Математические методы анализа текстов

Тематическое моделирование (часть 2)

К. В. Воронцов
vokov@forecsys.ru

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Математические методы анализа текстов (ВМиК МГУ) / 2017»

кафедра ММП • 7 апреля 2017

Содержание

Задача тематического моделирования

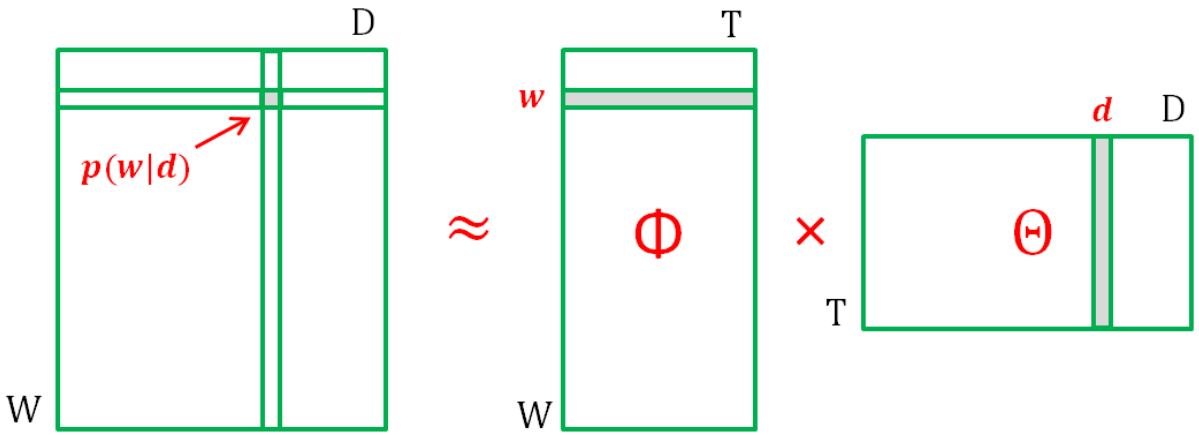
Дано: коллекция текстовых документов

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

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:



ARTM — Аддитивная Регуляризация Тематических Моделей

Максимизация log правдоподобия с регуляризатором R:

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

EM-алгоритм: метод простой итерации для системы уравнений

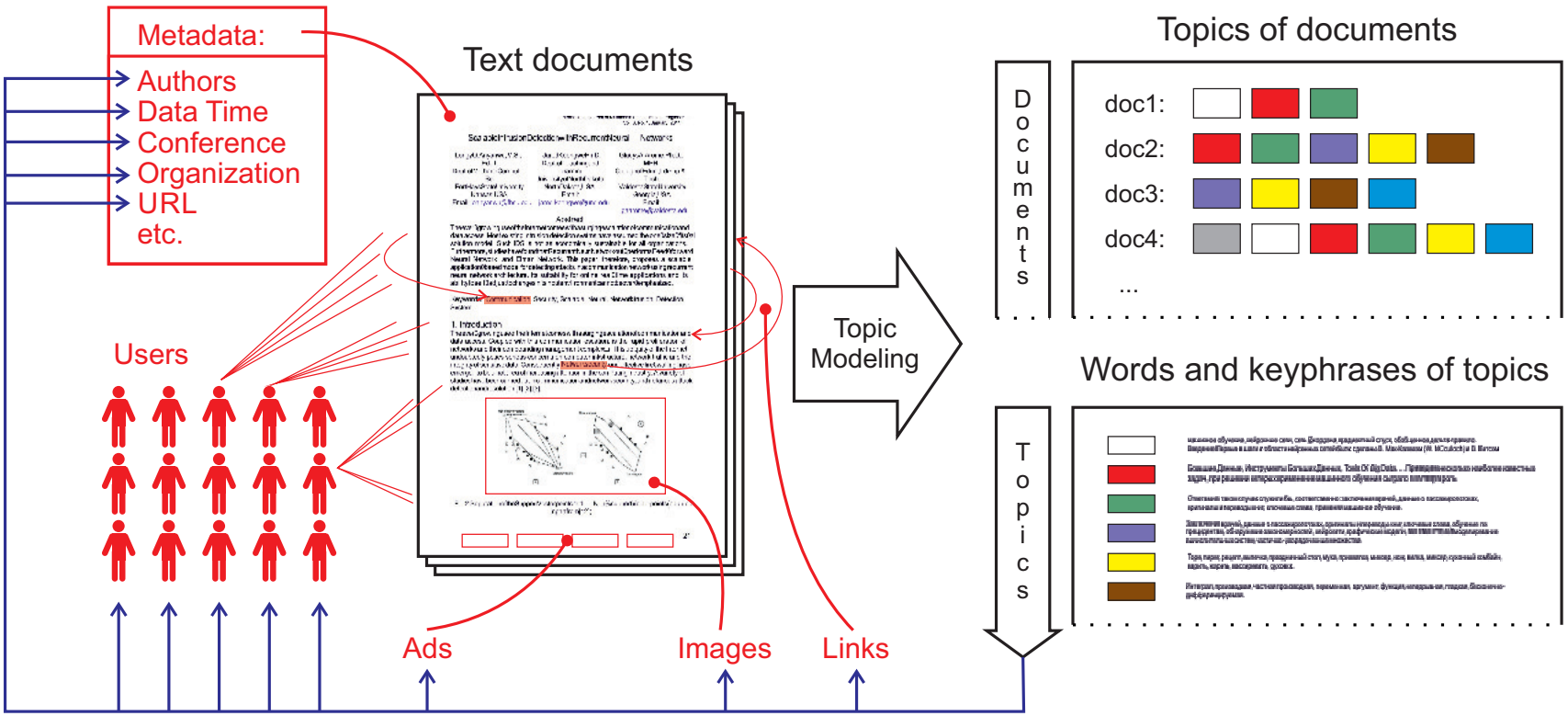
$$\text{E-шаг:} \quad p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

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

Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН. 2014.

