Математические методы анализа текстов. Тематическое моделирование (часть 2)

К. В. Воронцов, А. А. Потапенко, А. С. Попов, М. А. Апишев,
Р. Ю. Дербаносов, Н. А. Шаталов

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Математические методы анализа текстов
(курс лекций, К.В.Воронцов, А.А.Потапенко)»

7 ноября 2018

Содержание

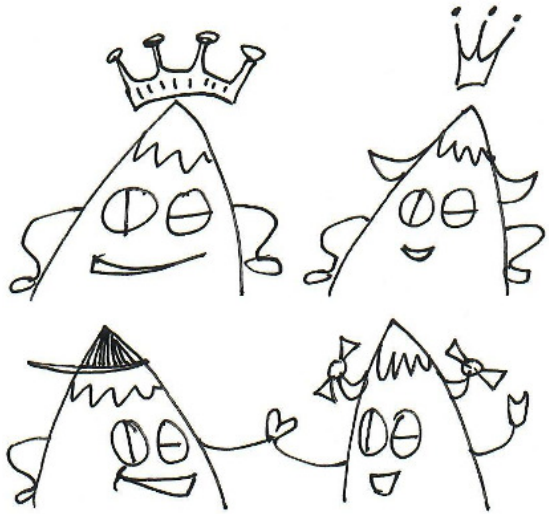
Обобщение №4: модели совстречаемости слов

Проблема

Тематические модели формируют векторные представления (эмбединги) слов, но почему-то они не способны решать задачи семантической близости слов, как word2vec.

Решение

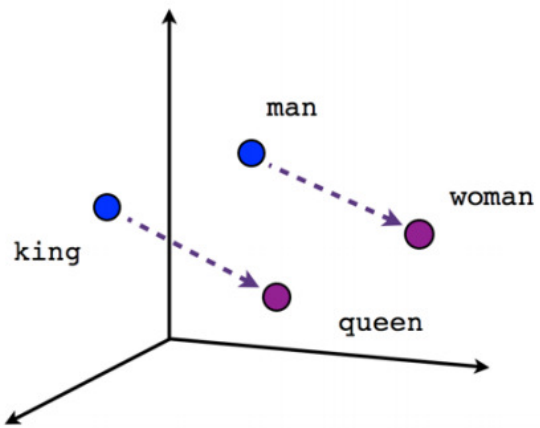
Понять, что такого есть в word2vec, и ввести это в TM.



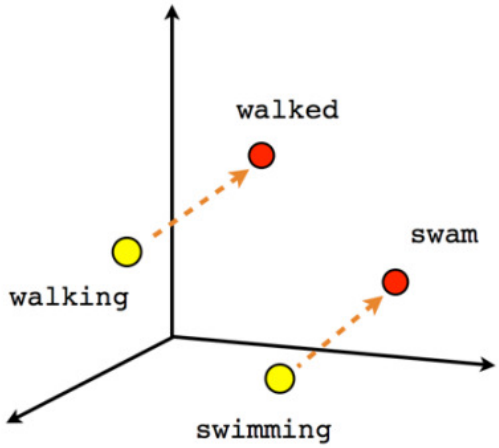
Задача семантического векторного представления слов

Найти для каждого слова w вектор $x_w \in \mathbb{R}^T$, чтобы близкие по смыслу слова имели близкие векторы.

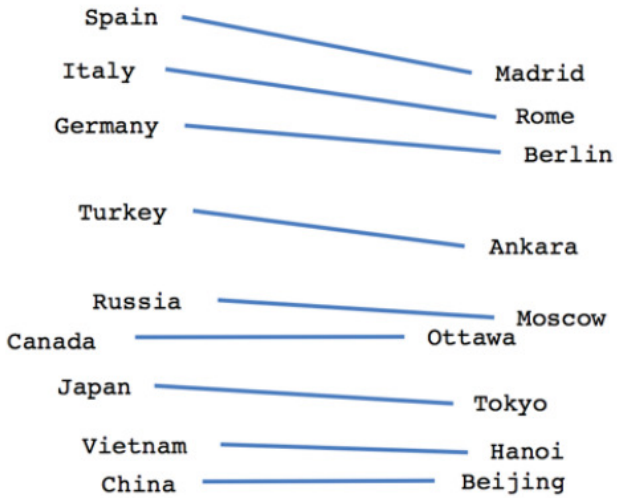
Задача семантической аналогии слов:
по трём словам угадать четвёртое.



Male-Female



Verb tense



Country-Capital

Дистрибутивная гипотеза

- Words that occur in the same contexts tend to have similar meanings [Harris, 1954].
- You shall know a word by the company it keeps [Firth, 1957].

Синтагматическая близость слов:
со-встречаемость слов в одном контексте.

здание–строитель, кран–вода, функция–точка

Парадигматическая близость слов:
взаимозаменяемость слов в одном контексте.

здание–дом, кран–смеситель, функция–отображение

Z.Harris. Distributional structure. 1954.
J.R.Firth. A synopsis of linguistic theory 1930-1955. Oxford, 1957.
P.D.Turney, P.Pantel. From frequency to meaning: Vector space models of semantics // Journal of Artificial Intelligence Research (JAIR). 2010.

Модели векторных представлений для текстов и графов

word2vec: эмбединги слов

T.Mikolov et al. Efficient estimation of word representations in vector space. 2013.

paragraph2vec: эмбединги фрагментов или документов

Q.Le, T.Mikolov. Distributed representations of sentences and documents. 2014.

sent2vec: эмбединги предложений

M.Pagliardini et al. Unsupervised learning of sentence embeddings using compositional n-gram features. 2017.

FastText: эмбединги символьных n-грамм

`https://github.com/facebookresearch/fastText`

node2vec: эмбединги вершин графа

A.Grover, J.Leskovec. Node2vec: scalable feature learning for networks. 2016.

graph2vec: более общие эмбединги на графах

A.Narayanan et al. Graph2vec: learning distributed representations of graphs. 2017.

StarSpace: эмбединги чего угодно от Facebook AI Research

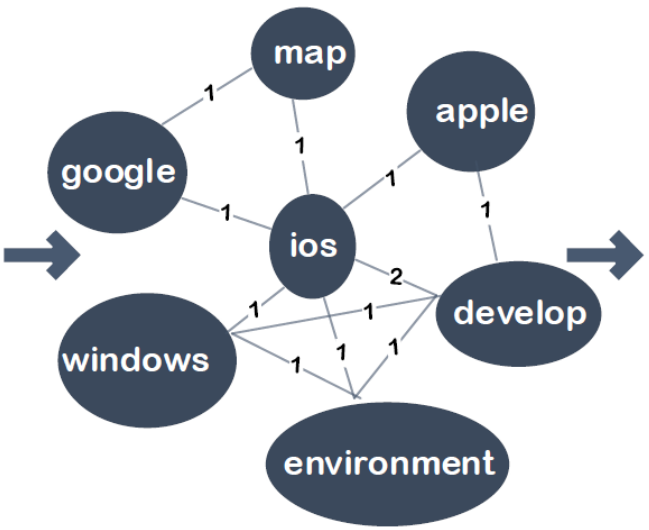
L.Wu, A.Fisch, S.Chopra, K.Adams, A.B.J.Weston. StarSpace: embed all the things! 2018.

Недостаток: координаты векторов не интерпретируемы

Модель сети слов WNTM для коротких текстов

Идея: моделировать не документы, а связи между словами.
d_u — псевдо-документ, объединение всех контекстов слова u.
n_{uw} — число вхождений слова w в псевдо-документ d_u.
Контекст — короткое сообщение / предложение / окно $\pm h$ слов.

doc1	Google Map for iOS
doc2	iOS is developed by Apple
doc3	iOS develop environment for Windows



word	pseudo-document
map	google ios
google	map ios
apple	develop ios
ios	google map apple windows environment develop develop
windows	environment ios develop
develop	apple ios ios windows environment
environment	ios windows develop

Yuan Zuo, Jichang Zhao, Ke Xu. Word Network Topic Model: a simple but general solution for short and imbalanced texts. 2014.

Модели WNTM и WTM (Word Topic Model)

Тематическая модель контекстов, разложение $W \times W$-матрицы:

$$p(w|d_u) = \sum_{t \in T} p(w|t)p(t|d_u) = \sum_{t \in T} \phi_{wt}\theta_{tu},$$

где d_u — псевдо-документ слова u.

