Информационный поиск: обучение ранжированию и 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)»

30 мая 2018

Содержание

Определения и обозначения

X — множество объектов
$X^\ell = \{x_1, \ldots, x_\ell\}$ — обучающая выборка
$i \prec j$ — правильный порядок на парах $(i, j) \in \{1, \ldots, \ell\}^2$

Задача:
построить ранжирующую функцию $a\colon X \to \mathbb{R}$ такую, что

$$i \prec j \quad \Rightarrow \quad a(x_i) < a(x_j)$$

Линейная модель ранжирования:

$$a(x; w) = \langle x, w \rangle$$

где $\big(f_1(x), \ldots, f_n(x)\big) \in \mathbb{R}^n$ — вектор признаков объекта x

Задача ранжирования поисковой выдачи

D — коллекция текстовых документов (documents)
Q — множество запросов (queries)
$D_q \subseteq D$ — множество документов, найденных по запросу q
$X = Q \times D$ — объектами являются пары «запрос, документ»:

$$x \equiv (q, d),\ q \in Q,\ d \in D_q$$

Y — упорядоченное множество рейтингов
$y : X \to Y$ — оценки релевантности, поставленные асессорами:
чем выше оценка $y(q, d)$, тем релевантнее документ d запросу q

Правильный порядок определён только между документами, найденными по одному и тому же запросу q:

$$(q, d) \prec (q, d') \quad \Leftrightarrow \quad y(q, d) < y(q, d')$$

Признаки в задачах ранжирования поисковой выдачи

Типы признаков

- функции только документа d
- функции только запроса q
- функции запроса и документа (q, d)

- текстовые
  — слова запроса q встречаются в d чаще обычного
  — слова запроса q есть в заголовках или выделены в d
- ссылочные
  — на документ d много ссылаются
  — документ d содержит много полезных ссылок
- кликовые
  — на документ d часто кликают
  — на документ d часто кликают по запросу q

TF-IDF(q, d) — мера релевантности документа d запросу q

n_{dw} (term frequency) — число вхождений слова w в текст d;
N_w (document frequency) — число документов, содержащих w;
N — число документов в коллекции D;

N_w/N — оценка вероятности встретить слово w в документе;
$(N_w/N)^{n_{dw}}$ — оценка вероятности встретить его n_{dw} раз;
$P(q,d) = \prod\limits_{w \in q} (N_w/N)^{n_{dw}}$ — оценка вероятности встретить
в документе d слова запроса $q = \{w_1, \ldots, w_k\}$ *чисто случайно*;

Оценка релевантности запроса q документу d:

$$-\log P(q,d) = \sum_{w \in q} \underbrace{n_{dw}}_{\mathrm{TF}(w,d)} \underbrace{\log(N/N_w)}_{\mathrm{IDF}(w)} \;\to\; \max.$$

$\mathrm{TF}(w,d) = n_{dw}$ — term frequency;
$\mathrm{IDF}(w) = \log(N/N_w)$ — inverted document frequency.

PageRank — классический ссылочный признак

Документ d тем важнее,
— чем больше других документов c ссылаются на d,
— чем важнее документы c, ссылающиеся на d,
— чем меньше других ссылок имеют эти документы c.

Вероятность попасть на страницу d, если кликать случайно:

$$\mathrm{PR}(d) = \frac{1-\delta}{N} + \delta \sum_{c \in D_d^{in}} \frac{\mathrm{PR}(c)}{|D_c^{out}|},$$

$D_d^{in} \subset D$ — множество документов, ссылающихся на d,
$D_c^{out} \subset D$ — множество документов, на которые ссылается c,
$\delta = 0.85$ — вероятность продолжать клики (damping factor),
N — число документов в коллекции D.

Sergey Brin, Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. 1998.

Точность и средняя точность

Пусть $Y = \{0, 1\}$, $y(q, d)$ — релевантность,
$a(q, d)$ — искомая функция ранжирования,
$d_q^{(i)}$ — i-й документ по убыванию $a(q, d)$.

Precision, точность — доля релевантных среди первых n:

$$P_n(q) = \frac{1}{n} \sum_{i=1}^{n} y(q, d_q^{(i)})$$

Average Precision, средняя P_n по позициям релевантных документов:

$$AP(q) = \sum_n y(q, d_q^{(n)}) P_n(q) \Big/ \sum_n y(q, d_q^{(n)})$$

Mean Average Precision, средняя AP по всем запросам:

$$MAP = \frac{1}{|Q|} \sum_{q \in Q} AP(q)$$

Доля «дефектных пар»

Пусть $Y \subseteq \mathbb{R}$, $y(q,d)$ — релевантность,
$a(q,d)$ — искомая функция ранжирования,
$d_q^{(i)}$ — i-й документ по убыванию $a(q,d)$.

Число инверсий порядка среди первых n документов:

$$DP_n(q) = \sum_{i<j}^{n} \Big[y(q, d_q^{(i)}) < y(q, d_q^{(j)}) \Big].$$

Связь с AUC (area under ROC-curve) в задачах классификации с двумя классами $Y = \{-1, +1\}$, $a\colon X \to Y$

$$AUC_n(q) = \frac{1}{\ell_- \ell_+} \sum_{i,j=1}^{n} \big[y_i < y_j\big] \big[a(x_i) < a(x_j)\big] = 1 - \frac{1}{\ell_- \ell_+} DP_n(q).$$

DCG — Discounted Cumulative Gain

Пусть $Y \subseteq \mathbb{R}$, $y(q,d)$ — релевантность,
$a(q,d)$ — искомая функция ранжирования,
$d_q^{(i)}$ — i-й документ по убыванию $a(q,d)$.

Дисконтированная (взвешенная) сумма выигрышей:

$$DCG_n(q) = \sum_{i=1}^{n} \underbrace{G_q(d_q^{(i)})}_{\text{gain}} \cdot \underbrace{D(i)}_{\text{discount}}$$

$G_q(d) = (2^{y(q,d)} - 1)$ — бóльший вес релевантным документам
$D(i) = 1/\log_2(i+1)$ — бóльший вес в начале выдачи

Нормированная дисконтированная сумма выигрышей:

$$NDCG_n(q) = \frac{DCG_n(q)}{\max DCG_n(q)}$$

$\max DCG_n(q)$ — это $DCG_n(q)$ при идеальном ранжировании

Яндекс pFound — модель поведения пользователя

Пусть $Y \subseteq [0,1]$,
$y(q,d)$ — релевантность, оценка вероятности найти ответ в d,
$a(q,d)$ — искомая функция ранжирования,
$d_q^{(i)}$ — i-й документ по убыванию $a(q,d)$.

Вероятность найти ответ в первых n документах:

$$pFound_n(q) = \sum_{i=1}^{n} P_i \cdot y\left(q, d_q^{(i)}\right),$$

где P_i — вероятность дойти до i-го документа:

$$P_1 = 1;$$

$$P_{i+1} = P_i \cdot \left(1 - y\left(q, d_q^{(i)}\right)\right) \cdot \left(1 - P_{out}\right),$$

где P_{out} — вероятность прекратить поиск без ответа

Яндекс pFound — модель поведения пользователя

Параметры критерия pFound:

- $P_{out} = 0.15$ — вероятность прекратить поиск без ответа;
- $y(q, d)$ — оценка вероятности найти ответ в документе:

оценка асессора	$y(q, d)$
Vital	0.61
Useful	0.41
Relevant+	0.14
Relevant−	0.07
Not Relevant	0.00

Гулин А., Карпович П., Расковалов Д., Сегалович И. Оптимизация алгоритмов ранжирования методами машинного обучения. РОМИП-2009.