Задачи мультимодального тематического моделирования

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



Мультимодальная ARTM

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

$W = W^1 \sqcup \cdots \sqcup W^M$ — объединённый словарь всех модальностей

Максимизация суммы log правдоподобий с регуляризацией:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

$$\theta_{td} = \underset{t \in T}{\operatorname{norm}} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right), \quad n_{td} = \sum_{w \in d} \tau_{m(w)} n_{dw} p_{tdw}$$

Регуляризатор для классификации и категоризации текстов

Y — множество классов;

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

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

Регуляризатор — правдоподобие модальности классов:

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

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):

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

Подставляем, получаем формулы M-шага:

$$\theta_{td} = \underset{t \in T}{\operatorname{norm}} \left(n_{td} + \tau v_t \theta_{td} \left(y_d - \sum_{t \in T} v_t \theta_{td} \right) \right);$$

$$v = (\Theta \Theta^{\mathsf{T}})^{-1} \Theta y.$$

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 изменений 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.

Параллельные и сравнимые корпуса текстов

Parallel — точный перевод (с выравниванием предложений), пример: EuroParl, протоколы европарламента, 21 язык.

Comparable — не перевод, а пересказ на другом языке, пример: Википедия.

Мультиязычные модели (ML-LDA, PLTM, BiLDA)

- каждый язык — отдельная модальность, W^ℓ — словарь языка ℓ из множества языков L.
- $\theta_{td} = p(t|d)$ общее для всех связных документов $d = \bigsqcup\limits_{\ell \in L} d^\ell$

Дополнительные данные — двуязычные словари:

- $\Pi_k(w) \subset W^k$ — все переводы слова $w \in W^\ell$ в языке k
- выравнивание документов по предложениям

I. Vulić, W. De Smet, J. Tang, M.-F. Moens. Probabilistic topic modeling in multilingual settings: an overview of its methodology and applications. 2015

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

Асессор 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 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.

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

Иерархические тематические модели

Стратегии построения тематических иерархий:

- структура иерархии: дерево / многодольный граф
- направление: снизу вверх / сверху вниз / одновременно
- наращивание: поверхинное / послойное

Открытые проблемы:

- "Despite recent activity in the field of HPTMs, determining the hierarchical model that best fits a given data set, in terms of the structure and size of the learned hierarchy, still remains a challenging task and an open issue."
- "The evaluation of hierarchical PTMs is also an open issue."

Zavitsanos E., Paliouras G., Vouros G. A. Non-Parametric Estimation of Topic Hierarchies from Texts with Hierarchical Dirichlet Processes. 2011.

Регуляризатор родительских тем (реализован в BigARTM)

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

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

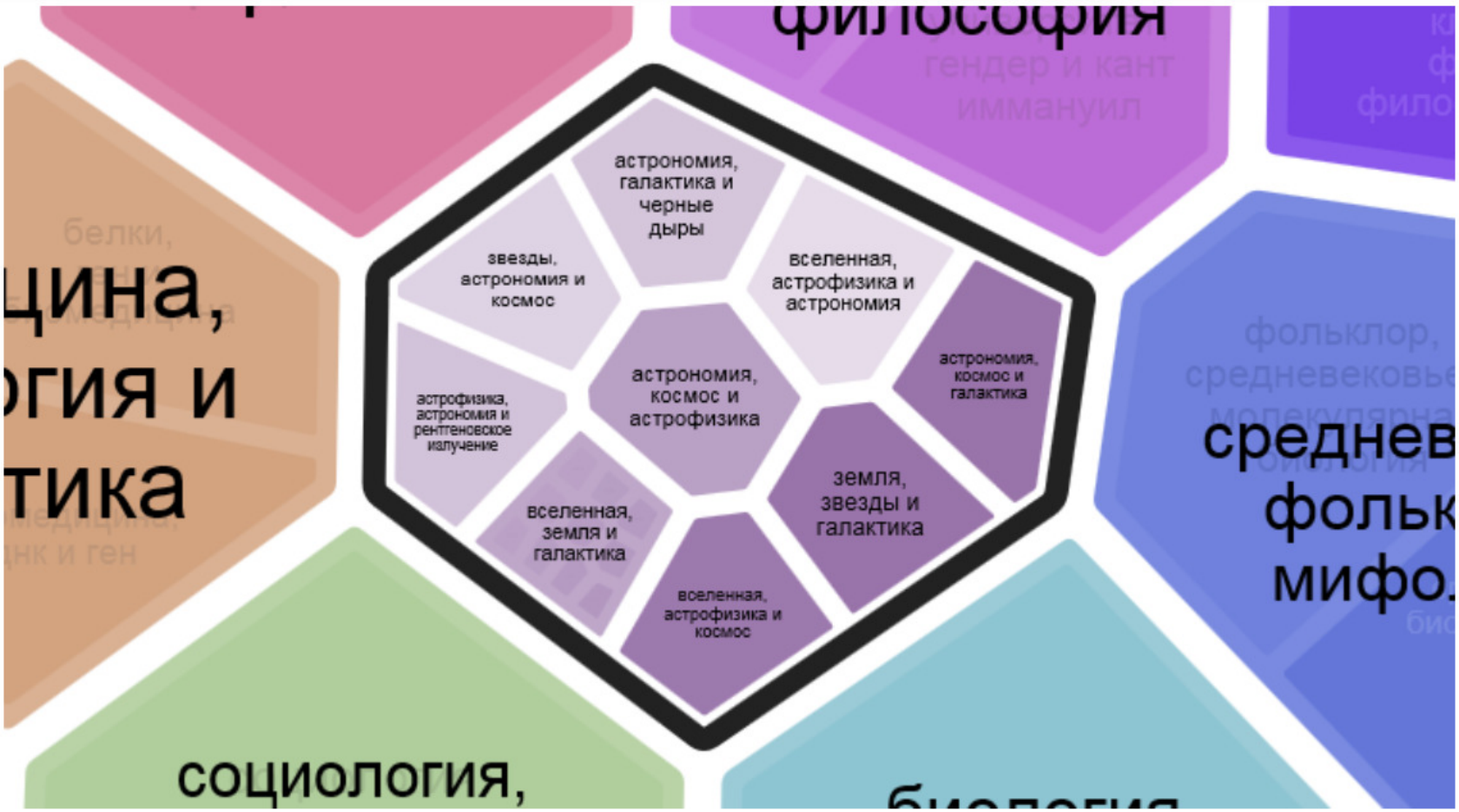
$$R(\Phi,\Psi) = \tau \sum_{t\in T}\sum_{w\in W} n_{wt} \ln \sum_{s\in S} \phi_{ws}\psi_{st} \;\to\; \max.$$

Родительские темы t — «документы» с частотами слов n_{wt}.