Максимизация логарифма правдоподобия:

$$\sum_{u,w \in W} n_{uw} \log \sum_{t \in T} \phi_{wt}\theta_{tu} \to \max_{\Phi,\Theta},$$

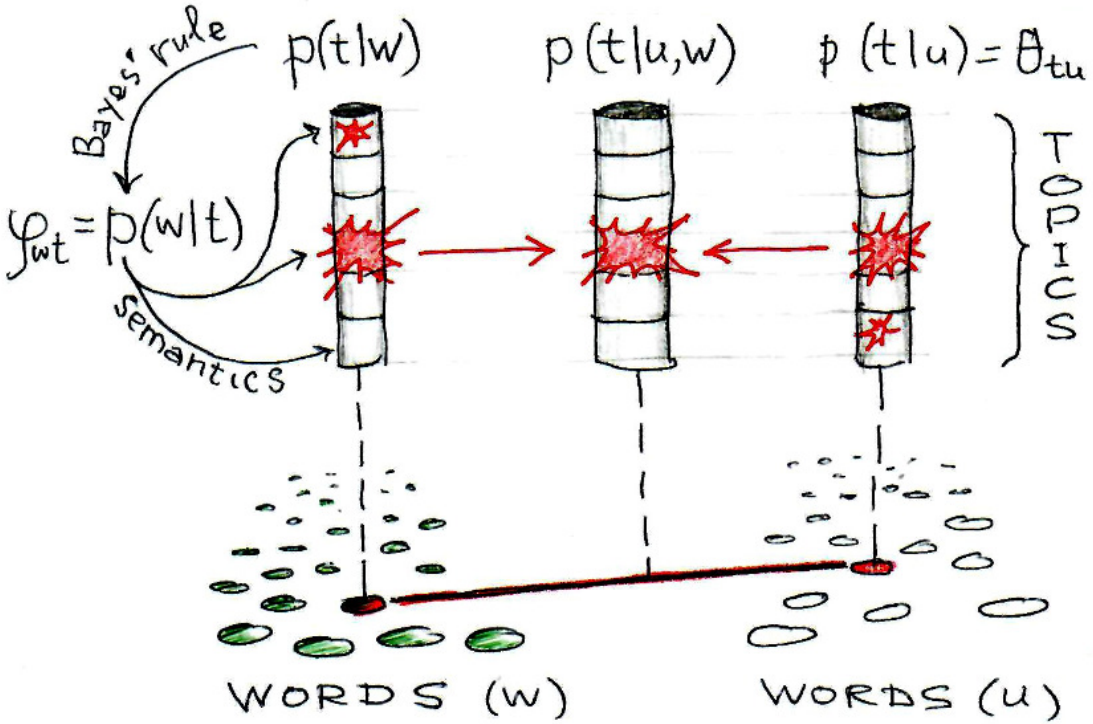
где n_{uw} — совстречаемость слов u, w (кстати, $n_{uw} = n_{wu}$).

Yuan Zuo, Jichang Zhao, Ke Xu. Word Network Topic Model: a simple but general solution for short and imbalanced texts. 2014.

Berlin Chen. Word Topic Models for spoken document retrieval and transcription. ACM Trans., 2009.

Интерпретируемые эмбединги совстречаемости слов

- Идея *дистрибутивной семантики*: “Words that occur in the same contexts tend to have similar meanings” [Harris, 1954].
- Слово индуцирует псевдо-документ всех его контекстов



word2vec и ARTM на задачах аналогии слов

Два подхода к синтезу векторных представлений слов:

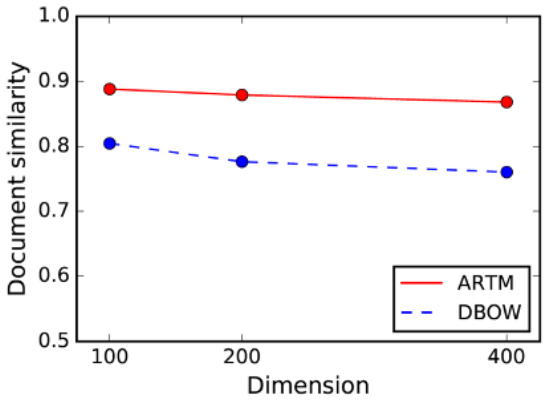
- **ARTM**: интерпретируемые разреженные компоненты
- **word2vec**: интерпретируемые векторные операции

Операция	Результат ARTM	Результат word2vec
king – boy + girl	*queen*, princess, lord, prince	*queen*, princess, regnant, kings
moscow – russia + spain	*madrid*, barcelona, aires, buenos	*madrid*, barcelona, valladolid, malaga
india – russia + ruble	*rupee*, birbhum, pradesh, madhaya	*rupee*, rupiah, devalued, debased
cars – car + computer	*computers*, software, servers, implementations	*computers*, software, hardware, microcomputers

A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL-6, 2017.

word2vec и ARTM в задаче семантической близости документов

ArXiv triplets dataset [Dai et. al, 2015]: 20K троек статей:
⟨ статья A, схожая статья B, непохожая статья C ⟩



- обучение по 1M текстов статей ArXiv
- тестирование на триплетах ArXiv
- Конкурент DBOW: paragraph2vec [Dai et. al, 2015]

ARTM превосходит модель DBOW (distributed bag-of-words).

Andrew Dai, Cristopher Olah, Quoc Le. Document Embedding with Paragraph Vectors, CoRR, 2015

A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL-6, 2017.

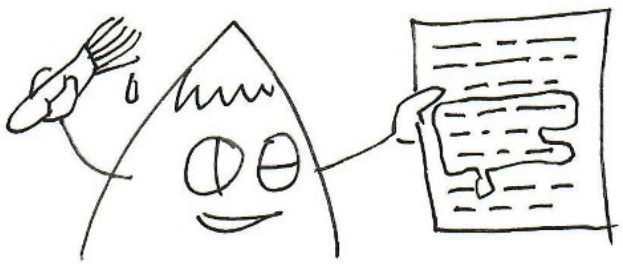
Обобщение №5: модели с регуляризацией E-шага

Проблема

Гипотеза «мешка слов» — самое часто критикуемое допущение тематического моделирования.
Как строить модели, учитывающие порядок слов?

Решение

Пост-обработка $p(t|d, w_i)$ как пучка временных рядов, например, сглаживание или сегментирование, с учётом предложений, секционирования, синтаксических связей, лексических цепочек, и т. д.



Сегментная структура текста и пост-обработка E-шага

Документ $d = \{w_1, \ldots, w_{n_d}\}$, n_d — длина документа d

Матрица *тематики слов в документах* $p(t|d, w_i)$ размера $T \times n_d$:

Регуляризация E-шага

Трёхмерная матрица $\Pi = \left(p_{tdw} = p(t|d,w)\right)_{T\times D\times W}$

Максимизация log правдоподобия с регуляризаторами R и $\tilde{R}$:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} + R(\Pi(\Phi,\Theta)) + \tilde{R}(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}.$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:}\quad \begin{cases} p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right) \\ \tilde{p}_{tdw} = p_{tdw}\left(1 + \frac{1}{n_{dw}}\left(\frac{\partial R(\Pi)}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R(\Pi)}{\partial p_{zdw}}\right)\right) \\ \text{M-шаг:}\quad \phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum_{d\in D} n_{dw}\tilde{p}_{tdw} + \phi_{wt}\frac{\partial \tilde{R}}{\partial \phi_{wt}}\right) \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum_{w\in d} n_{dw}\tilde{p}_{tdw} + \theta_{td}\frac{\partial \tilde{R}}{\partial \theta_{td}}\right) \end{cases}$$

Набросок доказательства: три леммы

Лемма 1. Для функции $p_{tdw}(\Phi,\Theta) = \frac{\phi_{wt}\theta_{td}}{\sum_z \phi_{wz}\theta_{zd}}$ и любого $z \in T$

$$\phi_{wt}\frac{\partial p_{zdw}}{\partial \phi_{wt}} = \theta_{td}\frac{\partial p_{zdw}}{\partial \theta_{td}} = p_{tdw}\big([z=t] - p_{zdw}\big).$$

Введём функцию от вспомогательных переменных Π:

$$Q_{tdw}(\Pi) = \frac{\partial R(\Pi)}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R(\Pi)}{\partial p_{zdw}}.$$

Лемма 2. Если $R(\Pi)$ не зависит от p_{tdw} при $w \notin d$, то

$$\phi_{wt}\frac{\partial R(\Pi)}{\partial \phi_{wt}} = \sum_{d\in D} p_{tdw}Q_{tdw}(\Pi); \qquad \theta_{td}\frac{\partial R(\Pi)}{\partial \theta_{td}} = \sum_{w\in d} p_{tdw}Q_{tdw}(\Pi).$$

Лемма 3. Формулы M-шага:

$$\phi_{wt} = \operatorname*{norm}_{w\in W}\Big(\sum_{d\in D} n_{dw}p_{tdw} + \sum_{d\in D} Q_{tdw}p_{tdw} + \phi_{wt}\frac{\partial \tilde{R}}{\partial \phi_{wt}}\Big);$$

$$\theta_{td} = \operatorname*{norm}_{t\in T}\Big(\sum_{w\in d} n_{dw}p_{tdw} + \sum_{w\in d} Q_{tdw}p_{tdw} + \theta_{td}\frac{\partial \tilde{R}}{\partial \theta_{td}}\Big).$$

Гипотеза: пост-обработка E-шага — это неявная регуляризация

Между E- и M-шагом добавляется обработка матрицы $p_{tdw} = p(t|d,w)$ тематики слов документа:

$$\tilde{p}_{tdw} = p_{tdw}\left(1 + \frac{1}{n_{dw}}\left(\frac{\partial R(\Pi)}{\partial p_{tdw}} - \sum_{z \in T} p_{zdw}\frac{\partial R(\Pi)}{\partial p_{zdw}}\right)\right) \qquad (1)$$

Пост-обработка E-шага позволяет учитывать порядок слов в каждом документе в обход гипотезы «мешка слов».