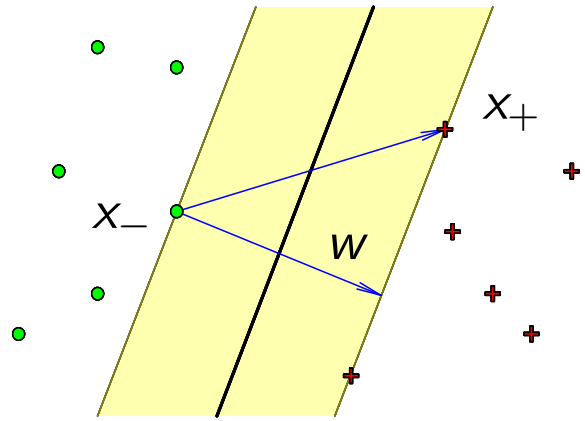
Напоминание: SVM — метод опорных векторов

Линейный классификатор:

$$a(x) = \text{sign}\big(\langle w, x\rangle - w_0\big), \quad w, x \in \mathbb{R}^n,\ w_0 \in \mathbb{R}.$$

Задача обучения SVM:

$$\begin{cases} \dfrac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{\ell}\xi_i \to \min\limits_{w,w_0,\xi}; \\ M_i(w,w_0) \geqslant 1 - \xi_i, \quad i = 1,\ldots,\ell; \\ \xi_i \geqslant 0, \quad i = 1,\ldots,\ell. \end{cases}$$



где $M_i(w,w_0) = y_i\big(\langle w, x_i\rangle - w_0\big)$ — отступ объекта x_i.

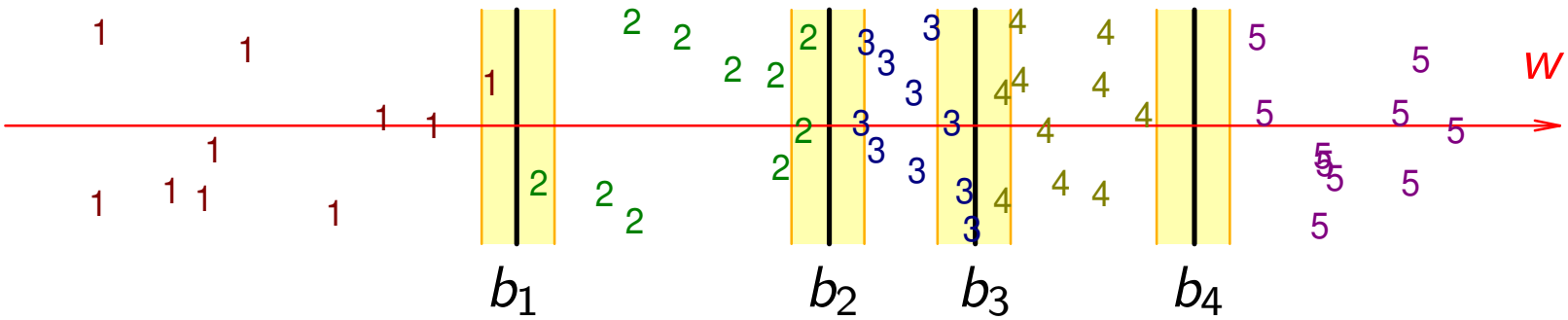
Эквивалентная задача безусловной минимизации:

$$Q(w,w_0) = \sum_{i=1}^{\ell}\big(1 - M_i(w,w_0)\big)_+ + \frac{1}{2C}\|w\|^2 \to \min_{w,w_0}.$$

Ранговая классификация OC-SVM (Ordinal Classification SVM)

Пусть $Y = \{1, \ldots, K\}$, функция ранжирования *линейная* с порогами $b_0 = -\infty$, $b_1, \ldots, b_{K-1} \in \mathbb{R}$, $b_K = +\infty$:

$$a(x) = y, \quad \text{если} \quad b_{y-1} < \langle w, x \rangle \leqslant b_y$$



Постановка задачи SVM для ранговой классификации:

$$\begin{cases} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell}[y_i \neq K]\,(\xi_i + \xi_i^*) \to \min\limits_{w,b,\xi}; \\ b_{y_i-1} + 1 - \xi_i^* \leqslant \langle w, x_i \rangle \leqslant b_{y_i} - 1 + \xi_i; \\ \xi_i^* \geqslant 0, \quad \xi_i \geqslant 0. \end{cases}$$

Три основных подхода к ранжированию

- Point-wise — поточечный (в частности, OC-SVM)
- Pair-wise — попарный
- List-wise — списочный

Переход к гладкому функционалу качества ранжирования:

$$Q(a) = \sum_{i \prec j} \big[\underbrace{a(x_j) - a(x_i)}_{\text{Margin}_{ij}} < 0\big] \leqslant \sum_{i \prec j} \mathscr{L}\big(a(x_j) - a(x_i)\big) \quad \to \quad \min_a$$

где $a(x)$ — функция ранжирования, $M = \text{Margin}_{ij}$ — отступ, $\mathscr{L}(M)$ — невозрастающая непрерывная функция отступа:

- $\mathscr{L}(M) = (1 - M)_+$ — RankSVM
- $\mathscr{L}(M) = \exp(-M)$ — RankBoost
- $\mathscr{L}(M) = \log(1 + e^{-M})$ — RankNet

Ranking SVM

Постановка задачи SVM для попарного подхода:

$$Q(a) = \tfrac{1}{2}\|w\|^2 + C\sum_{i \prec j} \mathscr{L}\big(\underbrace{a(x_j, w) - a(x_i, w)}_{\text{Margin}_{ij}(w)}\big) \quad \to \quad \min_w,$$

где $a(x) = \langle w, x\rangle$ — функция ранжирования,
$\mathscr{L}(M) = (1 - M)_+$ — функция потерь,
$M = \text{Margin}_{ij}(w) = \langle w, x_j - x_i\rangle$ — отступ,

Постановка задачи квадратичного программирования:

$$\begin{cases} \tfrac{1}{2}\|w\|^2 + C\sum_{i \prec j} \xi_{ij} \to \min\limits_{w,\xi}; \\ \langle w, x_j - x_i\rangle \geqslant 1 - \xi_{ij}, \quad i \prec j; \\ \xi_{ij} \geqslant 0, \quad i \prec j. \end{cases}$$

Переход от попарного подхода к списочному

RankNet: попарный подход, линейная модель $a(x) = \langle w, x \rangle$, гладкая функция потерь $\mathscr{L}(M) = \log(1 + e^{-\sigma M})$:

$$Q(a) = \sum_{i \prec j} \mathscr{L}\big(a(x_j) - a(x_i)\big) \quad \to \quad \min_w$$

Метод стохастического градиента: для случайной пары $i \prec j$

$$w := w + \eta \cdot \frac{\sigma}{1 + \exp(\sigma \langle x_j - x_i, w \rangle)} \cdot (x_j - x_i)$$

LambdaRank: оптимизация негладкого Q (MAP, NDCG, pFound)
Домножаем градиентный шаг на изменение функционала ΔQ_{ij} при перестановке объектов местами $x_i \leftrightarrows x_j$ в списке выдачи:

$$w := w + \eta \cdot \frac{\sigma}{1 + \exp(\sigma \langle x_j - x_i, w \rangle)} \cdot |\Delta Q_{ij}| \cdot (x_j - x_i);$$

C.Burges. From RankNet to LambdaRank to LambdaMART: an overview. 2010.

Резюме по ранжированию

- Ранжирование — особый класс задач машинного обучения
- Три подхода: поточечный, попарный, списочный
- Критерии качества не универсальны, зависят от приложения

Ранжирование в Яндексе:

- Ежемесячно добавляется более 50 000 оценок асессоров
- За 8 лет придумано и проверено более 2000 признаков
- PairWise подход лучше, чем PointWise и ListWise
- Технология MatrixNet — градиентный бустинг над ODT (небрежными решающими деревьями)
- CatBoost — свободно доступный аналог MatrixNet

Tie-Yan Liu. Learning to Rank for Information Retrieval. 2011.
Hang Li. A Short Introduction to Learning to Rank. 2011.