Визуализация древовидных иерархий (FoamTree)

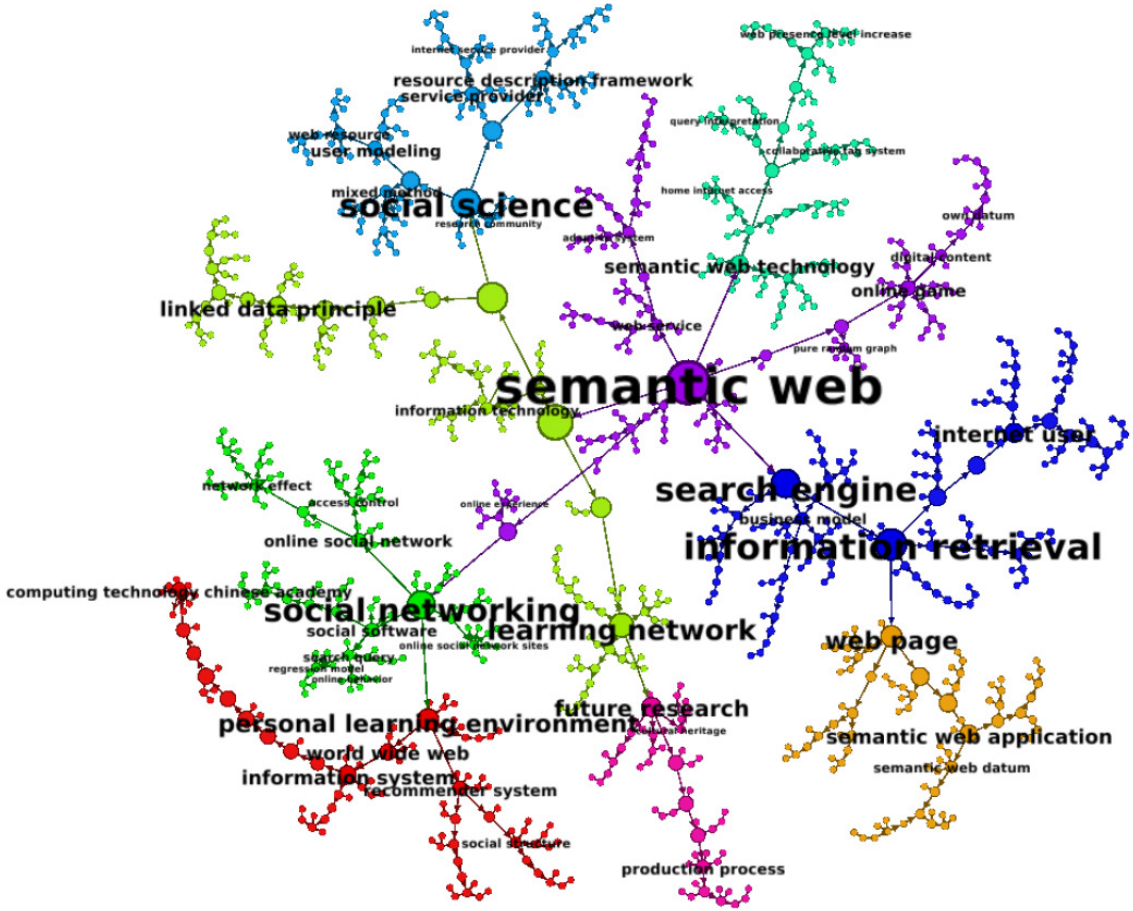
https://carrotsearch.com/foamtree-overview

Визуализация древовидных иерархий (FoamTree)



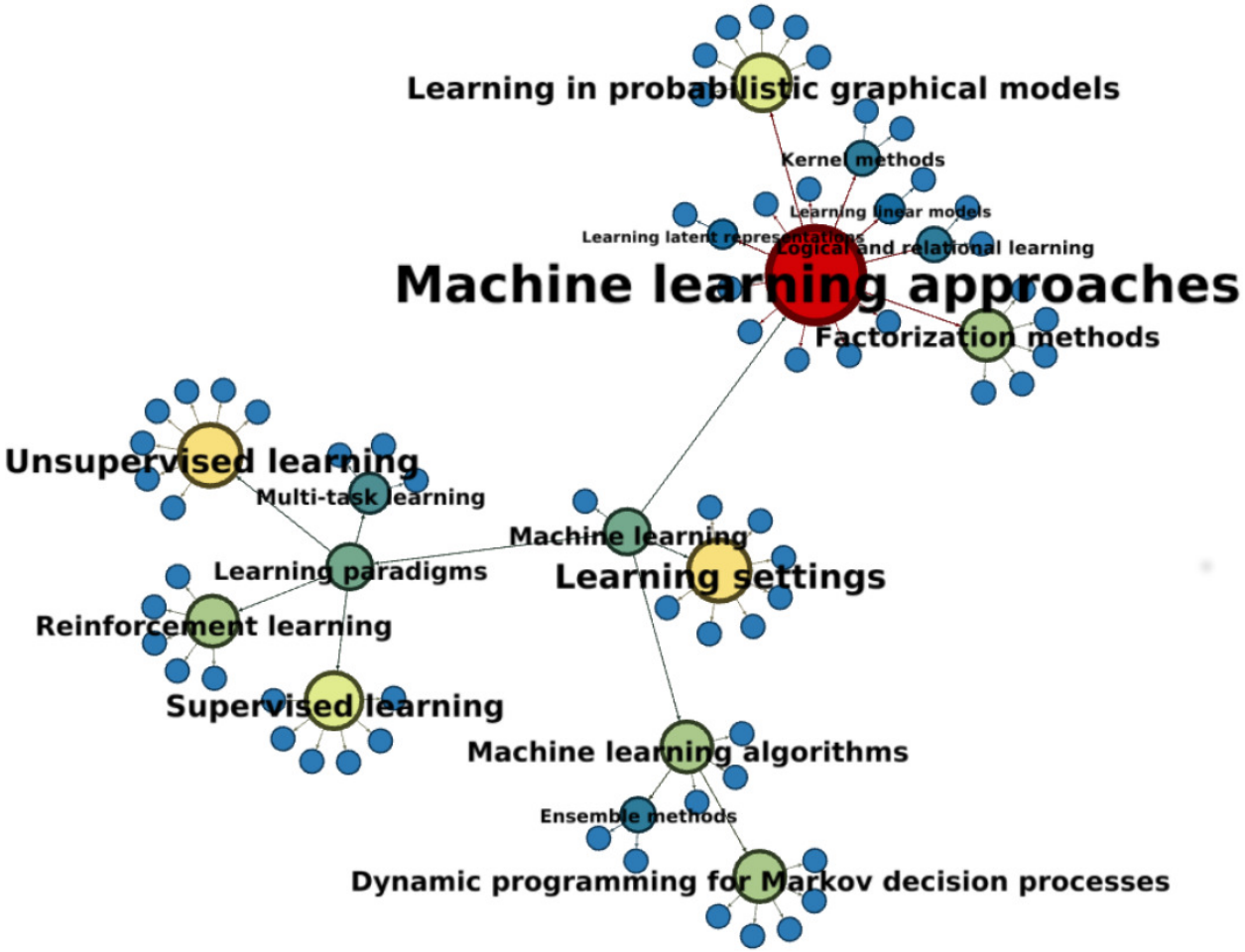
https://carrotsearch.com/foamtree-overview