Гипотеза

Любое «разумное» преобразование $p_{tdw} \to \tilde{p}_{tdw}$ эквивалентно некоторому регуляризатору $R(\Pi(\Phi,\Theta))$.

Открытый вопрос: при каких условиях по заданным p_{tdw} и $\tilde{p}_{tdw}$ возможно подобрать функцию $R(\Pi)$ так, чтобы выполнялось уравнение пост-обработки (1)?

Пример 1. Кросс-энтропийное разреживание $p(t|d,w)$

Путь каждый термин относится к небольшому числу тем:

$$\mathrm{KL}\Big(\frac{1}{|T|} \,\Big\|\, p(t|d,w)\Big) \to \max.$$

Суммируем по всем терминам всех документов:

$$R(\Pi) = -\frac{\tau}{|T|}\sum_{d\in D}\sum_{w\in d} n_{dw}\sum_{t\in T}\ln p_{tdw} \to \max.$$

Формула регуляризованного E-шага:

$$\tilde{p}_{tdw} = p_{tdw} - \tau\Big(\frac{1}{|T|} - p_{tdw}\Big).$$

Интерпретация: Если $p_{tdw} < \frac{1}{|T|}$, то p_{tdw} станет ещё меньше.
Тематика термина концентрируется в небольшом числе тем.

Недостаток: Тематика соседних слов разреживается независимо.

Пример 2. Тематическая модель сегментированного текста

S_d — множество микро-сегментов документа d

n_{sw} — число вхождений слова w в сегмент s длины n_s

Тематика сегмента $s \in S_d$ — средняя тематика его слов:

$$p_{tds} \equiv p(t|d,s) = \frac{1}{n_s} \sum_{w \in s} n_{sw} p_{tdw}.$$

Кросс-энтропийный регуляризатор разреживания $p(t|d,s)$:

$$R(\Pi) = -\sum_{d \in D} \sum_{s \in S_d} \sum_{t \in T} \ln \sum_{w \in s} n_{sw} p_{tdw} \to \max.$$

Формула регуляризованного E-шага:

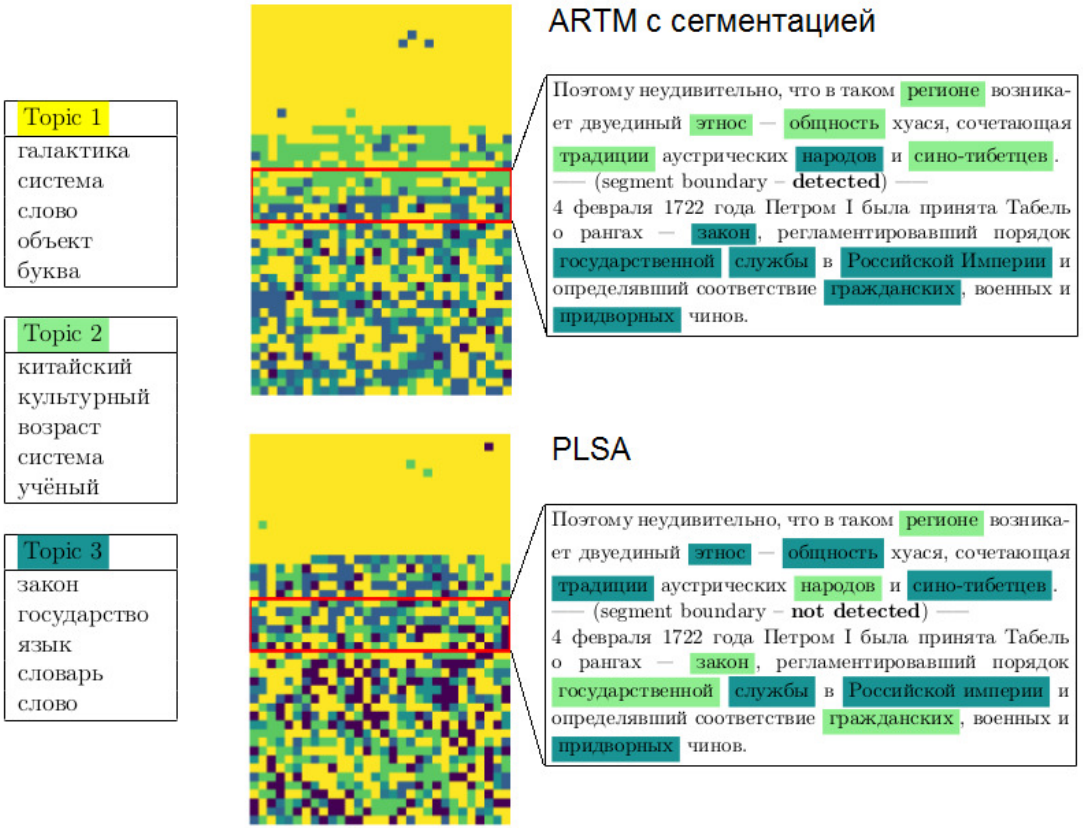
$$\tilde{p}_{tdw} = p_{tdw} \left(1 - \frac{\tau}{n_{dw}} \sum_{s \in S_d} \frac{n_{sw}}{n_s} \left(\frac{1}{p_{tds}} - \sum_{z \in T} \frac{p_{zdw}}{p_{zds}}\right)\right).$$

Интерпретация: если $p_{tds} < \frac{1}{|T|}$, то p_{tdw} уменьшатся $\forall\, w \in s$.

Тематика сегмента концентрируется в небольшом числе тем.

Пример. Регуляризатор E-шага для сегментации текста

Полусинтетическая коллекция из фрагментов postnauka.ru



N.Skachkov, K.Vorontsov. Improving topic models with segmental structure of texts. Dialogue, 2018.

Обобщение №6: модели транзакционных данных

Проблема

Исходные данные могут быть сложнее, чем парные взаимодействия (транзакции) между объектами

Решение

Тематическая модель должна описывать транзакции, состоящие из любых подмножеств объектов

Транзакционные данные

Выборка может содержать не только пары (d, w), но также тройки, четвёрки, ..., n-ки элементов разных модальностей.

Примеры:

- **Данные социальной сети:**
  (d, u, w) — пользователь u записал слово w в блоге d
- **Данные сети интернет-рекламы:**
  (u, d, b) — пользователь u кликнул баннер b на странице d
- **Данные рекомендательной системы:**
  (u, f, s) — пользователь u оценил фильм f в ситуации s
- **Данные финансовых организаций:**
  (b, s, g) — покупатель u купил у продавца s товар g

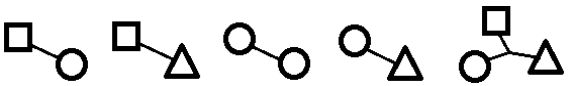
Задача: по наблюдаемой выборке рёбер гиперграфа выявить латентные темы его вершин.

Тематическая модель гиперграфа: определения и обозначения

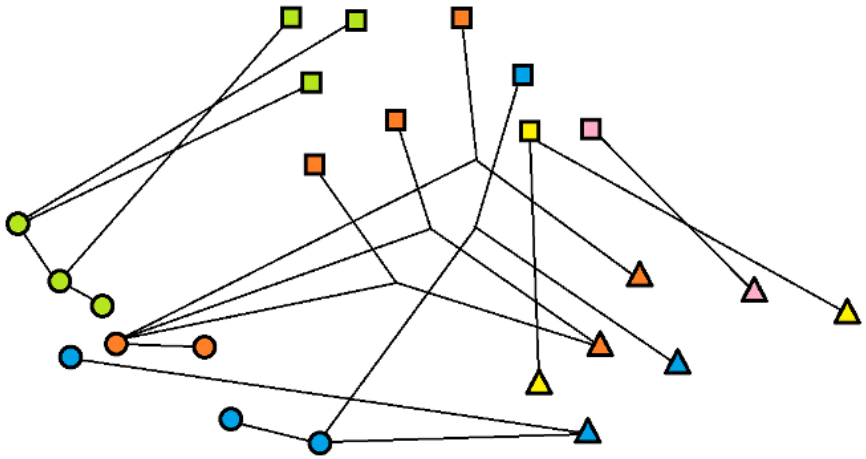
$\Gamma = \langle V, E \rangle$ — ориентированный гиперграф.
$V = V^1 \sqcup \cdots \sqcup V^M$ — разбиение вершин по модальностям

M — множество модальностей:

K — множество типов рёбер:



T — множество тем:



X^k — наблюдаемая выборка транзакций — рёбер типа k
ребро (d, x): вершина-контейнер $d \in V$ и вершины $x \subset V$,
n_{dx} — число вхождений ребра (d, x) в выборку X^k
$p(d, x)$ — неизвестное распределение на рёбрах типа k

Тематическая модель гиперграфа

Вероятностная тематическая модель рёбер типа k:

$$p(x|d) = \sum_{t \in T} \theta_{td} \prod_{v \in x} \phi_{vt},$$

$\theta_{td} = p(t|d)$ — тематика контейнера не зависит от типа ребра k
$\phi_{vt} = p(v|t)$ — распределение термов модальности v в теме t

Задача максимизации log правдоподобия:

$$\sum_{k \in K} \tau_k \sum_{(d,x) \in X^k} n_{dx} \ln \sum_{t \in T} \theta_{td} \prod_{v \in x} \phi_{vt} \to \max_{\Phi, \Theta},$$

$$\phi_{vt} \geqslant 0, \quad \sum_{v \in V^m} \phi_{vt} = 1; \qquad \theta_{td} \geqslant 0, \quad \sum_{t \in T} \theta_{td} = 1;$$

где $\tau_k > 0$ — веса типов рёбер.