Что такое «тема» в коллекции текстовых документов?

- *тема* — семантически однородный кластер текстов
- *тема* — специальная терминология предметной области
- *тема* — набор терминов (слов или словосочетаний), совместно часто встречающихся в документах

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.

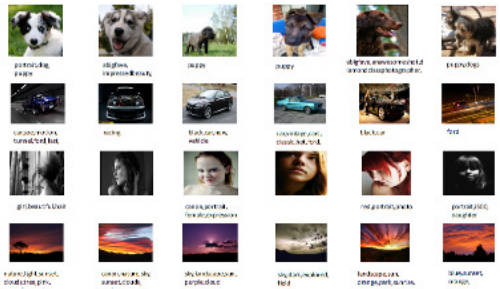
Тематическая модель выявляет латентные темы по наблюдаемым распределениям слов $p(w|d)$ в документах.

Приложения тематического моделирования

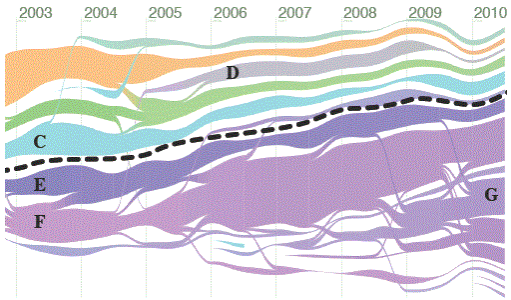
разведочный поиск в электронных библиотеках

персонализированный поиск в соцсетях

мультимодальный поиск текстов и изображений

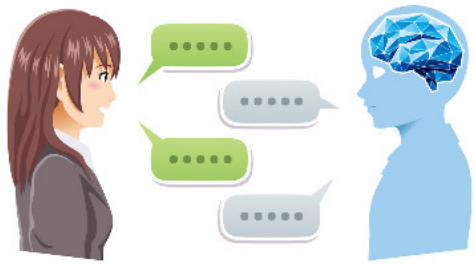


детектирование и трекинг новостных сюжетов



навигация по большим текстовым коллекциям

управлением 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
- порядок слов в документе не важен (bag of words)
- порядок документов в коллекции не важен
- $D \times W \times T$ — дискретное вероятностное пространство
- 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:

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

Задача построения тематической модели коллекции

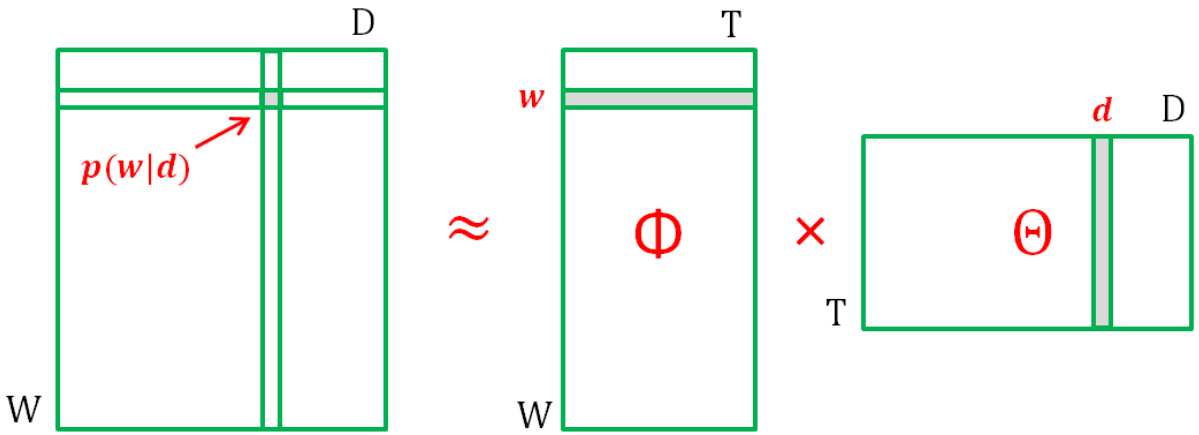
Дано: коллекция текстовых документов

- 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

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d) p(d) \to \max_{\Phi, \Theta}$$

эквивалентна максимизации функционала

$$\mathscr{L}(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \to \max_{\Phi, \Theta}$$

при ограничениях неотрицательности и нормировки

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

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

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

Регуляризация — стандартный приём доопределения решения с помощью дополнительных критериев.

ARTM: аддитивная регуляризация тематических моделей

Максимизация логарифма правдоподобия с регуляризатором:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

где $\underset{t\in T}{\text{norm}}(x_t) = \dfrac{\max\{x_t,0\}}{\sum\limits_{s\in T}\max\{x_s,0\}}$ — операция нормировки вектора.

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 $n_{tdw} = n_{dw}p(t|d,w)$:

$$\phi_{wt} = \frac{n_{wt}}{n_t}, \qquad n_{wt} = \sum_{d\in D} n_{tdw}, \qquad n_t = \sum_{w\in W} n_{wt};$$

$$\theta_{td} = \frac{n_{td}}{n_d}, \qquad n_{td} = \sum_{w\in d} n_{tdw}, \qquad n_d = \sum_{t\in T} n_{td}.$$

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 $w \in W$

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

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 $t \in T$

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

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:

$$\begin{cases} f(x) \to \min\limits_x; \\ g_i(x) \leqslant 0, \quad i = 1, \ldots, m; \\ h_j(x) = 0, \quad j = 1, \ldots, k. \end{cases}$$

Необходимые условия. Если x — точка локального минимума, то существуют множители μ_i, $i = 1, \ldots, m$, λ_j, $j = 1, \ldots, k$:

$$\begin{cases} \dfrac{\partial \mathscr{L}}{\partial x} = 0, \quad \mathscr{L}(x; \mu, \lambda) = f(x) + \sum\limits_{i=1}^{m} \mu_i g_i(x) + \sum\limits_{j=1}^{k} \lambda_j h_j(x); \\ g_i(x) \leqslant 0; \ h_j(x) = 0; \ \text{(исходные ограничения)} \\ \mu_i \geqslant 0; \ \text{(двойственные ограничения)} \\ \mu_i g_i(x) = 0; \ \text{(условие дополняющей нежёсткости)} \end{cases}$$

Вывод системы уравнений из условий Каруша–Куна–Таккера

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

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

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 $\lambda_t \leqslant 0$, то тема t вырождена, $\phi_{wt} \equiv 0$ для всех w.
4. Если $\lambda_t > 0$, то либо $\phi_{wt} = 0$, либо $n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} > 0$:

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

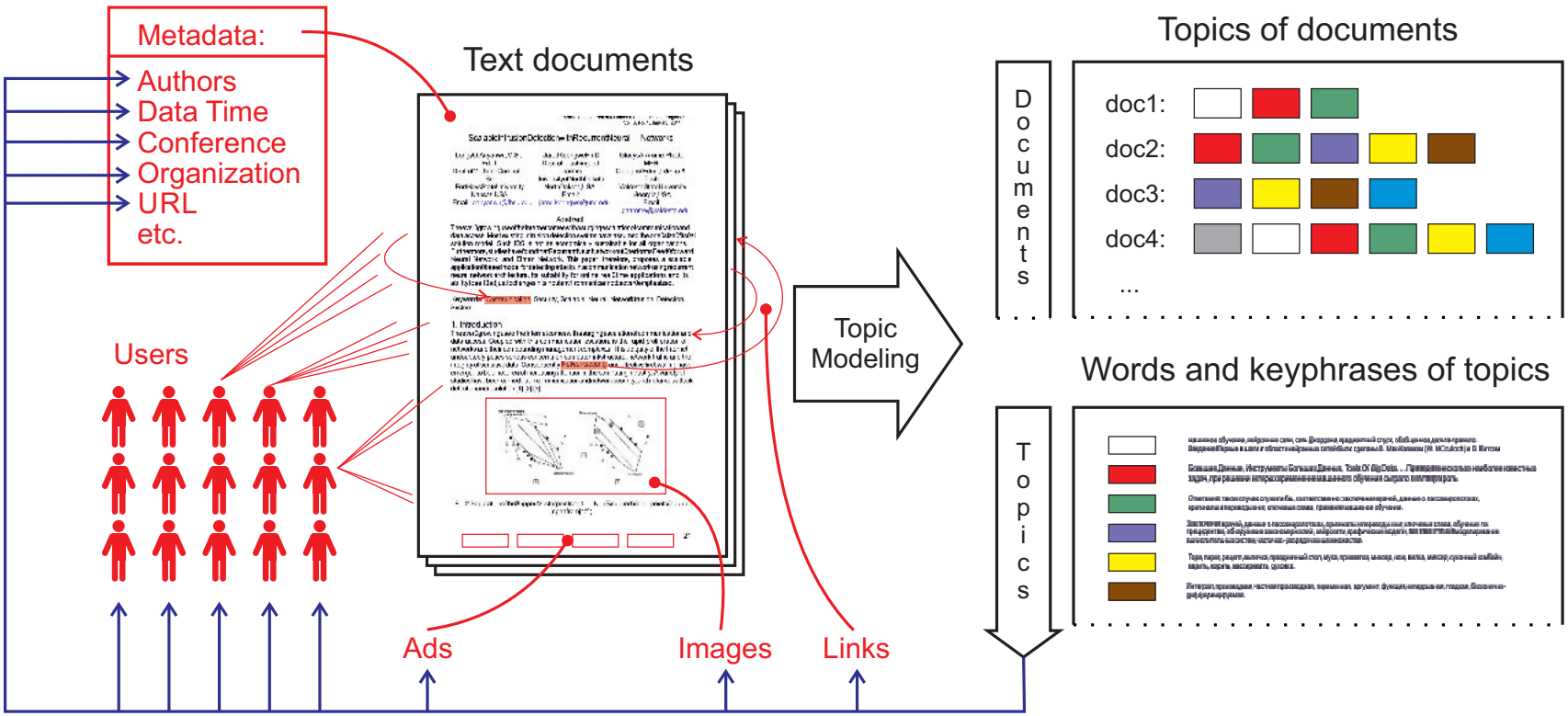
5. Суммируем обе части равенства по $w \in W$:

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

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

Задачи мультимодального тематического моделирования

Темы определяют распределения не только терминов $p(w|t)$, но и других *модальностей*: $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{ссылка}|t)$, $p(\text{баннер}|t)$, $p(\text{элемент\_изображения}|t)$, $p(\text{пользователь}|t), \ldots$



Мультимодальная ARTM

Пусть документы содержат токены разных модальностей.
W^m — словарь токенов m-й модальности, $m \in M$

Максимизация суммы log правдоподобий с регуляризацией:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

Классические 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:

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

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:

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

Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999.
Blei D., Ng A., Jordan M. Latent Dirichlet allocation. 2003.

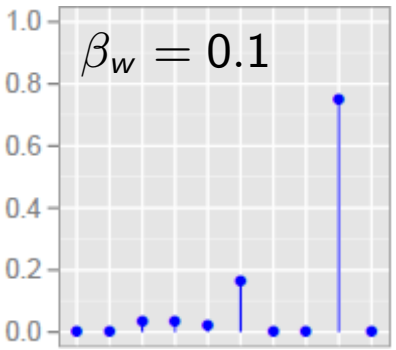
Вероятностная байесовская интерпретация LDA [Blei, 2003]

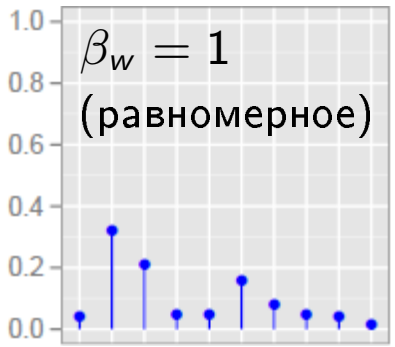
Гипотеза. Вектор-столбцы $\phi_t = (\phi_{wt})_{w \in W}$ и $\theta_d = (\theta_{td})_{t \in T}$ порождаются распределениями Дирихле, $\alpha \in \mathbb{R}^{|T|}$, $\beta \in \mathbb{R}^{|W|}$:

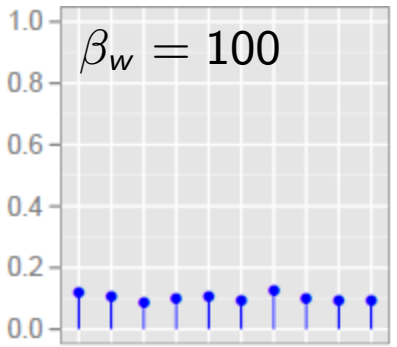
$$\mathrm{Dir}(\phi_t|\beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1}, \quad \phi_{wt} > 0; \quad \beta_0 = \sum_w \beta_w, \ \beta_t > 0;$$

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

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







Максимизация апостериорной вероятности для модели LDA

Совместное правдоподобие данных и модели:

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

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 слабо разреженные оценки:

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

при $\beta_w > 1$, $\alpha_t > 1$ — сглаживание,
при $0 < \beta_w < 1$, $0 < \alpha_t < 0$ — слабое разреживание,
при $\beta_w = 1$, $\alpha_t = 1$ априорное распределение равномерно, PLSA.

Обобщённая не-байесовская интерпретация LDA

Сглаживание распределений по KL-дивергенции:
приблизить $\phi_{wt} \equiv p(w|t)$ к заданным распределениям $\beta_t(w)$,
приблизить $\theta_{td} \equiv p(t|d)$ к заданным распределениям $\alpha_d(t)$:

$$\sum_{t\in T} \tau_t \mathrm{KL}\big(\beta_t(w)\|\phi_{wt}\big) \to \min_{\Phi}; \qquad \sum_{d\in D} \tau_d \mathrm{KL}\big(\alpha_d(t)\|\theta_{td}\big) \to \min_{\Theta}.$$

Взвешенная сумма регуляризаторов:

$$R(\Phi,\Theta) = \sum_{t\in T} \tau_t \sum_{w\in W} \beta_t(w) \ln \phi_{wt} + \sum_{d\in D} \tau_d \sum_{t\in T} \alpha_d(t) \ln \theta_{td}.$$

Формулы M-шага:

$$\phi_{wt} = \underset{w}{\mathrm{norm}}\big(n_{wt} + \underbrace{\tau_t\beta_t(w)}_{\beta_{wt}}\big), \qquad \theta_{td} = \underset{t}{\mathrm{norm}}\big(n_{td} + \underbrace{\tau_d\alpha_d(t)}_{\alpha_{td}}\big).$$

Сглаживание, разреживание и частичное обучение тем

Формулы M-шага (теперь нет ограничений на β_{wt}, α_{td}):

$$\phi_{wt} = \underset{w}{\text{norm}}\left(n_{wt} + \beta_{wt}\right), \qquad \theta_{td} = \underset{t}{\text{norm}}\left(n_{td} + \alpha_{td}\right).$$

Разреживание и сглаживание описывается общей формулой:

- разреживание — максимизация KL, $\beta_{wt} < 0$, $\alpha_{td} < 0$
- сглаживание — минимизация KL, $\beta_{wt} > 0$, $\alpha_{td} > 0$