Визуализация древовидных иерархий



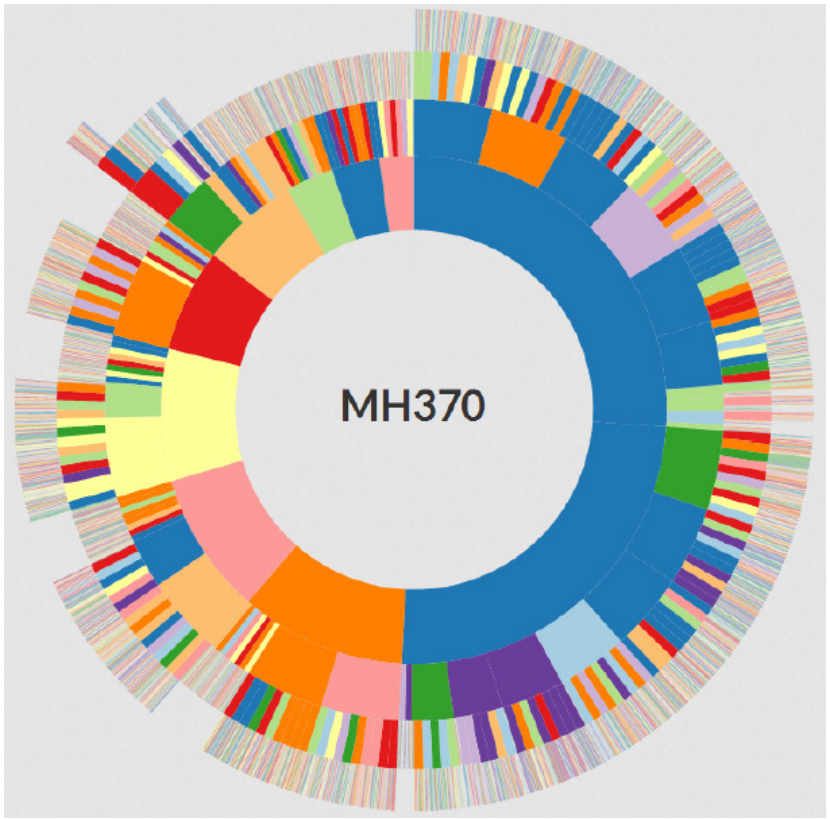
Georgeta Bordea. Domain adaptive extraction of topical hierarchies for Expertise Mining. 2013.

Визуализация древовидных иерархий



Georgeta Bordea. Domain adaptive extraction of topical hierarchies for Expertise Mining. 2013.

Визуализация древовидных иерархий



Smith A., Hawes T., Myers M.. Hiérarchie: interactive visualization for hierarchical topic models. Workshop on Interactive Language Learning, Visualization, and Interfaces, ACL, 2014.

Биграммная тематическая модель

n_{dvw} — частота пары слов «vw» в документе d
$\phi^v_{wt} = p(w|v,t)$ — распределение слов после слова v в теме t

Модель BTM (Bigram Topic Model):

$$\sum_{d\in D}\sum_{vw\in d} n_{dvw} \ln \sum_{t\in T} \phi^v_{wt}\theta_{td} \to \max_{\Phi,\Theta}$$

Это мультимодальная модель:
$M = W$, каждому слову v соответствует отдельная модальность,
$W^v = W$ — все слова, которые могут следовать за v.

Недостатки биграммной модели BTM:

- все пары соседних слов образуют биграммы;
- модель не описывает отдельные слова (униграммы);
- общее число токенов $O(|W|^2)$.

Hanna Wallach. Topic modeling: beyond bag-of-words // ICML 2006

Объединение униграмм и биграмм в одной модели

Модель TNG (Topical n-grams):

$$\sum_{d\in D}\sum_{vw\in d} n_{dvw} \ln \sum_{t\in T} \underbrace{\big(x_{vwt}\phi^{v}_{wt} + (1-x_{vwt})\phi_{wt}\big)}_{p(w|v,t)} \theta_{td} \to \max_{\Phi,\Theta}$$

$x_{vwt} = \mathrm{P}($пара слов «vw» является биграммой в теме $t)$.

Частные случаи:

- $x_{vwt} = x_{vt}$ — матрица параметров в модели TNG.
- $x_{vwt} \equiv 1$ — модель BTM;
- $x_{vwt} =$ [пара слов «vw» из словаря биграмм];

Xuerui Wang, Andrew McCallum, Xing Wei. Topical *N*-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval. 2007.