EM-алгоритм для гиперграфовой ARTM

Задача максимизации регуляризованного правдоподобия:

$$\sum_{k\in K}\tau_k \sum_{(d,x)\in X^k} n_{dx} \ln \sum_{t\in T} \theta_{td} \prod_{v\in x} \phi_{vt} \;+\; R(\Phi,\Theta) \to \max_{\Phi,\Theta}.$$

EM-алгоритм: метод простой итерации для системы уравнений со вспомогательными переменными $p_{tdx} = p(t|d,x)$:

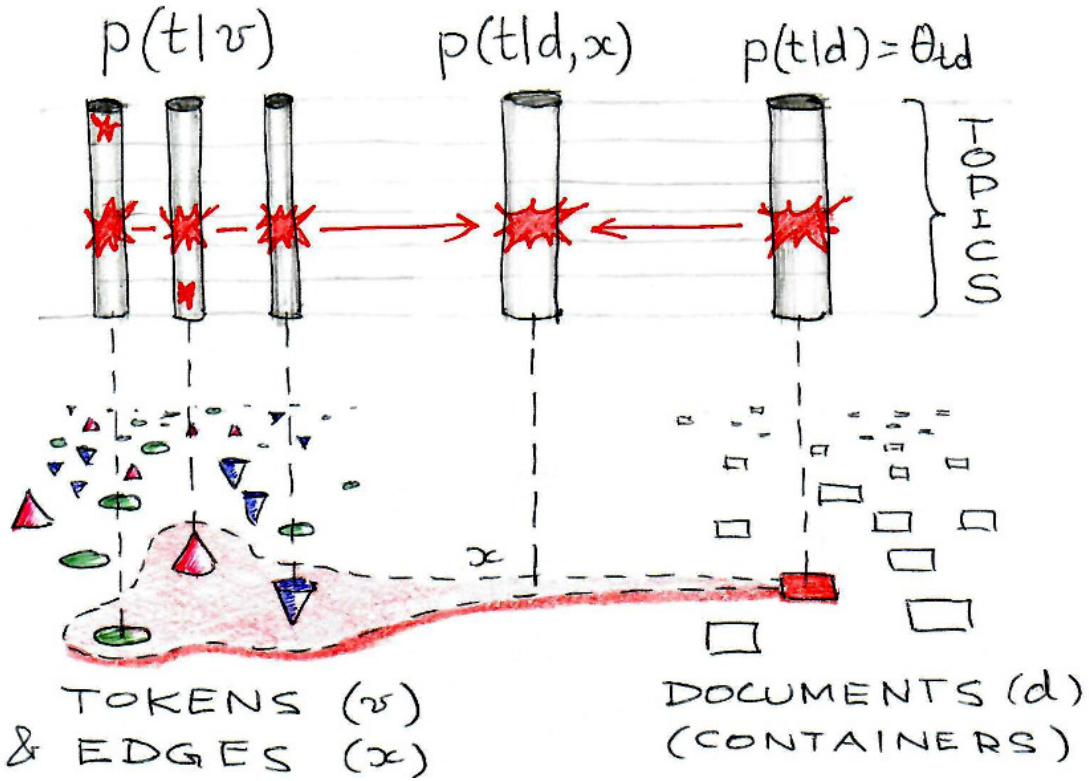
E-шаг:
$$p_{tdx} = \underset{t\in T}{\operatorname{norm}}\Big(\theta_{td} \prod_{v\in x} \phi_{vt}\Big)$$

M-шаг:
$$\phi_{vt} = \underset{v\in V^m}{\operatorname{norm}}\Big(\sum_{k\in K}\tau_k \sum_{(d,x)} \big[v\in x\big] n_{dx} p_{tdx} + \phi_{vt}\frac{\partial R}{\partial \phi_{vt}} \Big)$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\Big(\sum_{k\in K}\tau_k \sum_{(d,x)} n_{dx} p_{tdx} + \theta_{td}\frac{\partial R}{\partial \theta_{td}} \Big)$$

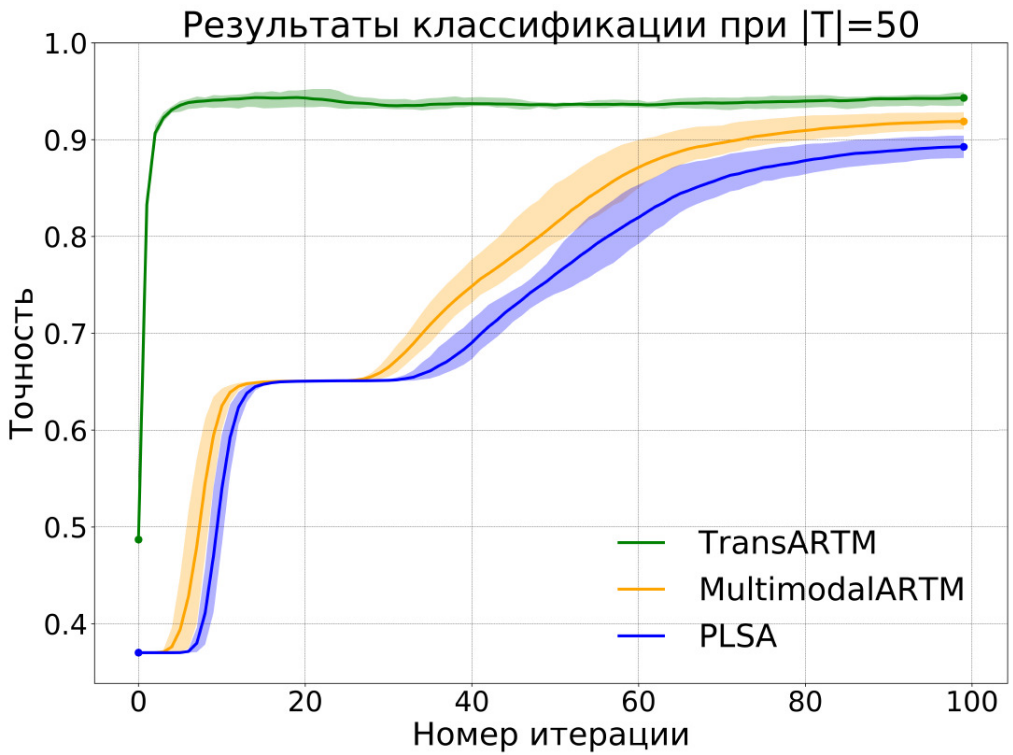
Интерпретируемые эмбединги транзакционных данных

- *Гиперграф* — множество подмножеств вершин-токенов
- Транзакция = подмножество токенов = ребро гиперграфа
- Транзакция происходит, когда токены имеют общие темы



Эксперименты на модельных данных

13M транзакций, 3 модальности, 5 классов, 9 типов рёбер



Вывод: обычные модели не могут восстановить гиперграф.

Илья Кариков. Гиперграфовые тематические модели транзакционных данных. Магистерская диссертация, МФТИ, 2018.

Модели предложений и коротких текстов TwitterLDA, senLDA

S_d — множество предложений документа d

n_{sw} — сколько раз терм w встречается в предложении s

Тематическая модель предложения s:

$$p(s|d) = \sum_{t\in T} p(t|d) \prod_{w\in s} p(w|t)^{n_{sw}} = \sum_{t\in T} \theta_{td} \prod_{w\in s} \phi_{wt}^{n_{sw}}$$

Максимизация регуляризованного log-правдоподобия

$$\sum_{d\in D} \sum_{s\in S_d} \ln \sum_{t\in T} \theta_{td} \prod_{w\in s} \phi_{wt}^{n_{sw}} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