Частичное обучение темы t:

- $\beta_{wt} = +\tau_{\text{бт}}[w \in W_t]$ — «белый список» терминов
- $\beta_{wt} = -\tau_{\text{чт}}[w \in W_t]$ — «чёрный список» терминов
- $\alpha_{td} = +\tau_{\text{бд}}[d \in D_t]$ — «белый список» документов
- $\alpha_{td} = -\tau_{\text{чд}}[d \in D_t]$ — «чёрный список» документов

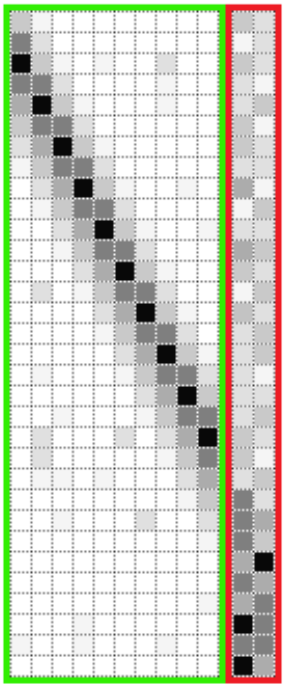
Разделение тем на предметные и фоновые

$T = S \sqcup B$ — множество всех тем

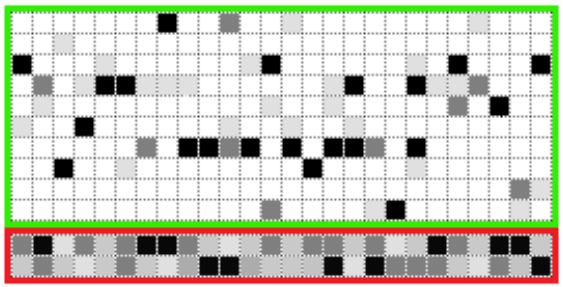
S — разреженные *предметные* темы, специальная лексика

B — сглаженные *фоновые* темы, общая лексика языка

$\Phi_{W \times T}$



$\Theta_{T \times D}$



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:

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

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

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

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

Правдоподобие и перплексия (perplexity)

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

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

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

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

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):

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

$d = d' \sqcup d''$ — 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''.

Пример тем. Мультиязычная модель Википедии

216 175 русско-английских пар статей. Языки — модальности.
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

Дударенко М. А. Регуляризация многоязычных тематических моделей. Вычислительные методы и программирование. 2015. Т. 16. С. 26–36.

Пример тем. Мультиязычная модель Википедии

216 175 русско-английских пар статей. Языки — модальности.
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

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

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	тупиковое множество	комитет	задача MASC

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

Интерпретируемость 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
Wikipedia	RACO	0.62	0.69
	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 тематической модели

Когерентность (согласованность) темы t по k топовым словам:

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

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

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

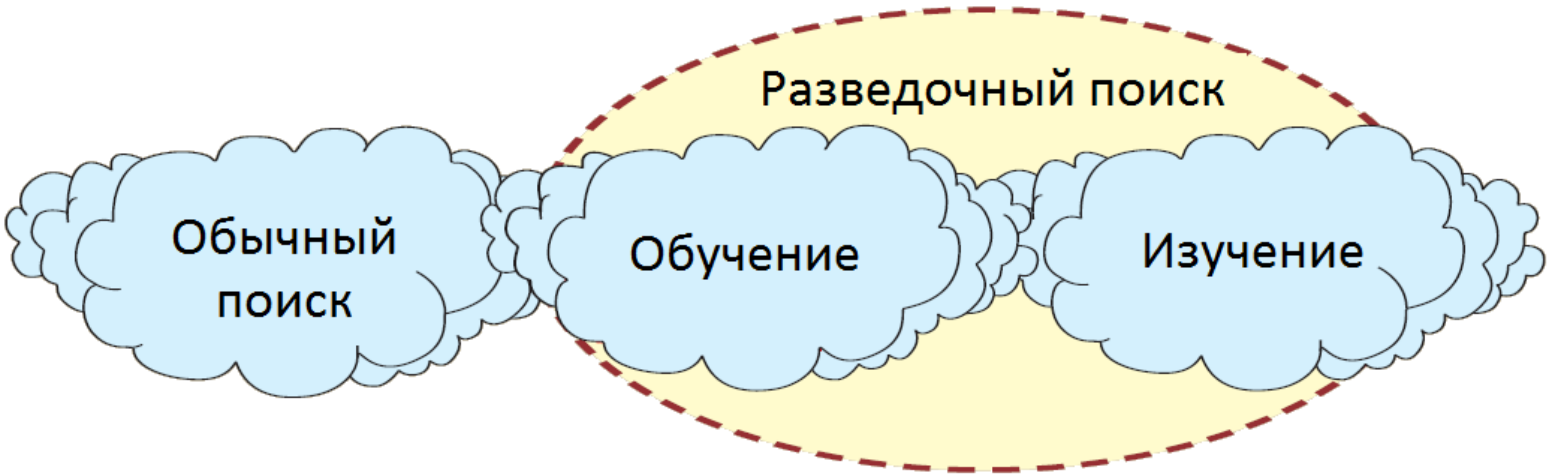
N_{uv} — число документов, в которых термины u, v хотя бы один раз встречаются рядом (в окне 10 слов),

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.

Концепция разведочного поиска (exploratory search)

- пользователь может не знать ключевых терминов,
- запросом может быть текст произвольной длины,
- информационной потребностью — систематизация знаний



навигация в сети,	самообразование,	исследование,
поиск фактов,	тематический поиск	экспертиза,
упоминаний,	систематизация	реферирование,
конкретных ответов	знаний	мониторинг тем

Gary Marchionini. Exploratory Search: from finding to understanding. 2006.

Разведочный тематический поиск

Запрос q — текст произвольной длины

$\theta_{tq} = p(t|q)$ — тематический профиль запроса q
$\theta_{td} = p(t|d)$ — тематические профили документов $d \in D$

Косинусная мера близости документа d и запроса q:

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

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

Выдача тематического поиска — k первых документов.

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

Две коллекции новостей про технологии

Habrahabr.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

Предобработка текстов

- отброшены 5% наиболее частотных слов (общая лексика)
- удаление пунктуации
- нижний регистр, ё→е
- лемматизация pymorphy2

Методика оценивания качества разведочного поиска

Поисковый запрос

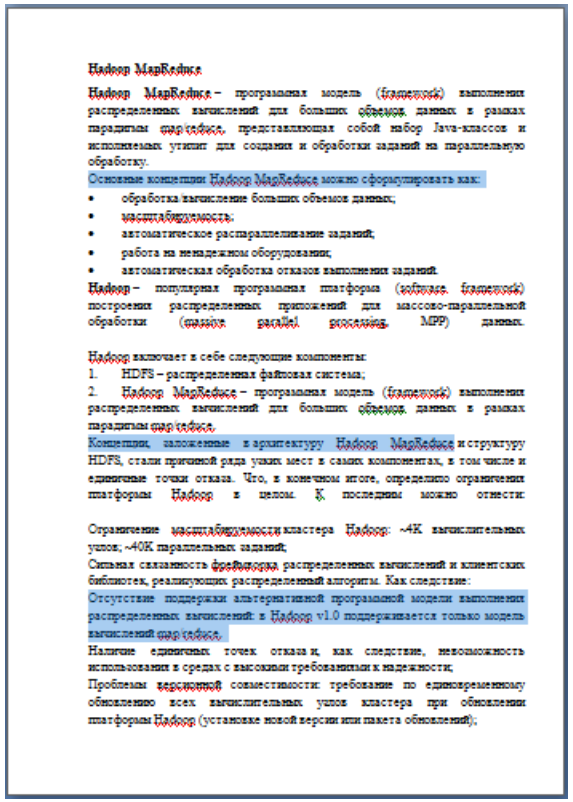
набор ключевых слов или фрагментов текста, около одной страницы A4

Поисковая выдача

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

Два задания асессорам

1. найти как можно больше статей, пользуясь любыми средствами поиска (и засечь время)
2. оценить релевантность поисковой выдачи на том же запросе



Hadoop MapReduce

Hadoop MapReduce – программная модель (framework) выполнения распределенных вычислений для больших объемов данных в рамках парадигмы map/reduce, представляющая собой набор Java-классов и исполняемых утилит для создания и обработки заданий на параллельную обработку.