Мультимодальная мультиграммная ARTM

W^n — словари n-грамм, отфильтрованные по трём критериям:
1) наличие подчинительных синтаксических связей,
2) статистически значимая совстречамость (коллокация),
3) высокая тематичность $\mathrm{KL}\left(\frac{1}{|T|} \,\middle\|\, p(t|vw)\right)$.

Связь с моделью TNG.
При $x_{vwt} = \lambda[vw \in W^2]$ сумма log-правдоподобий модальностей является оценкой снизу для log-правдоподобия модели TNG:

$$\sum_{d\in D}\sum_{vw\in d} n_{dvw} \ln \sum_{t\in T}\left(x_{vwt}\phi^v_{wt} + (1-x_{vwt})\phi_{wt}\right)\theta_{td} =$$

$$\sum_{d\in D}\sum_{vw\in d} n_{dvw} \ln\left(\lambda \sum_{t\in T}\phi^v_{wt}\theta_{td} + (1-\lambda)\sum_{t\in T}\phi_{wt}\theta_{td}\right) \geqslant$$

$$\lambda \sum_{d,vw} n_{dvw} \ln \sum_{t\in T}\phi^v_{wt}\theta_{td} + (1-\lambda)\sum_{d,w} n_{dw} \ln \sum_{t\in T}\phi_{wt}\theta_{td} \to \max_{\Phi,\Theta}$$

Биграммы радикально улучшают интерпретируемость тем

Коллекция 1000 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

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

Проблема коротких текстов

Короткие тексты (short text): Twitter и другие микроблоги, социальные медиа, заголовки статей и новостных сообщений.

Основные проблемы коротких сообщений:

- огромный объём ($\sim 10^9$ твитов в день)
- концентрация распределения $p(t|d)$ в одной теме
- неустойчивость определения темы $p(t|d)$
- выделение редких тем на фоне основных тем микроблогов (личная переписка, life style, репосты новостей)
- опечатки и намеренное искажение слов языка
- раннее обнаружение новых тем

Битермы: модель совстречаемости слов в коротких текстах

Битерм — пара слов, встречающихся рядом:
в одном коротком сообщении / предложении / окне $\pm h$ слов.

Тематическая модель битермов (Biterm topic model):

$$p(u,w) = \sum_{t\in T} p(w|t)p(u|t)p(t) = \sum_{t\in T} \phi_{wt}\phi_{ut}\pi_t,$$

где $\phi_{wt} = p(w|t)$, $\pi_t = p(t)$ — параметры модели.

Критерий максимума логарифма правдоподобия:

$$\sum_{u,w} n_{uw} \ln \sum_t \phi_{wt}\phi_{ut}\pi_t \;\to\; \max_{\Phi,\pi},$$

$$\phi_{wt} \geqslant 0; \quad \textstyle\sum_w \phi_{wt} = 1; \qquad \pi_t \geqslant 0; \quad \textstyle\sum_t \pi_t = 1$$

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Xueqi Cheng. A Biterm Topic Model for Short Texts // WWW 2013.

Необходимые условия точки максимума правдоподобия

Максимизация log правдоподобия с регуляризатором R:

$$\sum_{u,w} n_{uw} \ln \sum_{t} \phi_{wt}\phi_{ut}\pi_t + R(\Phi,\pi) \;\to\; \max_{\Phi,\pi},$$

n_{uw} — частота битерма (u,w) в документах коллекции.

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad p_{tuw} \equiv p(t|u,w) = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\phi_{ut}\pi_t\right)$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w\in W}{\text{norm}}\left(n_{wt} + \phi_{wt}\dfrac{\partial R}{\partial \phi_{wt}} \right), \quad n_{wt} = \sum\limits_{u\in W} n_{uw} p_{tuw} \\ \pi_t = \underset{t\in T}{\text{norm}}\left(n_t + \pi_t \dfrac{\partial R}{\partial \pi_t} \right), \quad n_t = \sum\limits_{u,w\in W} n_{uw} p_{tuw} \end{cases}$$

Битермы как регуляризатор для обычной ΦΘ-модели

1. Регуляризатор битермов для матрицы Φ:

$$R(\Phi) = \tau \sum_{u,w \in W} n_{uw} \ln \sum_{t \in T} n_t \phi_{ut} \phi_{wt} \to \max$$

Подставляем в формулу M-шага, получаем сглаживание:

$$\phi_{wt} = \underset{w}{\operatorname{norm}} \Big(n_{wt} + \tau \sum_{u \in W} n_{uw} p_{tuw} \Big);$$

$$p_{tuw} \equiv p(t|u,w) = \underset{t \in T}{\operatorname{norm}} \big(n_t \phi_{wt} \phi_{ut} \big).$$

2. Регуляризатор разреживания для матрицы Θ:

$$R(\Theta) = -\tau' \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td} \to \max.$$

Модель сети слов WNTM для коротких текстов

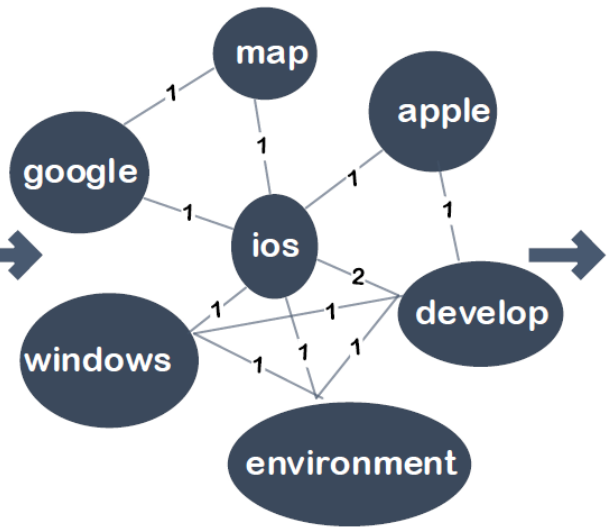
Идея: моделировать не документы, а связи между словами.

d_w — псевдо-документ, объединение всех контекстов слова w.

n_{wu} — число вхождений слова u в псевдо-документ d_w.