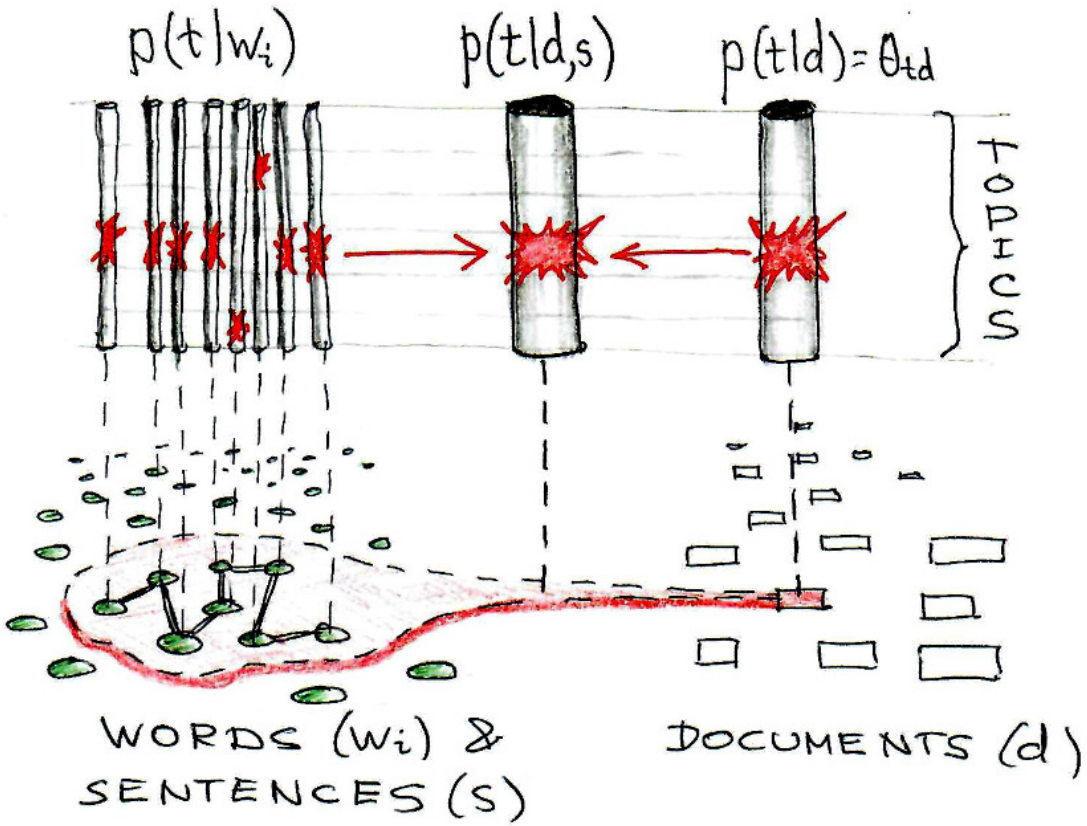
это частный случай гиперграфовой модели, в которой предложения являются «транзакциями» или гипер-рёбрами.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee Peng Lim et al. Comparing Twitter and traditional media using topic models. ECIR 2011.

G.Balikas, M.-R.Amini, M.Clausel. On a topic model for sentences. SIGIR 2016.

Интерпретируемые эмбединги предложений

- Предложение — семантически однородная единица языка
- Предложение образуется из слов, имеющих общие темы
- Предложение = подмножество слов = ребро гиперграфа



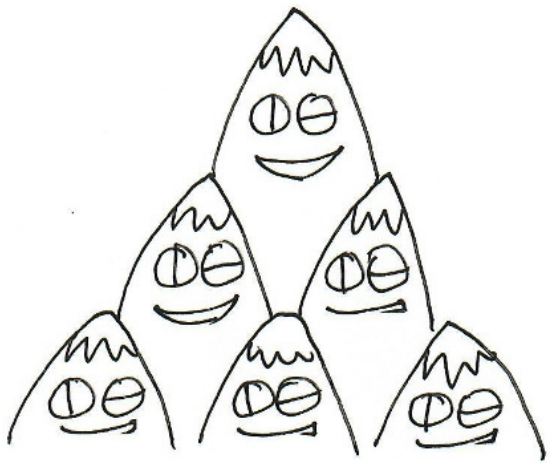
Обобщение №7: иерархические модели

Проблема

Невозможно определить оптимальное число тем.
Хотелось бы разделять темы на подтемы иерархически.
Придумано много иерархических моделей, но они либо ограниченные, либо тормозные, либо замороченные.

Решение

Придумать что-то радикально простое



Послойное построение уровней тематической иерархии

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:

$$\sum_{t\in T} n_t \,\mathrm{KL}_w\Big(p(w|t) \,\Big\|\, \sum_{s\in S} p(w|s)p(s|t)\Big) \;\to\; \min_{\Phi,\Psi},$$

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

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

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

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

Пример тематической иерархии

Тексты научно-просветительского ресурса Postnauka.ru:
2976 документов, 43196 слов, 1799 тэгов

N.A.Chirkova, K.V.Vorontsov. Additive Regularization for Hierarchical Multimodal Topic Modeling. JMLDA, 2016.

A.V.Belyy, M.S.Seleznova, A.K.Sholokhov, K.V.Vorontsov. Quality Evaluation and Improvement for Hierarchical Topic Modeling. Dialogue 2018.

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''.

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
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 как внутренняя мера интерпретируемости

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),

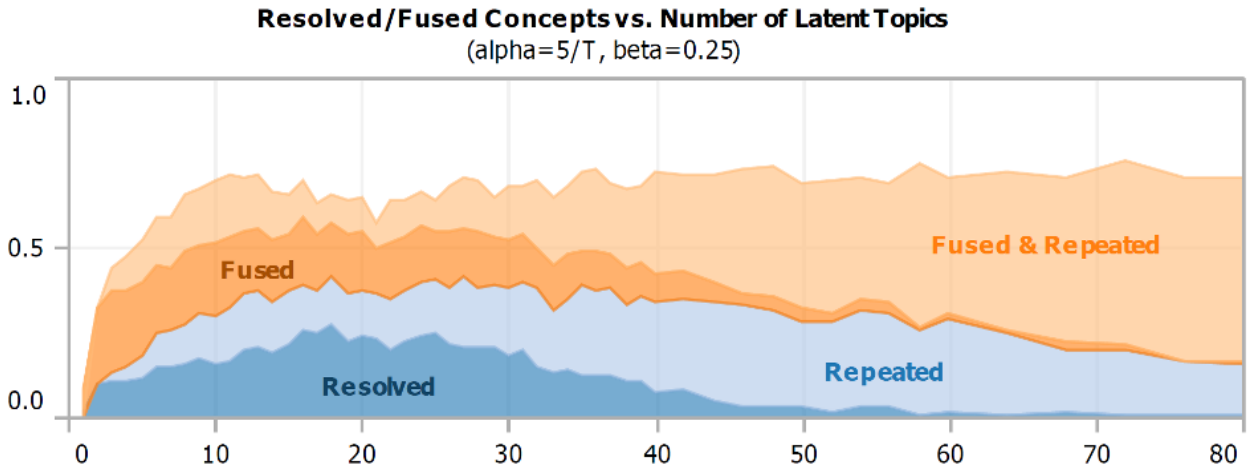
N_{uv} — число документов, в которых термины u, v хотя бы один раз встречаются рядом (в окне 10 слов),

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.

Внешние критерии качества

- Полнота и точность тематического поиска
- Качество классификации / сегментации / суммаризации
- Экспертное оценивание тем *методом интрузий*
- Точность соответствия тем заданным *концептам*
  (число ненайденных и расщеплённых тем и концептов)



Chuang J., Gupta S., Manning C., Heer J. Topic Model Diagnostics: Assessing Domain Relevance via Topical Alignment. ICML-2013.

Методика оценивания качества разведочного поиска

Поисковый запрос

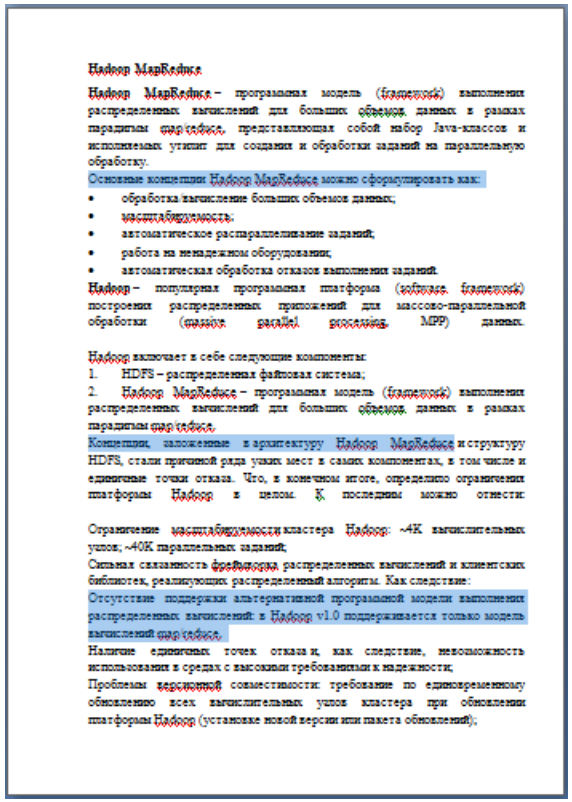
набор ключевых слов или фрагментов текста, около одной страницы A4

Поисковая выдача

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

Два задания асессорам

- найти как можно больше статей, пользуясь любыми средствами поиска (и засечь время)
- 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

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)