Основные концепции Hadoop MapReduce можно сформулировать как:

- обработка/вычисление больших объемов данных;
- масштабируемость;
- автоматическое распараллеливание заданий;
- работа на ненадежном оборудовании;
- автоматическая обработка отказов выполнения заданий.

Hadoop – популярная программная платформа (software framework) построения распределенных приложений для массово-параллельной обработки (massive parallel processing, MPP) данных.

Hadoop включает в себе следующие компоненты:

1. HDFS – распределенная файловая система;
2. Hadoop MapReduce – программная модель (framework) выполнения распределенных вычислений для больших объемов данных в рамках парадигмы map/reduce.

Концепции, заложенные в архитектуру Hadoop MapReduce и структуру HDFS, стали причиной ряда узких мест в самих компонентах, в том числе и единичные точки отказа. Что, в конечном итоге, определило ограничения платформы Hadoop в целом. К последним можно отнести:

Ограничение масштабируемости кластера Hadoop: ~4K вычислительных узлов; ~40K параллельных заданий;

Сильная связанность фреймворка распределенных вычислений и клиентских библиотек, реализующих распределенный алгоритм. Как следствие:

Отсутствие поддержки альтернативной программной модели выполнения распределенных вычислений: в Hadoop v1.0 поддерживается только модель вычислений map/reduce.

Наличие единичных точек отказа и, как следствие, невозможность использования в средах с высокими требованиями к надежности;

Проблемы версионной совместимости: требование по единовременному обновлению всех вычислительных узлов кластера при обновлении платформы Hadoop (установке новой версии или пакета обновлений);

Пример запроса для разведочного поиска

Пример: фрагмент запроса «Система IBM Watson»

IBM Watson — суперкомпьютер фирмы IBM, оснащённый вопросно-ответной системой искусственного интеллекта, созданный группой исследователей под руководством Дэвида Феруччи. Его создание — часть проекта DeepQA. Основная задача Уотсона — понимать вопросы, сформулированные на естественном языке, и находить на них ответы в базе данных. Назван в честь основателя IBM Томаса Уотсона.

IBM Watson представляет собой когнитивную систему, которая способна понимать, делать выводы и обучаться. Она также позволяет преобразовывать целые отрасли, различные направления науки и техники. Например, предсказывать появление эпидемий или возникновения очагов природных катастроф в различных регионах, вести мониторинг состояния атмосферы больших городов, оптимизировать бизнес-процессы, узнавать, какие товары будут в тренде в ближайшее время.

... ...

Релевантные тексты: примеры сервисов и приложений, основа которых — когнитивная платформа IBM Watson, используемые в IBM Watson технологии, вопрос-ответные системы, сопоставление IBM Watson с Wolfram-Alpha.

Нерелевантные тексты: общие вопросы искусственного интеллекта, другие коммерческие решения на рынке бизнес-аналитики.

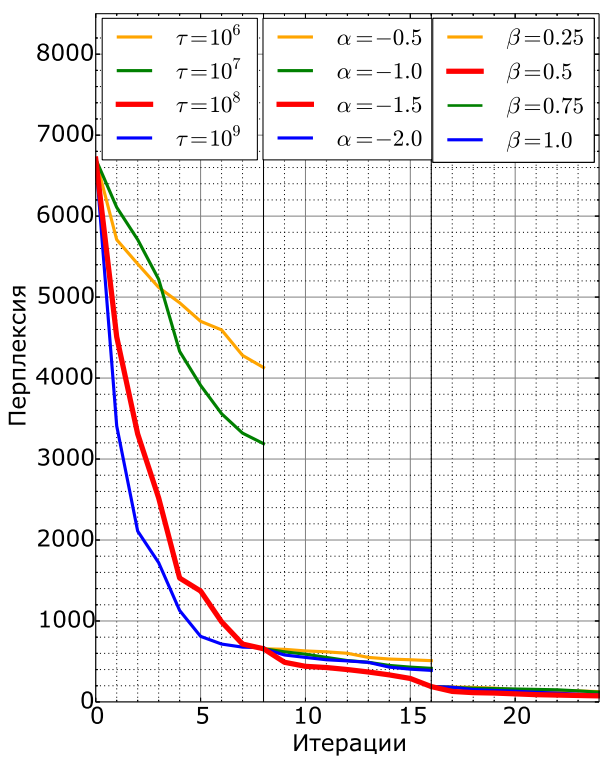
Тематика запросов разведочного поиска

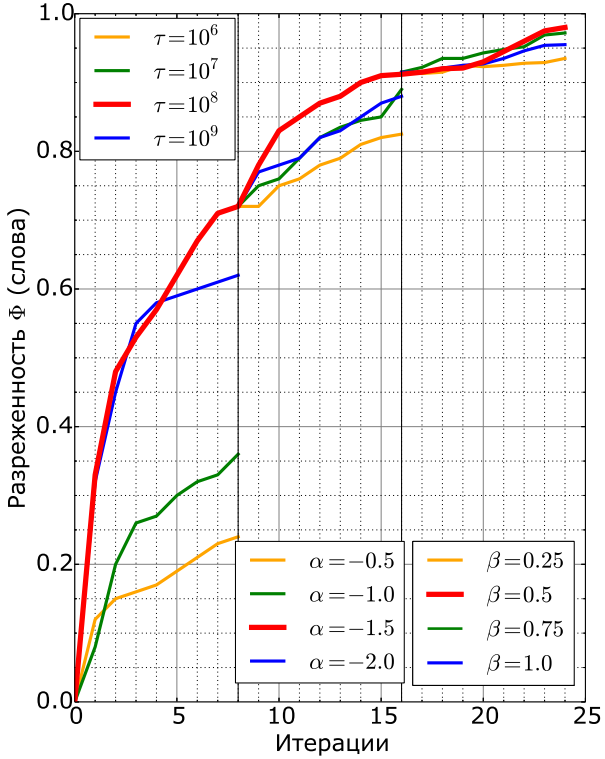
Примеры заголовков разведочных запросов к Хабру (объём каждого запроса — около одной страницы A4):

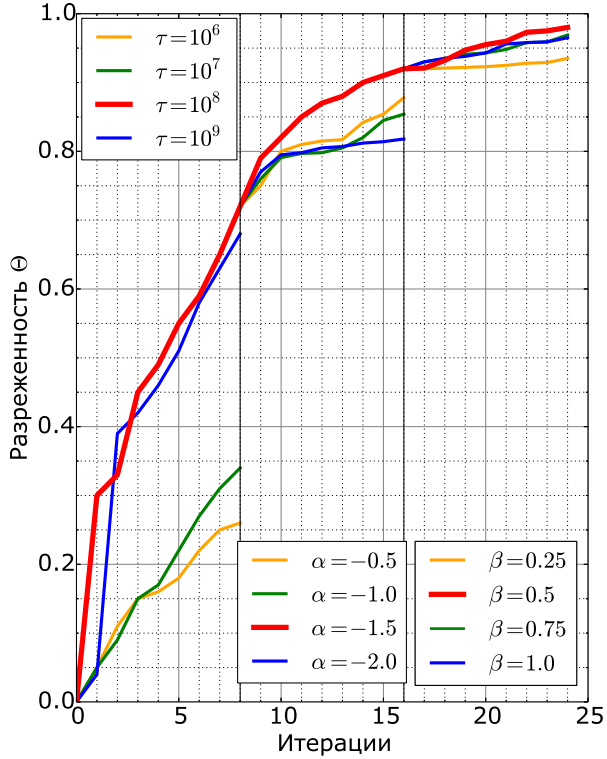
Алгоритмы раскраски графов	Система IBM Watson
Рекомендательная система Netflix	3D-принтеры
Методики быстрого набора текста	CERN-кластер
Космические проекты Илона Маска	AB-тестирование
Технологии Hadoop MapReduce	Облачные сервисы
Беспилотный автомобиль Google car	Контекстная реклама
Криптосистемы с открытым ключом	Марсоход Curiosity
Обзор платформ онлайн-курсов	Видеокарты NVIDIA
Data Science Meetups в Москве	Распознавание образов
Образовательные проекты mail.ru	Сервисы Google scholar
Межпланетная станция New horizons	MIT MediaLab Research
Языковая модель word2vec	Платформа Microsoft Azure