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:

$$p(u|d_w) = \sum_{t \in T} p(u|t)p(t|d_w) = \sum_{t \in T} \phi_{ut}\theta_{tw},$$

где d_w — псевдо-документ слова w.

Максимизация логарифма правдоподобия:

$$\sum_{u,w \in W} n_{wu} \log \sum_{t \in T} \phi_{ut}\theta_{tw} \to \max_{\Phi,\Theta},$$

где n_{wu} — совстречаемость слов w, u.

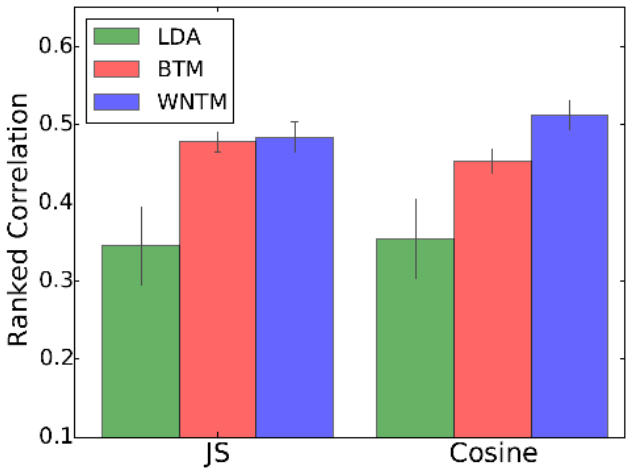
Отличие от модели битермов: там $\Theta = \mathrm{diag}(\pi_1, \ldots, \pi_t)\Phi^{\mathrm{T}}$.

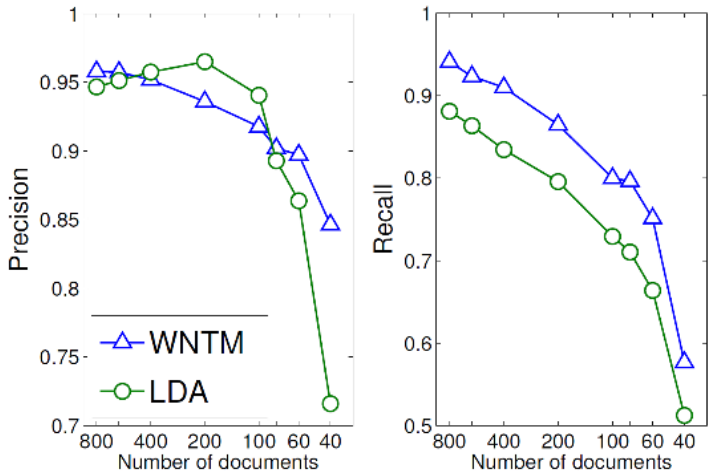
Yuan Zuo, Jichang Zhao, Ke Xu. Word Network Topic Model: a simple but general solution for short and imbalanced texts. 2014.

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

Результаты оценивания модели WNTM

- Когерентность на коротких текстах лучше, чем у LDA и Biterm TM; на длинных текстах преимуществ нет.
- *Слева:* оценивание семантической близости слов по $p(t|w)$, корреляция с 10-балльными экспертными оценками.
- *Справа:* полнота и точность распознавания новой темы в зависимости от числа документов.





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

Стандартная методика оценивания моделей языка

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

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

$d = d' \sqcup d''$ — 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''.

Интерпретации перплексии:
1) $\mathcal{P}(D') \to |d''|$ при $n \to \infty$, если слова равновероятны;
2) насколько хорошо мы предсказываем слова в документах (чем меньше перплексия, тем лучше).

Оценки разреженности темы

- Разреженность:
  - доля нулевых элементов в Φ
  - доля нулевых элементов в Θ
- Характеристики различности тем:
  - размер ядра темы: $|W_t|$, ядро $W_t = \{w : p(t|w) > 0.25\}$
  - чистота темы: $\sum_{w \in W_t} p(w|t)$
  - контрастность темы: $\frac{1}{|W_t|} \sum_{w \in W_t} p(t|w)$
- Вырожденность тематической модели:
  - доля фоновых слов: $\frac{1}{n} \sum_{d \in D} \sum_{w \in d} \sum_{t \in B} p(t|d,w)$
  - доля нетематичных документов: $\frac{1}{|D|} \sum_{d \in D} \Big[\sum_{t \in B} p(t|d) > 0.95 \Big]$
  - доля нетематичных терминов: $\frac{1}{|W|} \sum_{w \in W} \Big[\sum_{t \in B} p(t|w) > 0.95 \Big]$

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:

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

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.

Регуляризатор для максимизации когерентности тем

Гипотеза: тема лучше интерпретируется, если она содержит *когерентные* (часто встречающиеся рядом) слова $u, w \in W$.

Пусть C_{uw} — оценка когерентности, например $\hat{p}(w|u) = \frac{N_{uw}}{N_u}$.
Согласуем ϕ_{wt} с оценками $\hat{p}(w|t)$ по когерентным словам,

$$\hat{p}(w|t) = \sum_u p(w|u)p(u|t) = \frac{1}{n_t}\sum_u C_{uw} n_{ut};$$

$$R(\Phi) = \tau \sum_{t\in T} n_t \sum_{w\in W} \hat{p}(w|t) \ln \phi_{wt} \to \max.$$

Подставляем в формулу M-шага, получаем сглаживание:

$$\phi_{wt} = \underset{w}{\text{norm}}\left(n_{wt} + \tau \sum_{u\in W} C_{uw} n_{ut}\right).$$

Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A. Optimizing semantic coherence in topic models // EMNLP-2011. — Pp. 262–272.

