_{t \in S} \alpha_t \ln \theta_{td} \rightarrow \max.$$

где β_0, α_0 — коэффициенты регуляризации.

Регуляризатор декоррелирования тем

Цель: усилить различность тем; выделить в каждой теме лексическое ядро, отличающее её от других тем; вывести слова общей лексики из предметных тем в фоновые.

Минимизируем ковариации между вектор-столбцами ϕ_t :

$$R(\Phi) = -\frac{\tau}{2} \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \phi_{wt} \phi_{ws} \rightarrow \max.$$

Подставляем, получаем ещё один вариант разреживания — постепенное контрастирование строк матрицы Φ :

$$\phi_{wt} = \operatorname{norm}_w \left(n_{wt} - \tau \phi_{wt} \sum_{s \in T \setminus t} \phi_{ws} \right).$$

Tan Y., Ou Z. Topic-weak-correlated latent Dirichlet allocation // 7th Int'l Symp. Chinese Spoken Language Processing (ISCSLP), 2010. — Pp. 224–228.

Иерархическая тематическая модель: послойное построение

Шаг 1. Строим модель с небольшим числом тем.

Шаг k . Пусть модель с множеством тем T уже построена.
Строим множество дочерних тем S (subtopics), $|S| > |T|$.

Родительские темы приближаются смесями дочерних тем:

$$\sum_{t \in T} n_{wt} \ln p(w|t) = \sum_{t \in T} n_{wt} \ln \sum_{s \in S} p(w|s)p(s|t) \rightarrow \max_{\Phi, \Psi}$$

где $p(s|t) = \psi_{st}$, $\Psi = (\psi_{st})_{S \times T}$ — матрица связей.

Родительская $\Phi^p \approx \Phi\Psi$, отсюда регуляризатор матрицы Φ :

$$R(\Phi, \Psi) = \tau \sum_{t \in T} \sum_{w \in W} n_{wt} \ln \sum_{s \in S} \phi_{ws} \psi_{st} \rightarrow \max.$$

Родительские темы t — псевдо-документы с частотами слов n_{wt} .

Правдоподобие и перплексия (perplexity)

Правдоподобие языковой модели $p(w|d)$ (чем выше, тем лучше):

$$\mathcal{L}(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d), \quad p(w|d) = \sum_t \phi_{wt} \theta_{td}$$

Перплексия языковой модели $p(w|d)$ (чем меньше, тем лучше):

$$\mathcal{P}(D) = \exp\left(-\frac{1}{n} \sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d)\right), \quad n = \sum_{d \in D} \sum_{w \in d} n_{dw}$$

Интерпретация перплексии:

- если распределение $p(w|d) = \frac{1}{|W|}$ равномерное, то $\mathcal{P} = |W|$
- мера различности или неопределённости слов в тексте
- коэффициент ветвления (branching factor) текста

Перплексия тестовой (отложенной) коллекции

Перплексия тестовой коллекции D' (hold-out perplexity):

$$\mathcal{P}(D') = \exp\left(-\frac{1}{n''} \sum_{d \in D'} \sum_{w \in d''} n_{dw} \ln p(w|d)\right), \quad n'' = \sum_{d \in D'} \sum_{w \in d''} n_{dw}$$

$d = d' \sqcup d''$ — случайное разбиение тестового документа на две половины равной длины;

параметры ϕ_{wt} оцениваются по обучающей коллекции D ;

параметры θ_{td} оцениваются по первой половине d' ;

перплексия вычисляется по второй половине d'' .

Интерпретируемости и когерентность

Тема интерпретируемая, если по топовым словам темы эксперт может определить, о чём эта тема, и дать ей название.

- Экспертные оценки:
 - интерпретируемость темы по балльной шкале;
 - каждую тему оценивают несколько экспертов.
- Метод интрузий (intrusion):
 - в список топовых слов внедряется лишнее слово;
 - измеряется доля ошибок экспертов его при определении

Нужна автоматически вычисляемая мера интерпретируемости, коррелирующая с экспертными оценками.

Ею оказалась *когерентность* (согласованность, coherence).

Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

Эксперимент. Связь когерентности и интерпретируемости

Измерялась ранговая корреляция Спирмена между 15 метрикам и экспертными оценками интерпретируемости.

PMI — лучшая метрика.