Последовательный подбор коэффициентов регуляризации

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







Оценки качества поиска

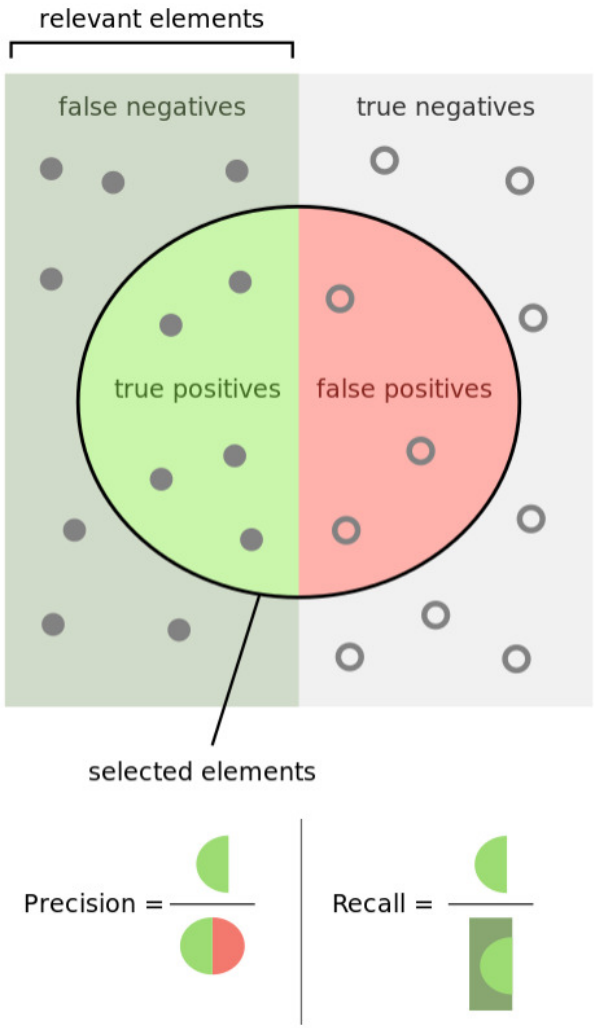
Precision — доля релевантных среди найденных
Recall — доля найденных среди релевантных

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \text{ — точность (precision)}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} \text{ — полнота, (recall)}$$

$$F_1 = \frac{P + R}{2PR} \text{ — F1-мера}$$

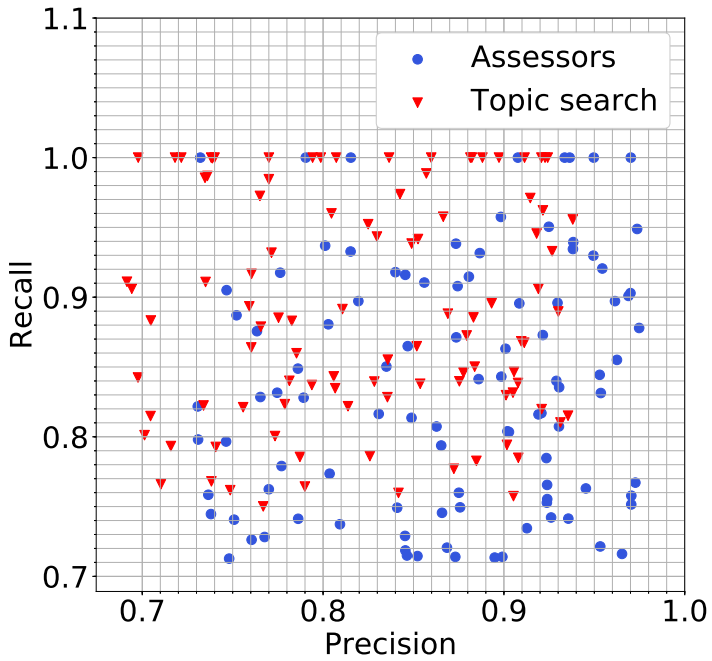
TP (true positive) — найденные релевантные
FP (false positive) — найденные нерелевантные
FN (false negative) — ненайденные релевантные



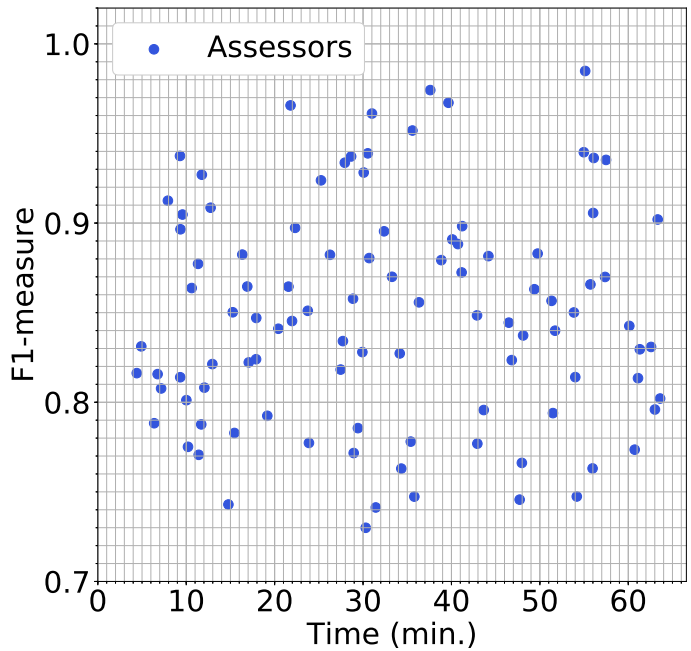
Результаты измерения точности и полноты по запросам

100 запросов, 3 асессора на запрос

точность и полнота поиска



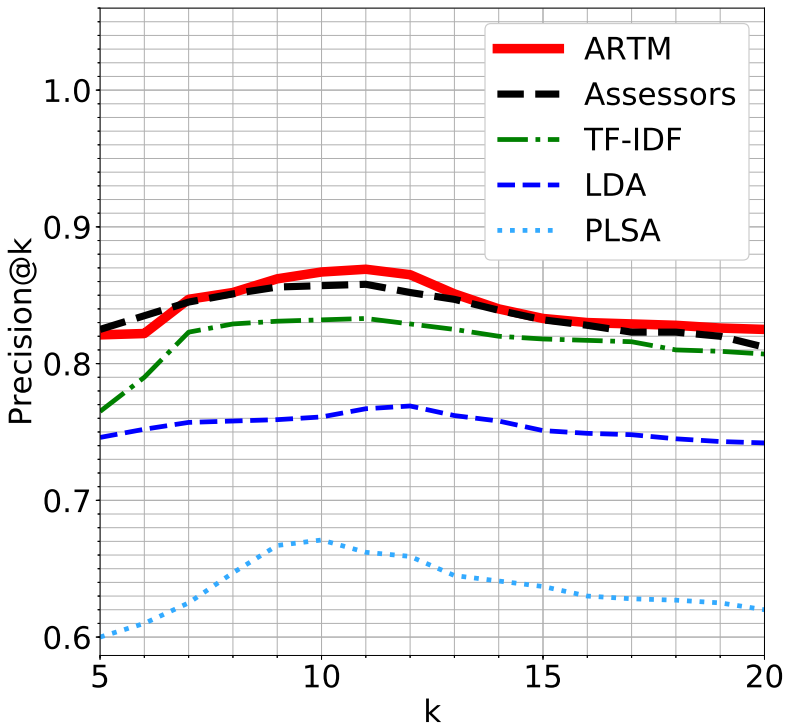
время и F_1-мера (асессоры)