Альтернативный регуляризатор когерентности

Квадратичный регуляризатор Quad-Reg:

$$R(\Phi) = \tau \sum_{t \in T} \ln \sum_{u,v \in W} C_{uv} \phi_{ut} \phi_{vt} \to \max,$$

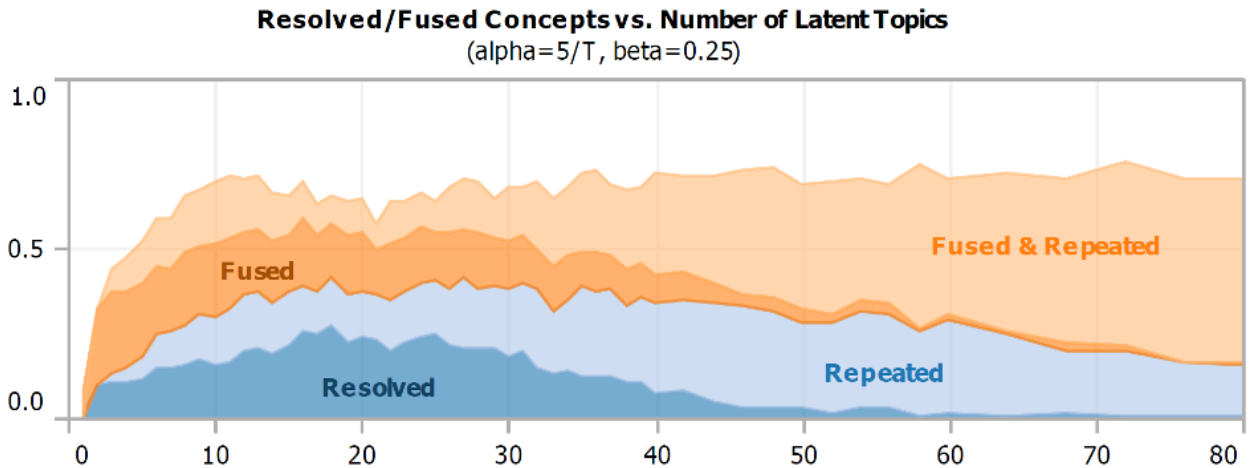
$C_{uv} = N_{uv} \left[\mathrm{PMI}(u,v) > 0 \right]$,
N_{uv} — число документов, в которых u, v хотя бы раз встречаются на расстоянии не более 10 слов,
N_u — число документов, в которых u встречается хотя бы раз,
$\mathrm{PMI}(u,v) = \ln \frac{|D| N_{uv}}{N_u N_v}$ — поточечная взаимная информация.

В литературе пока не выработан окончательный вариант регуляризатора когерентности.

Newman D., Bonilla E. V., Buntine W. L. Improving topic coherence with regularized topic models. 2011.

Внешние критерии

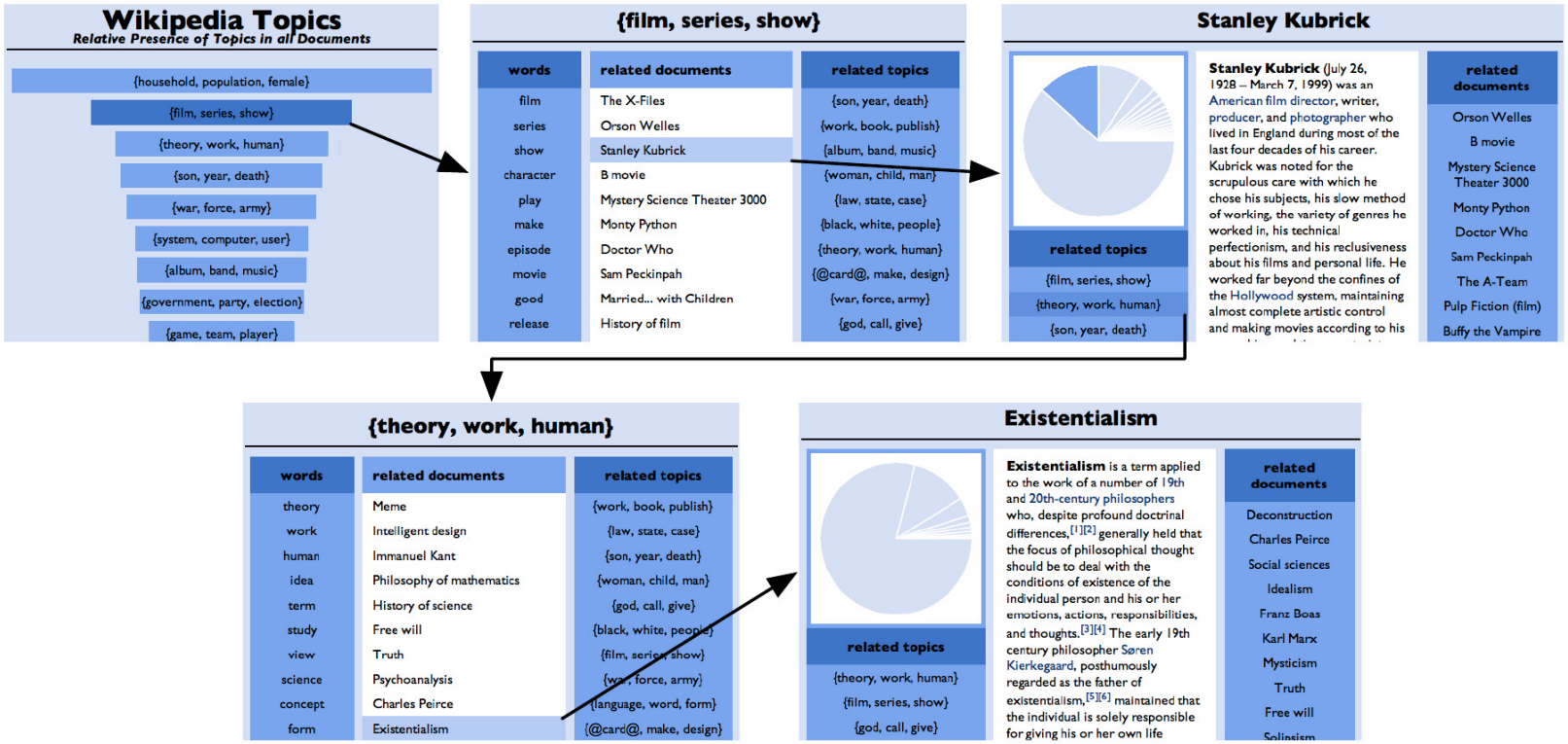
- Полнота и точность тематического поиска
- Качество ранжирования при тематическом поиске
- Качество категоризации документов
- Несоответствие тем и *концептов* — число пропущенных, смешаных, дублированных, расщеплённых концептов