Gold-standard — средняя корреляция Спирмена между оценками разных экспертов.

| Resource | Method | Median | Mean |
|---------------|------------|-------------|-------------|
| WordNet | HSO | 0.15 | 0.59 |
| | JCN | -0.20 | 0.19 |
| | LCH | -0.31 | -0.15 |
| | LESK | 0.53 | 0.53 |
| | LIN | 0.09 | 0.28 |
| | PATH | 0.29 | 0.12 |
| | RES | 0.57 | 0.66 |
| | VECTOR | -0.08 | 0.27 |
| | WuP | 0.41 | 0.26 |
| | RACO | 0.62 | 0.69 |
| Wikipedia | MiW | 0.68 | 0.70 |
| | DOCsim | 0.59 | 0.60 |
| | PMI | 0.74 | 0.77 |
| Google | TITLES | 0.51 | |
| | LOGHITS | -0.19 | |
| Gold-standard | IAA | 0.82 | 0.78 |

Вывод: когерентность близка к «золотому стандарту».

Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

Когерентность как внутренняя мера интерпретируемости

Когерентность (согласованность) темы t по k топовым словам:

$$\text{PMI}_t = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^k \text{PMI}(w_i, w_j)$$

где w_i — i -й термин в порядке убывания ϕ_{wt} .

$\text{PMI}(u, v) = \ln \frac{|D|N_{uv}}{N_u N_v}$ — поточечная взаимная информация (pointwise mutual information),

N_{uv} — число документов, в которых термины u, v хотя бы один раз встречаются рядом (в окне 10 слов),

N_u — число документов, в которых u встретился хотя бы 1 раз.

Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

Качества разведочного поиска документов по документам

Поисковый запрос

набор ключевых слов или фрагментов текста, около одной страницы A4

Поисковая выдача

документы d с распределением $p(t|d)$, близким к распределению $p(t|q)$ запроса

Два задания ассессорам

- найти как можно больше статей, пользуясь любыми средствами поиска (и засесть время)
- оценить релевантность поисковой выдачи на том же запросе

Поиск MapReduce

Поиск MapReduce – программа поиска (поиск) написанная распределенно: выделены для больших объемов данных и разнородных шардов, представляющих собой набор Java-классов и исполняемых утилит для создания и обработки данных на параллельной обработке.

Основные компоненты Поиск MapReduce можно сформулировать как:

- обработка вычислением больших объемов данных;
- масштабируемость;
- автоматическое распределение данных;
- работа на неоднородном оборудовании;
- автоматическая обработка отказов вычислений данных.

Поиск – популярная программная платформа (**поиск**, **поиск**) построена распределенными приложениями для массово-параллельной обработки (**поиск**, **поиск**, **поиск**, **поиск**).

Поиск включает в себе следующие компоненты:

1. HDFS – распределенная файловая система;
2. **Поиск MapReduce** – программная платформа (**поиск**) написанная распределенно, выделенная для больших объемов данных и разнородных шардов.

Ключевые, **поиск** и архитектура **Поиск MapReduce** и структура HDFS, стали прототипом ряда других систем в области вычислений, в том числе и основные точки отказа. Это, в конечном итоге, определило направление платформ **Поиск** и сейчас. В целом, в последние годы можно отметить:

Ограничение масштабируемости кластера **Поиск** – это вычислительный узел, – это параллельные данные.

Сильная взаимосвязь **Поиск** распределенно вычислений и клиентских выделенных распределенных алгоритмов. Как следствие:

Существенная поддержка алгоритмической программы выделенно распределенно вычислений в **Поиск** v1.0 поддерживается только только выделенно выделенно.

Клиент выделенно, точки отказа и как следствие, выделенность выделенности в средстве с выделенными требованиями к выделенности.

Проблема **выделенности** совместности: требование по выделенности выделенности всех выделенных узлов кластера при выделенности платформ **Поиск** (установка новых версий или пакета обновлений).

Пример запроса для разведочного поиска

Две коллекции новостей про технологии

habr.com/ru

175 143 статей на русском
10 552 слов (униграмм)
742 000 биграмм
524 авторов статей
10 000 авторов комментариев
2546 тегов
123 хаба (категории)

TechCrunch.com

759 324 статей на английском
11 523 слов (униграмм)
1.2 млн. биграмм
605 авторов
184 категорий



Векторный поиск тематически близких документов

$\theta_{tq} = p(t|q)$ — тематический вектор запроса q

$\theta_{td} = p(t|d)$ — тематические векторы документов $d \in D$

Косинусная мера близости документа d и запроса q :

$$\text{sim}(q, d) = \frac{\sum_t \theta_{tq} \theta_{td}}{(\sum_t \theta_{tq}^2)^{1/2} (\sum_t \theta_{td}^2)^{1/2}}.$$