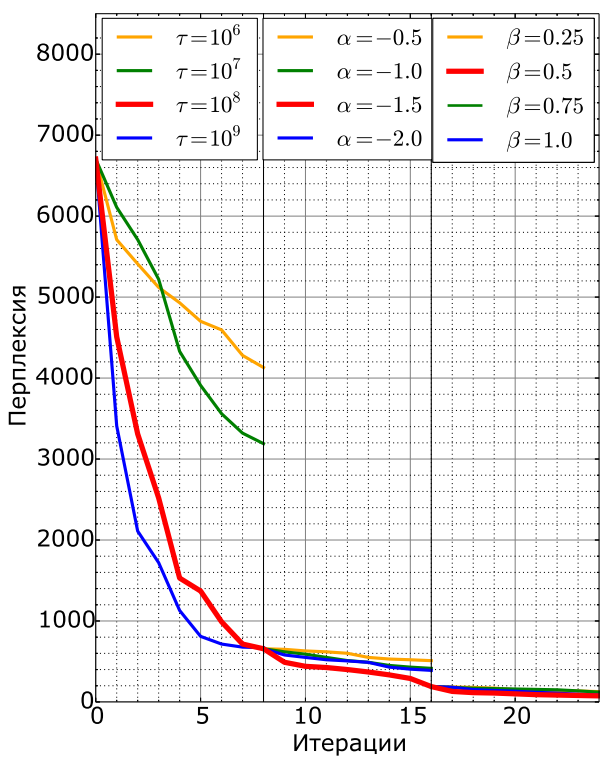
TechCrunch.com

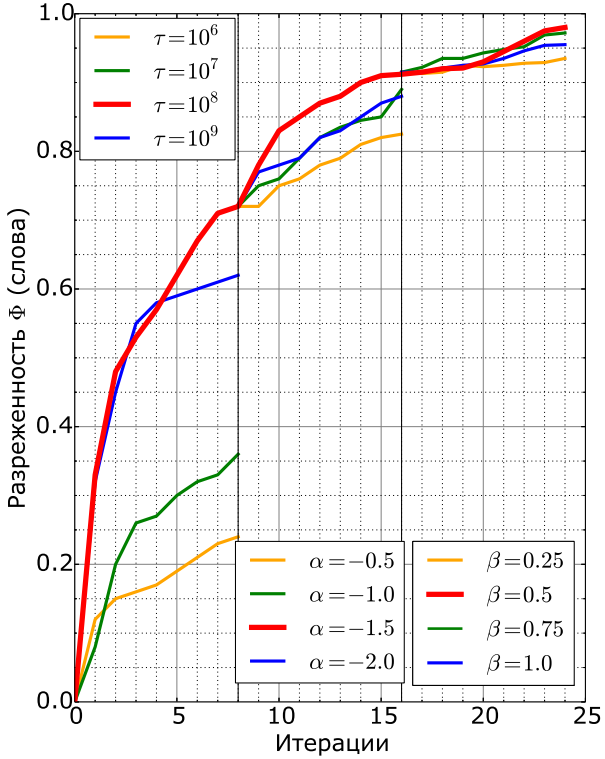
759 324 статей на английском
11 523 слов (униграмм)
1.2 млн. биграмм
605 авторов
184 категорий

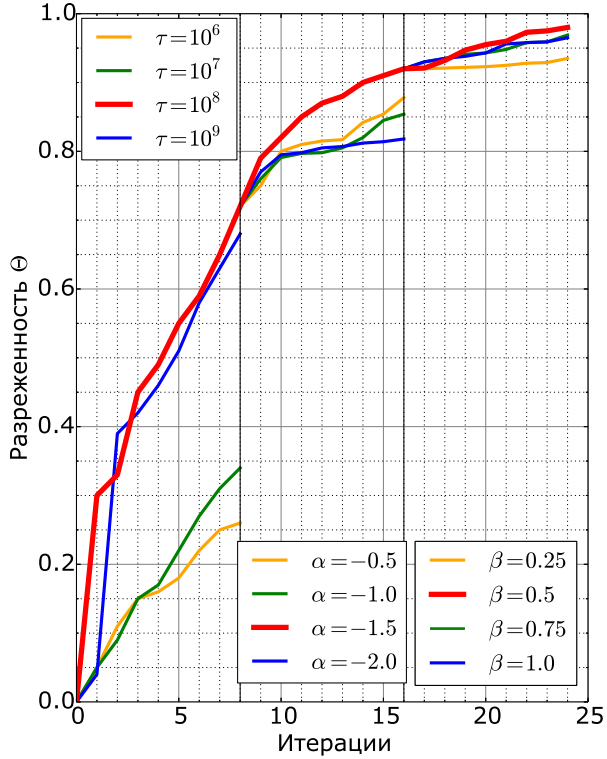
TechCrunch

Последовательный подбор коэффициентов регуляризации

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

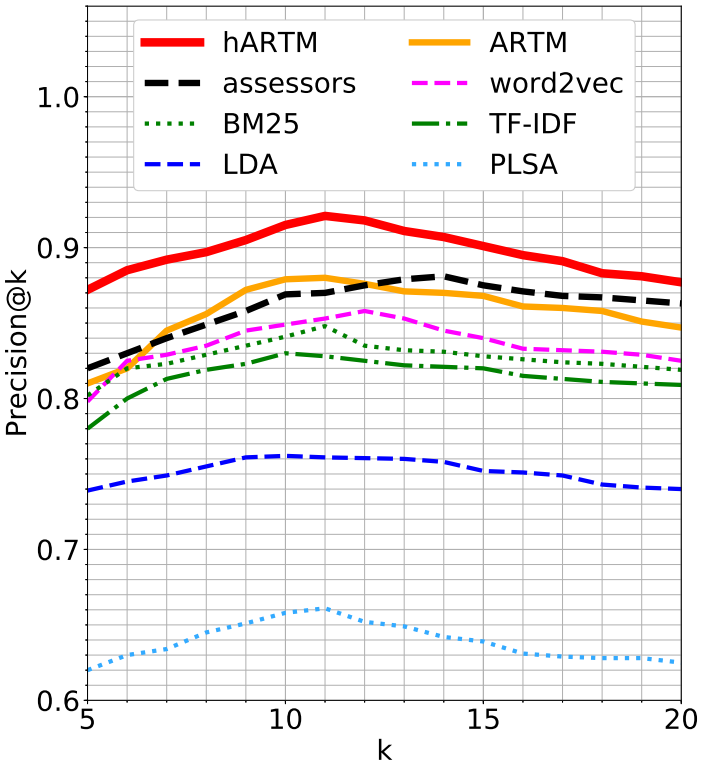


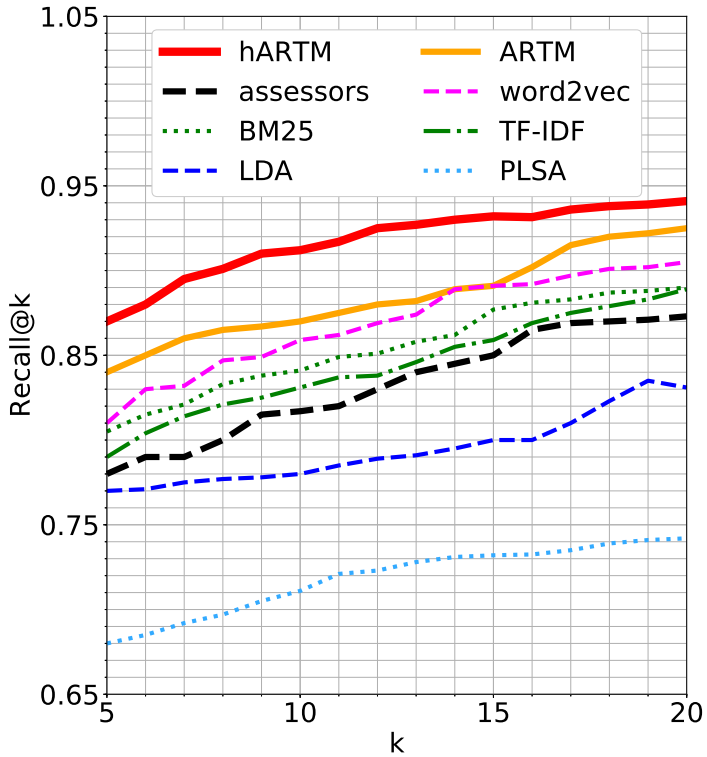




Сравнение качества поиска с асессорами и простыми моделями

Точность и полнота по первым k позициям поисковой выдачи (коллекция Habrahabr.ru)