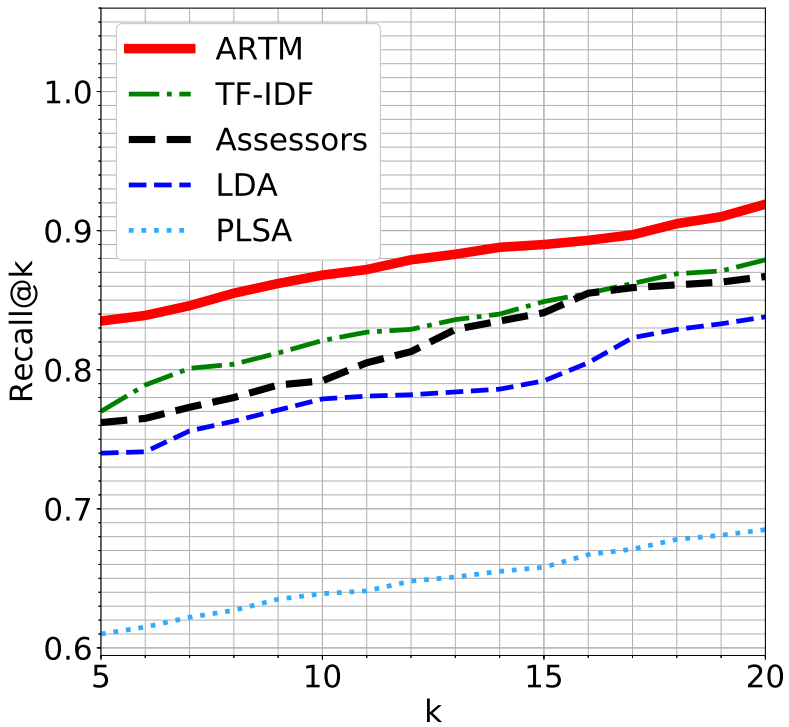


- среднее время обработки запроса асессором — 30 минут
- точность выше у асессоров, полнота — у поисковика

Сравнение с асессорами по качеству поиска

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 меры близости документа и запроса на качество поиска

Меры близости распределений:
Euclidean, Cosine, Manhattan, Hellinger, Kullback–Leibler

	Коллекция Habrahabr.ru					Коллекция TechCrunch.com				
	Eu	**cos**	Ma	He	KL	Eu	**cos**	Ma	He	KL
Prec@5	0.612	**0.810**	0.682	0.709	0.721	0.635	**0.819**	0.673	0.732	0.715
Prec@10	0.657	**0.879**	0.697	0.735	0.749	0.665	**0.867**	0.683	0.752	0.732
Prec@15	0.627	**0.868**	0.635	0.727	0.711	0.643	**0.833**	0.642	0.742	0.724
Prec@20	0.619	**0.847**	0.627	0.728	0.707	0.638	**0.825**	0.638	0.729	0.708
Recall@5	0.672	**0.840**	0.692	0.721	0.803	0.658	**0.835**	0.669	0.733	0.775
Recall@10	0.682	**0.870**	0.707	0.775	0.856	0.671	**0.868**	0.682	0.753	0.787
Recall@15	0.705	**0.891**	0.725	0.791	0.878	0.715	**0.890**	0.708	0.785	0.809
Recall@20	0.703	**0.925**	0.732	0.812	0.888	0.712	**0.919**	0.715	0.808	0.812

- Наилучшее качество поиска — при косинусной мере
- Одни и те же асессорские оценки можно использовать для оценивания новых моделей и поисковых движков

Влияние комбинаций регуляризаторов на качество поиска

Декоррелирование, Θ-разреживание, Φ-сглаживание

	Коллекция Habrahabr.ru				Коллекция TechCrunch.com			
	$R = 0$	Д	ДΘ	ДΘΦ	$R = 0$	Д	ДΘ	ДΘΦ
Prec@5	0.628	0.748	0.771	**0.810**	0.652	0.775	0.779	**0.819**
Prec@10	0.653	0.776	0.812	**0.879**	0.679	0.787	0.819	**0.867**
Prec@15	0.642	0.765	0.792	**0.868**	0.669	0.773	0.798	**0.833**
Prec@20	0.643	0.759	0.783	**0.847**	0.673	0.777	0.792	**0.825**
Recall@5	0.692	0.784	0.805	**0.840**	0.673	0.812	0.812	**0.835**
Recall@10	0.714	0.814	0.834	**0.870**	0.685	0.821	0.845	**0.868**
Recall@15	0.725	0.835	0.867	**0.891**	0.712	0.859	0.869	**0.890**
Recall@20	0.735	0.862	0.891	**0.925**	0.723	0.882	0.895	**0.919**

- Комбинирование регуляризаторов улучшает качество поиска,
- хотя исходно все регуляризаторы нацелены на улучшение интерпретируемости тем и не оптимизируют поиск явно

Влияние сочетания модальностей на качество поиска

Коллекция TechCrunch.com. Число тем $|T| = 450$.
Модальности: Слова, Категории, Биграммы, Авторы.

	асессоры	С	К	СБ	СБК	все
Prec@5	0.822	0.711	0.557	0.767	0.808	**0.819**
Prec@10	0.851	0.721	0.581	0.783	0.818	**0.867**
Prec@15	0.835	0.733	0.594	0.793	0.833	**0.833**
Prec@20	0.813	0.727	0.566	0.772	0.822	**0.825**
Recall@5	0.762	0.752	0.657	0.775	0.825	**0.835**
Recall@10	0.792	0.776	0.669	0.808	0.855	**0.868**
Recall@15	0.835	0.782	0.684	0.825	0.877	**0.890**
Recall@20	0.867	0.825	0.702	0.837	0.901	**0.919**

- Наилучшее качество поиска — по всем модальностям
- Наиболее полезные модальности — слова и категории

Влияние числа тем на качество поиска

Коллекция TechCrunch.com

Используем все 4 модальности, меняем $|T|$

	асессоры	350	400	450	**475**	500
Prec@5	0.822	0.653	0.725	0.752	**0.819**	0.777
Prec@10	0.851	0.663	0.732	0.762	**0.867**	0.811
Prec@15	0.835	0.682	0.743	0.787	**0.833**	0.793
Prec@20	0.813	0.650	0.743	0.773	**0.825**	0.793
Recall@5	0.762	0.731	0.762	0.793	**0.835**	0.817
Recall@10	0.792	0.763	0.793	0.812	**0.868**	0.855
Recall@15	0.835	0.782	0.807	0.855	**0.890**	0.882
Recall@20	0.867	0.792	0.823	0.862	**0.919**	0.903

- Наилучшее качество поиска — при 475 темах
- Тематический поиск превосходит асессоров по полноте

Резюме

- *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 очень разнообразны
- На практике *внешние* (extrinsic) критерии качества модели важнее *внутренних* (intrinsic) — перплексии и когерентности