Ранжируем документы коллекции $d \in D$ по убыванию $\text{sim}(q, d)$

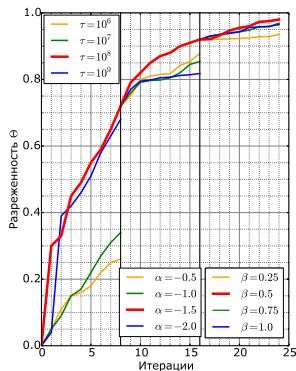
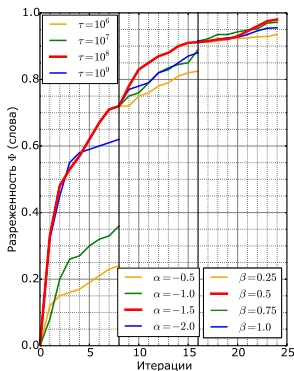
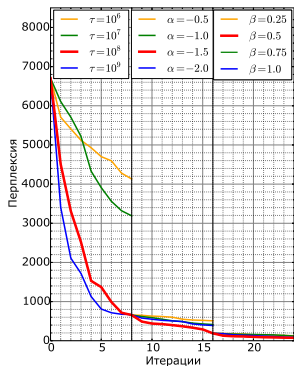
Выдача тематического поиска — k первых документов.

Реализация: *векторный индекс* для быстрого поиска документов d по каждой из тем t запроса

A.Ianina, K.Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

Последовательный подбор коэффициентов регуляризации

- декоррелирование распределений термов в темах (τ),
- разреживание распределений тем в документах (α),
- сглаживание распределений термов в темах (β).



Какие модели поиска сравнивались

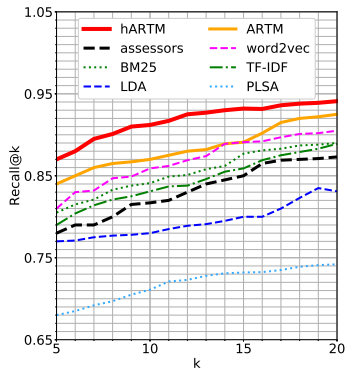
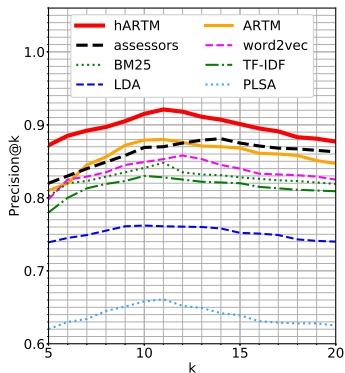
- **assessors**: результаты поиска, выполненного ассессорами
- **TF-IDF, BM25**: сравнение документов по частотам слов
- **word2vec**: нетематические векторные представления слов
- **PLSA**: Probabilistic Latent Semantic Analysis (1999)
- **LDA**: Latent Dirichlet Allocation (2003)
- **ARTM**: тематическая модель с тремя регуляризаторами
- **hARTM**: трёхуровневая иерархическая модель ARTM

Задачи регуляризаторов в ARTM и hARTM:

- сделать темы как можно более различными
- сделать векторы $p(t|d)$ как можно более разреженными
- не допустить вырождения распределений $p(w|t)$

Сравнение качества поиска с ассессорами и простыми моделями

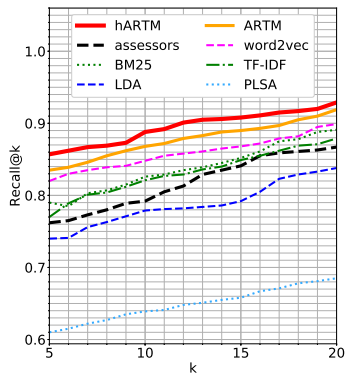
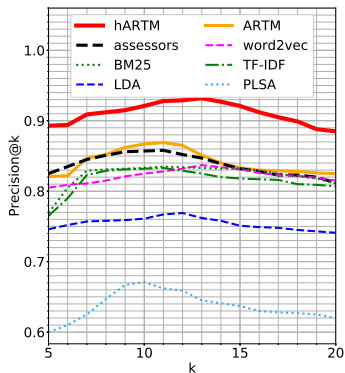
Точность и полнота по первым k позициям поисковой выдачи (коллекция **habr.com/ru**)



A. Ianina, K. Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

Сравнение качества поиска с ассессорами и простыми моделями

Точность и полнота по первым k позициям поисковой выдачи (коллекция TechCrunch.com)