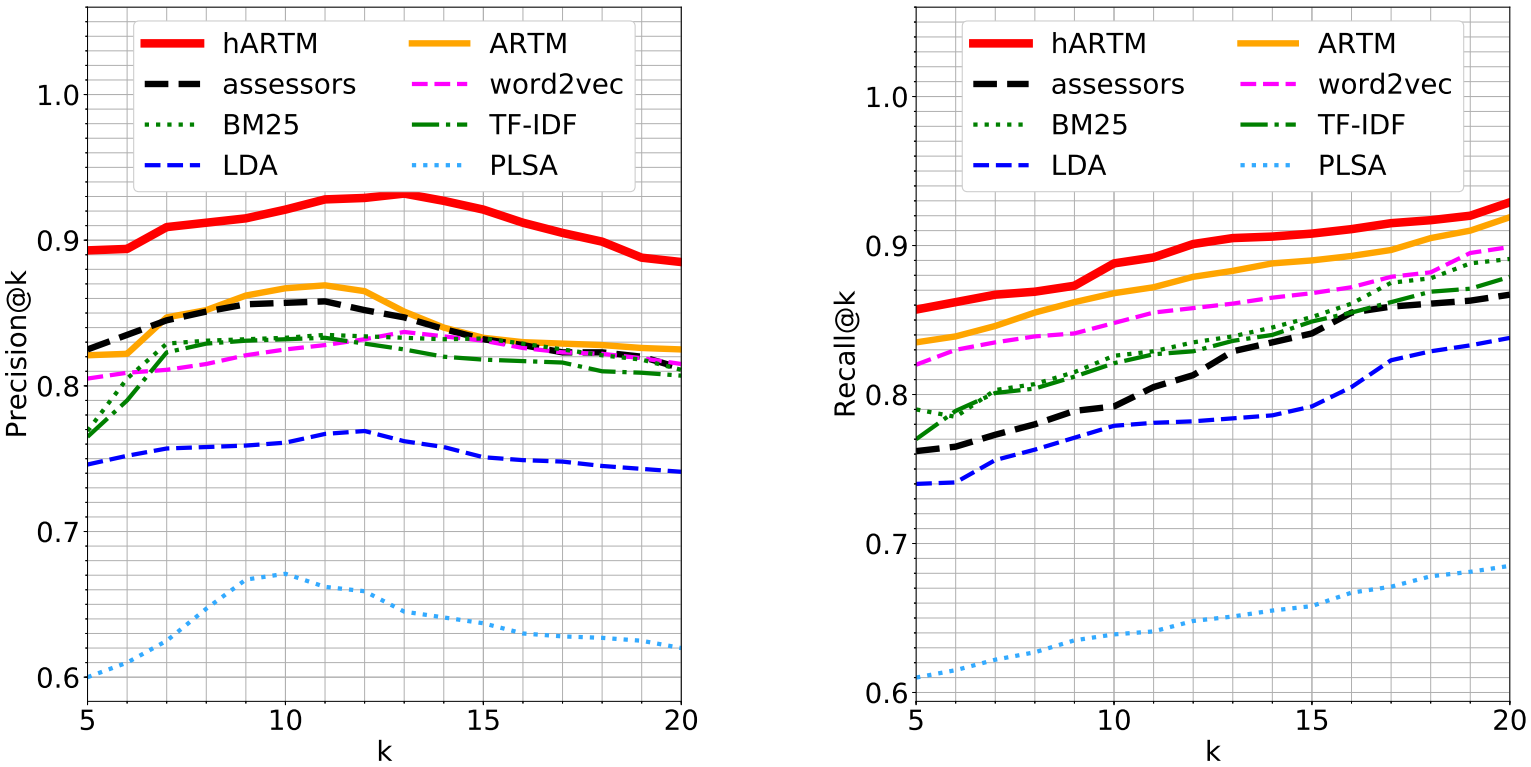




A.Ianina, K.Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

Сравнение качества поиска с асессорами 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

Коллекция Habrahabr.ru

Используем 3 регуляризатора, 5 модальностей, меняем $|T|$

	acессоры	100	150	**200**	250	400
Prec@5	0.821	0.662	0.721	**0.810**	0.761	0.693
Prec@10	0.869	0.761	0.812	**0.879**	0.825	0.673
Prec@15	0.875	0.733	0.795	**0.868**	0.791	0.651
Prec@20	0.863	0.724	0.795	**0.847**	0.792	0.642
Recall@5	0.780	0.732	0.807	**0.840**	0.821	0.721
Recall@10	0.817	0.771	0.843	**0.870**	0.851	0.751
Recall@15	0.850	0.824	**0.895**	0.891	0.871	0.773
Recall@20	0.873	0.857	0.905	**0.925**	0.892	0.771

- Наилучшее качество поиска — при 200 темах

Влияние числа тем на качество поиска

Коллекция TechCrunch.com

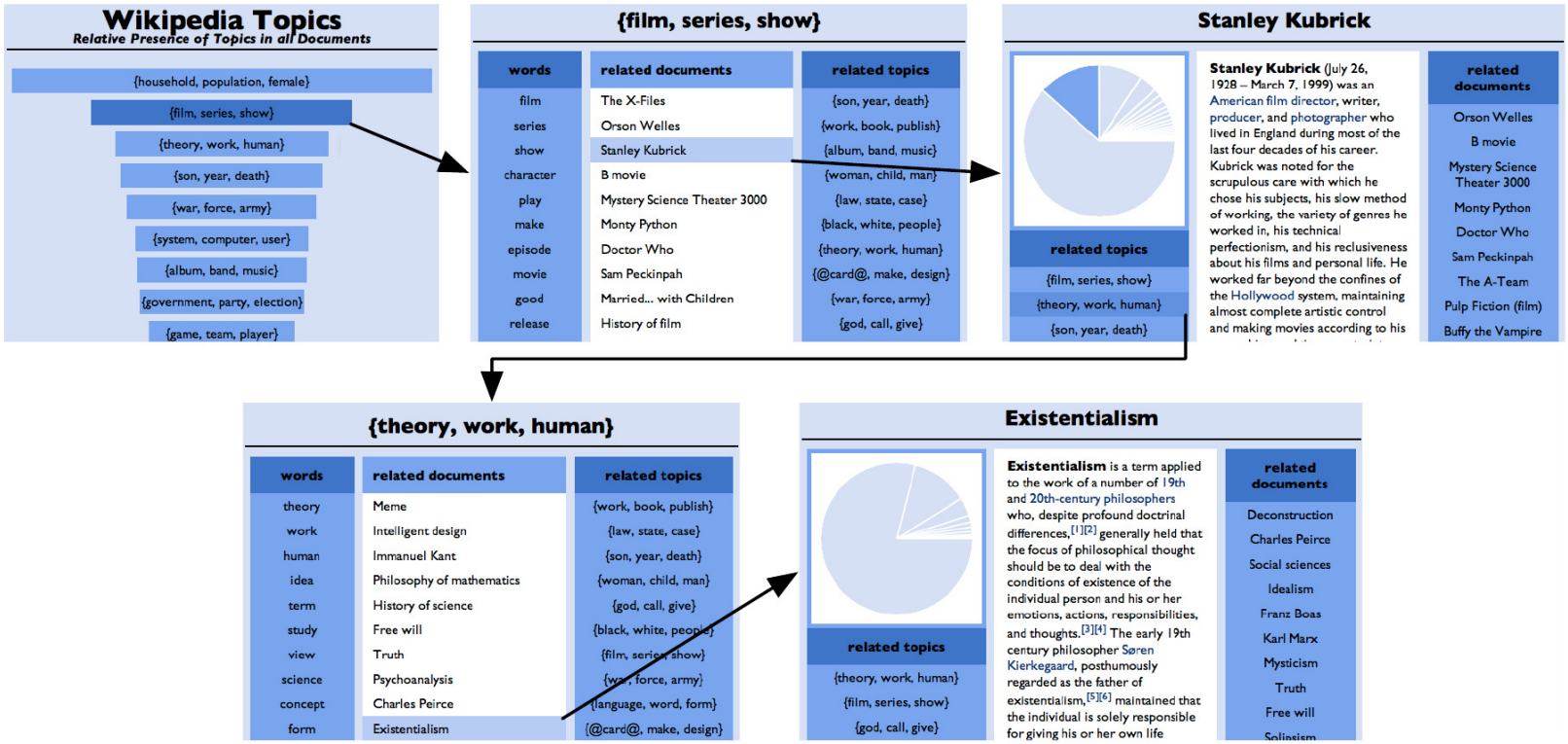
Используем 3 регуляризатора, 4 модальности, меняем $|T|$

	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

Система TMVE — Topic Model Visualization Engine

Тематический навигатор с веб-интерфейсом:

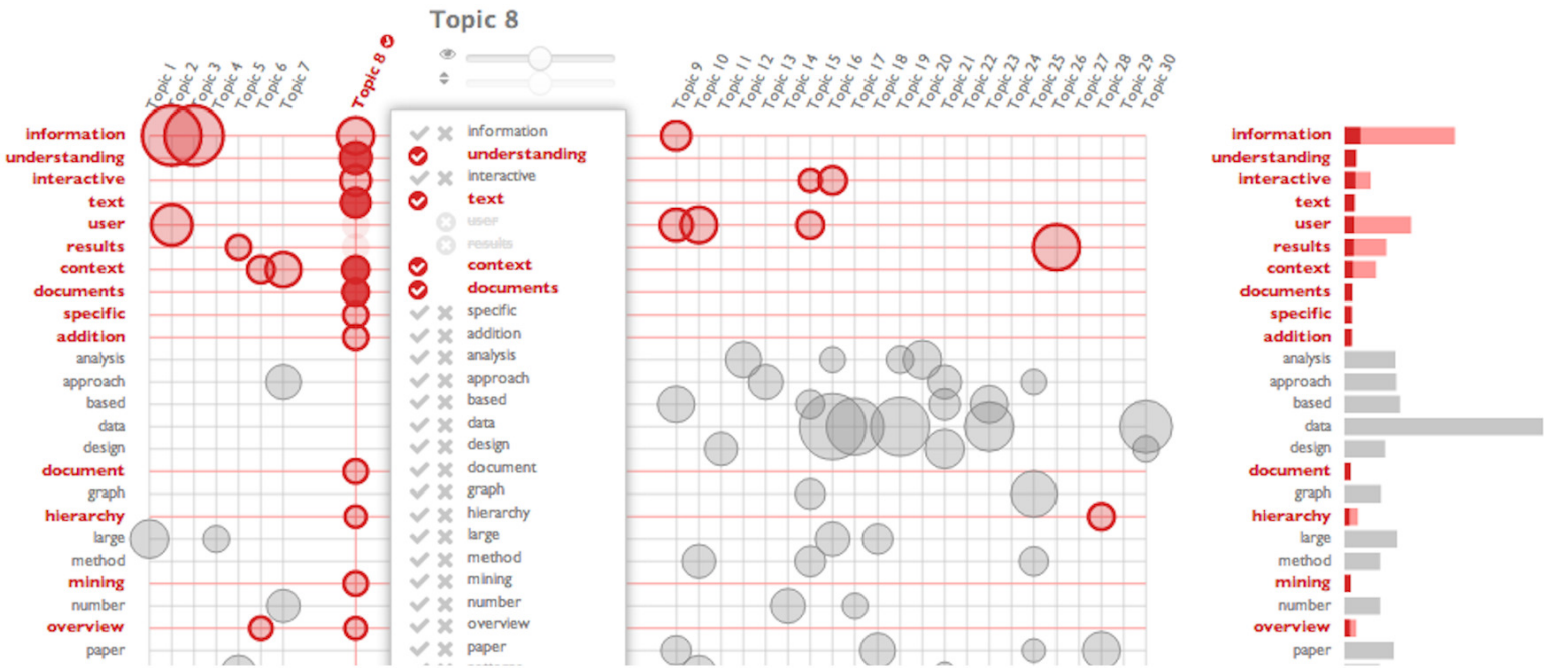


https://github.com/ajbc/tmv

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:

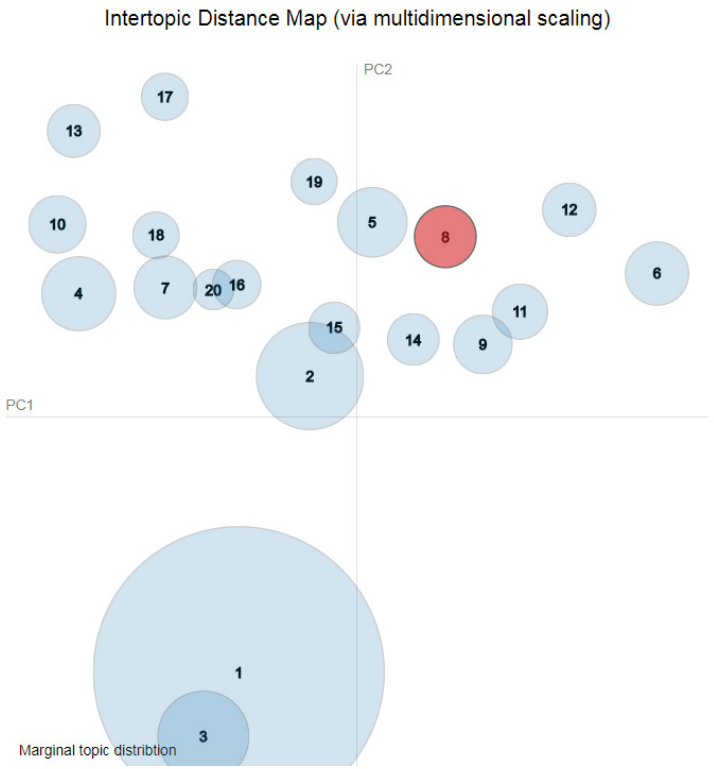


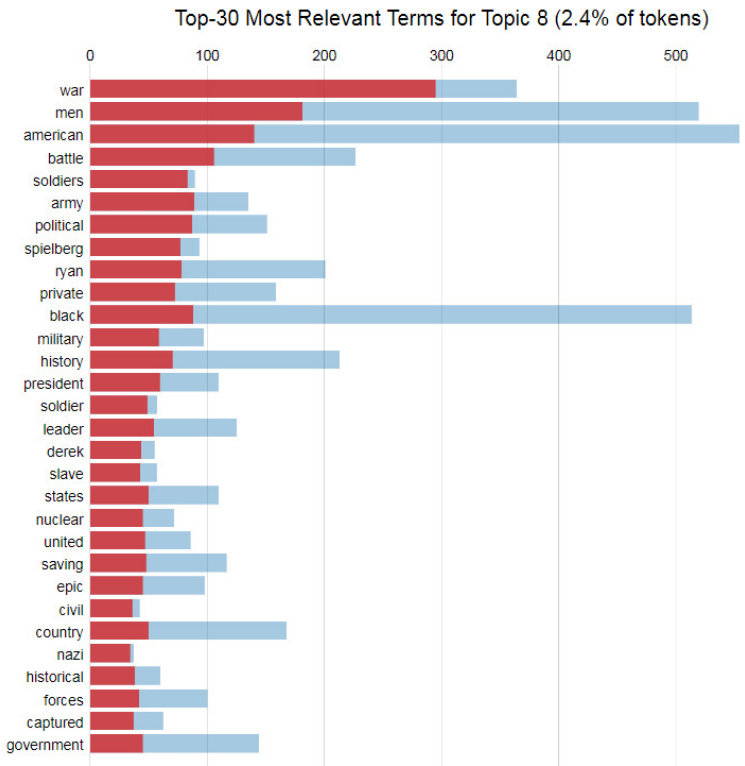
https://github.com/uwdata/termite-visualizations

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 $p(w)$:

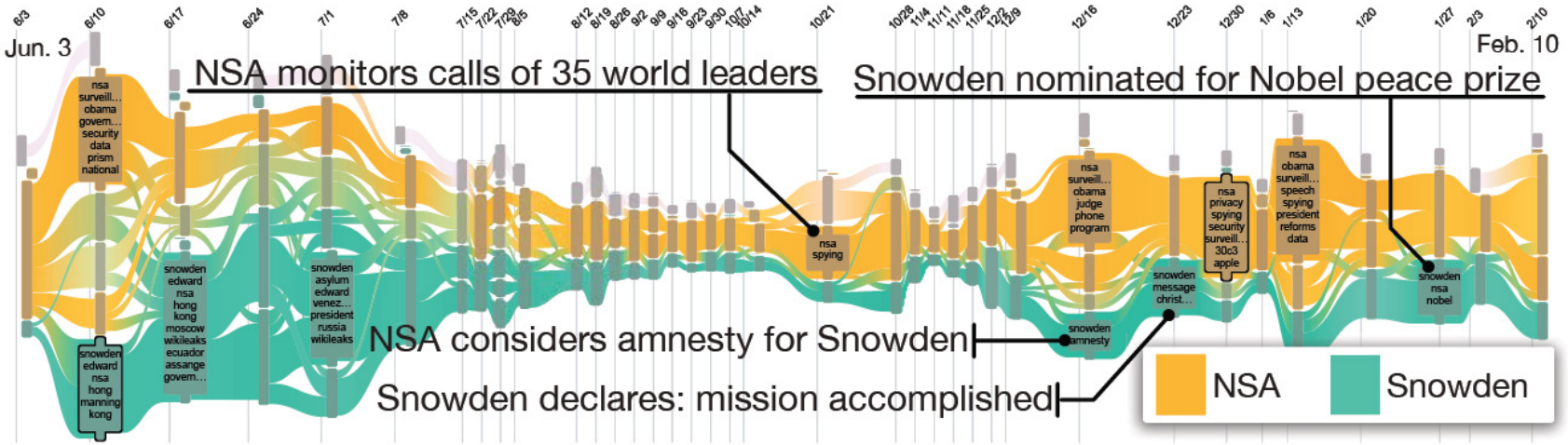




https://github.com/cpsievert/LDAvis

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.

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 весьма разнообразны
- На практике важны внешние критерии качества моделей

Литература

- *К.В.Воронцов.* Обзор вероятностных тематических моделей. 2018. – NEW!
  http://www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf
- *К.В.Воронцов.* Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН. 2014.
- *K.Vorontsov, A.Potapenko.* Additive regularization of topic models. Machine Learning, 2015.
- *O.Frei, M.Apishev.* Parallel non-blocking deterministic algorithm for online topic modeling. AIST 2016.
- *N.Chirkova, K.Vorontsov.* Additive regularization for hierarchical multimodal topic modeling. JMLDA, 2016.
- *A.Ianina, K.Vorontsov.* Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.
- *A.Potapenko, A.Popov, K.Vorontsov.* Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL, 2017.
- *V.Alekseev, V.Bulatov, K.Vorontsov.* Intra-Text Coherence as a Measure of Topic Models Interpretability. Dialogue, 2018.
- *A.Belyy, M.Seleznova, A.Sholokhov, K.Vorontsov.* Quality Evaluation and Improvement for Hierarchical Topic Modeling. Dialogue, 2018.
- *N.Skachkov, K.Vorontsov.* Improving topic models with segmental structure of texts. Dialogue, 2018.