A. Ianina, K. Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

Влияние числа тем на качество поиска

habr.com/ru. Все регуляризаторы и модальности, **три уровня**

| $ T_1 $ | 20 | | 25 | | | | | | 30 | | |
|---------|-------|-------|-------|-------|-------|--------------|--------------|-------|-------|-------|-------|
| $ T_2 $ | 150 | 200 | 250 | | 275 | | | 300 | | 400 | 450 |
| $ T_3 $ | 750 | 800 | 1200 | 1300 | 1300 | 1400 | 1500 | 1500 | 1600 | 3000 | 3500 |
| Pr@5 | 0.625 | 0.743 | 0.840 | 0.852 | 0.869 | 0.872 | 0.870 | 0.805 | 0.771 | 0.705 | 0.672 |
| Pr@10 | 0.648 | 0.754 | 0.851 | 0.867 | 0.882 | 0.915 | 0.901 | 0.811 | 0.799 | 0.722 | 0.694 |
| Pr@15 | 0.632 | 0.752 | 0.850 | 0.872 | 0.878 | 0.895 | 0.889 | 0.809 | 0.785 | 0.729 | 0.703 |
| Pr@20 | 0.629 | 0.745 | 0.845 | 0.861 | 0.871 | 0.877 | 0.882 | 0.803 | 0.778 | 0.710 | 0.681 |
| R@5 | 0.632 | 0.780 | 0.845 | 0.869 | 0.883 | 0.889 | 0.872 | 0.851 | 0.841 | 0.721 | 0.695 |
| R@10 | 0.654 | 0.792 | 0.859 | 0.873 | 0.905 | 0.922 | 0.881 | 0.873 | 0.850 | 0.749 | 0.703 |
| R@15 | 0.675 | 0.805 | 0.874 | 0.892 | 0.932 | 0.942 | 0.905 | 0.889 | 0.863 | 0.787 | 0.725 |
| R@20 | 0.684 | 0.824 | 0.889 | 0.901 | 0.958 | 0.961 | 0.912 | 0.904 | 0.878 | 0.805 | 0.734 |

- существует оптимальное число тем на каждом уровне
- три уровня лучше, чем один или два
- на нижнем уровне оптимальное число тем увеличивается

Влияние числа тем на качество поиска

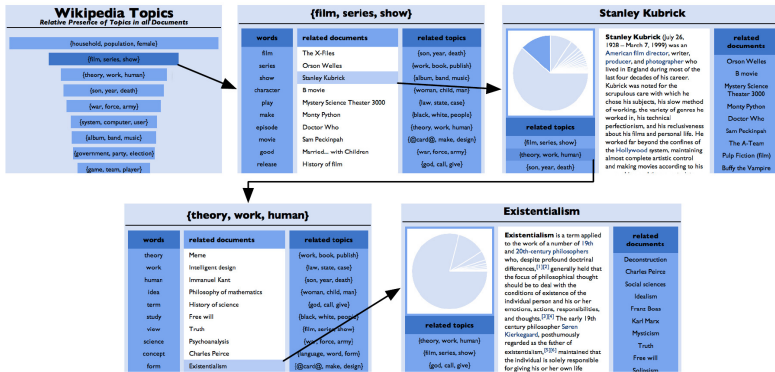
TechCrunch. Все регуляризаторы и модальности, три уровня

| $ T_1 $ | 80 | | 100 | | | | 120 | | | | |
|---------|-------|-------|-------|-------|-------|--------------|--------------|-------|-------|-------|-------|
| $ T_2 $ | 300 | 350 | 500 | 550 | | 600 | 700 | 750 | | | |
| $ T_3 $ | 1500 | 1700 | 2500 | 2600 | 2600 | 2800 | 3000 | 3000 | 3200 | 4500 | 4700 |
| Pr@5 | 0.655 | 0.707 | 0.751 | 0.792 | 0.887 | 0.893 | 0.890 | 0.789 | 0.722 | 0.703 | 0.678 |
| Pr@10 | 0.678 | 0.712 | 0.773 | 0.823 | 0.895 | 0.922 | 0.905 | 0.805 | 0.741 | 0.722 | 0.692 |
| Pr@15 | 0.692 | 0.715 | 0.775 | 0.831 | 0.902 | 0.921 | 0.907 | 0.821 | 0.743 | 0.725 | 0.703 |
| Pr@20 | 0.687 | 0.709 | 0.761 | 0.819 | 0.889 | 0.885 | 0.898 | 0.809 | 0.736 | 0.719 | 0.683 |
| R@5 | 0.751 | 0.795 | 0.802 | 0.856 | 0.871 | 0.877 | 0.863 | 0.852 | 0.831 | 0.738 | 0.705 |
| R@10 | 0.767 | 0.812 | 0.825 | 0.875 | 0.892 | 0.908 | 0.879 | 0.871 | 0.842 | 0.751 | 0.711 |
| R@15 | 0.772 | 0.824 | 0.841 | 0.887 | 0.912 | 0.927 | 0.901 | 0.893 | 0.854 | 0.772 | 0.721 |
| R@20 | 0.783 | 0.830 | 0.854 | 0.892 | 0.931 | 0.949 | 0.935 | 0.905 | 0.871 | 0.790 | 0.732 |

- существует оптимальное число тем на каждом уровне
- три уровня лучше, чем один или два
- на нижнем уровне оптимальное число тем увеличивается

Система TMVE — Topic Model Visualization Engine

Тематический навигатор с веб-интерфейсом:

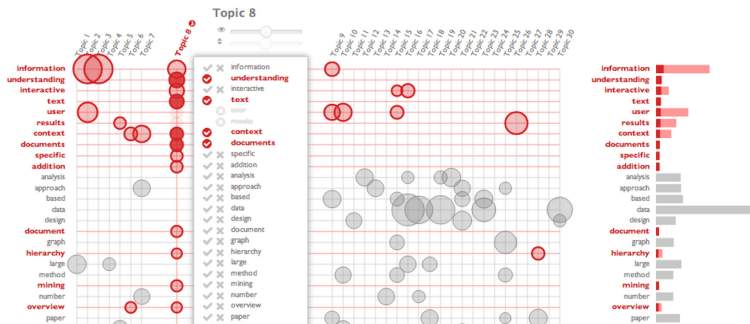


<https://github.com/ajbc/tmv>

Chaney A., Blei D. Visualizing Topic Models // Frontiers of computer science in China, 2012. — 55(4), pp. 77–84.

Система Termite

Интерактивная визуализация матрицы Φ и сравнение тем:

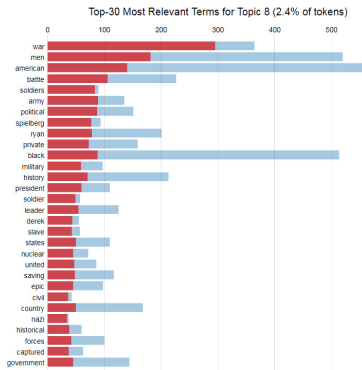
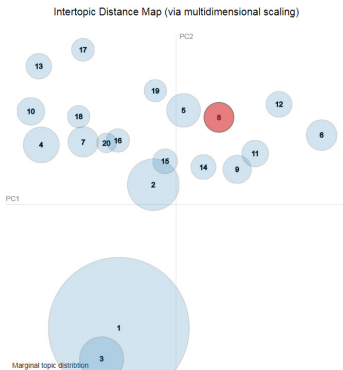


<https://github.com/uwdata/termite-visualizations>

Chuang J., Manning C., Heer J. Termite: Visualization Techniques for Assessing Textual Topic Models. IWCAVI 2012.

Система LDAvis

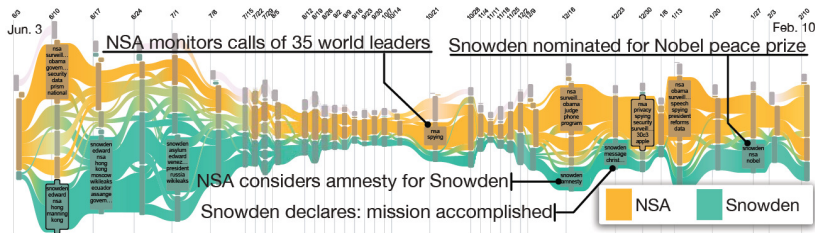
Карта сходства тем и сравнение $p(w|t)$ с $p(w)$:



<https://github.com/cpsievert/LDAvis>

C.Sievert, K.Shirley. LDAvis: A method for visualizing and interpreting topics. 2014.

Динамика тем: эволюция предметной области



Эволюция выбранных тем иерархии. Данные Prism (2013/06/03–2014/02/09)

- эксперт задаёт сечение иерархии (дерева) тем,
- интерактивно выбирает подмножество тем и событий,
- генерирует отчёт.

Weiwei Cui, Shixia Liu, Zhuofeng Wu, Hao Wei. How hierarchical topics evolve in large text corpora. 2014.

- Тематическое моделирование — это восстановление латентных тем по коллекции текстовых документов
- Задача сводится к стохастическому матричному разложению
- Стандартные методы — PLSA и LDA.
- Задача является некорректно поставленной, так как множество её решений в общем случае бесконечно
- Аддитивная регуляризация позволяет комбинировать модели и строить модели с заданными свойствами
- В отличие от классических задач машинного обучения, регуляризаторы весьма разнообразны
- На практике важны внешние критерии качества моделей