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

Система TMVE — Topic Model Visualization Engine

Тематический навигатор с веб-интерфейсом:

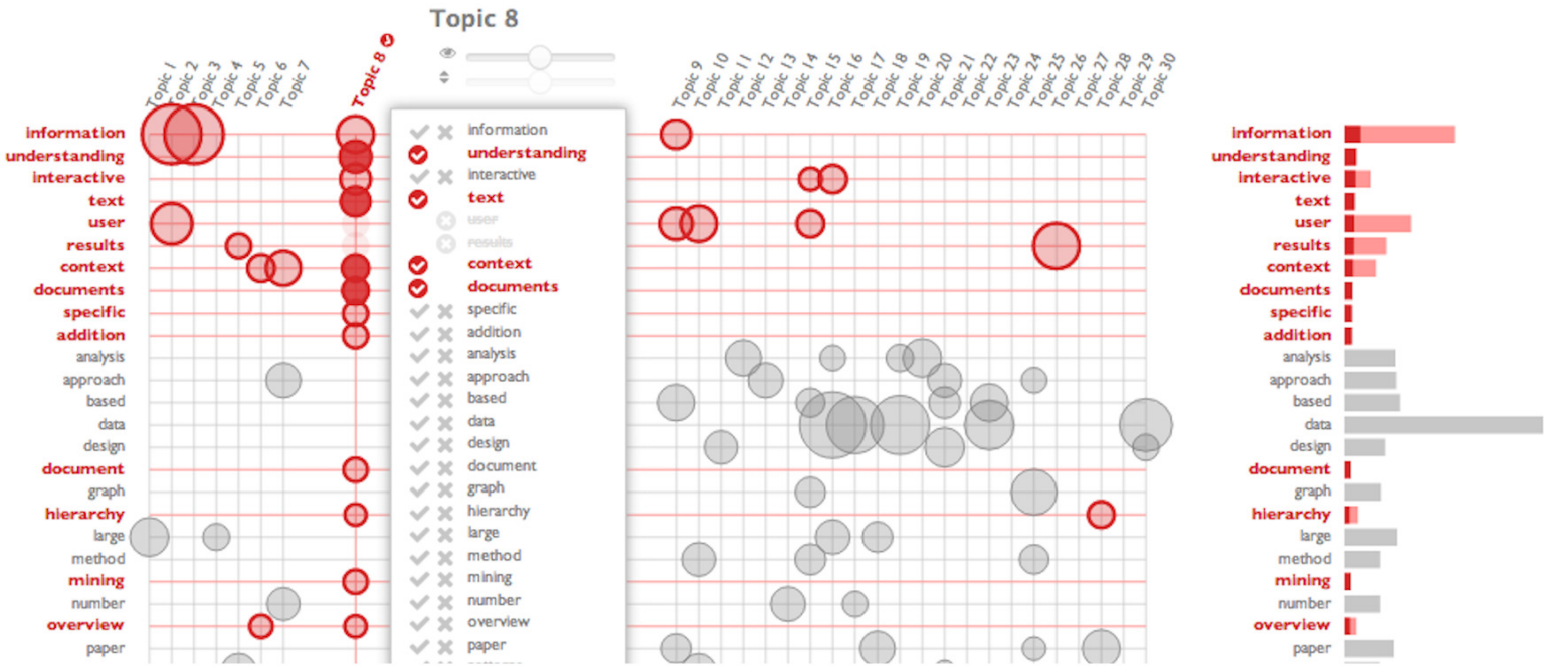


https://github.com/ajbc/tmv

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

Система Termite

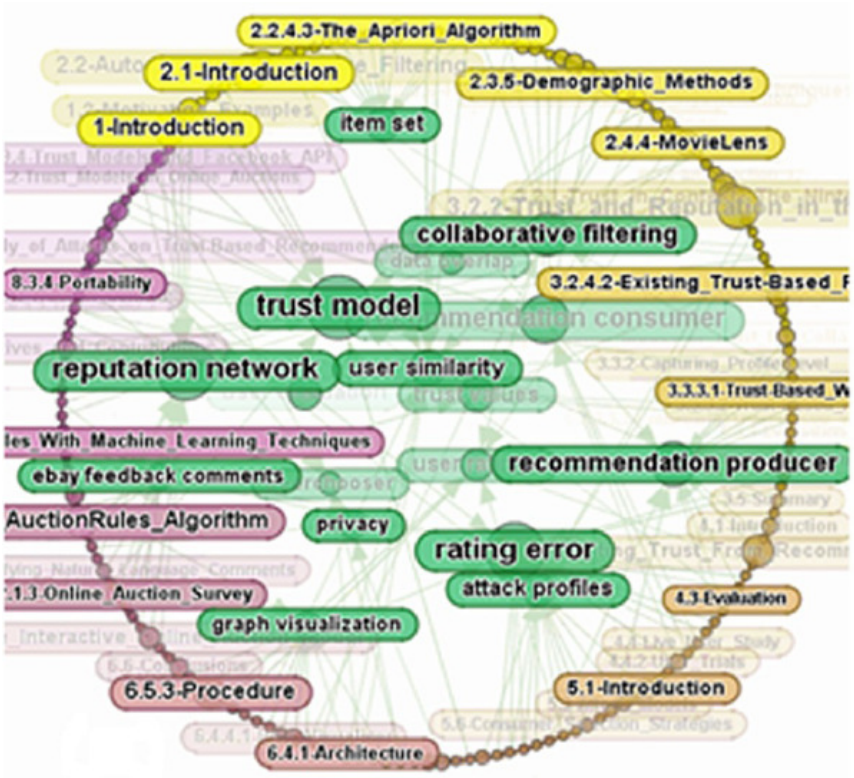
Интерактивная визуализация матрицы Φ и сравнение тем:

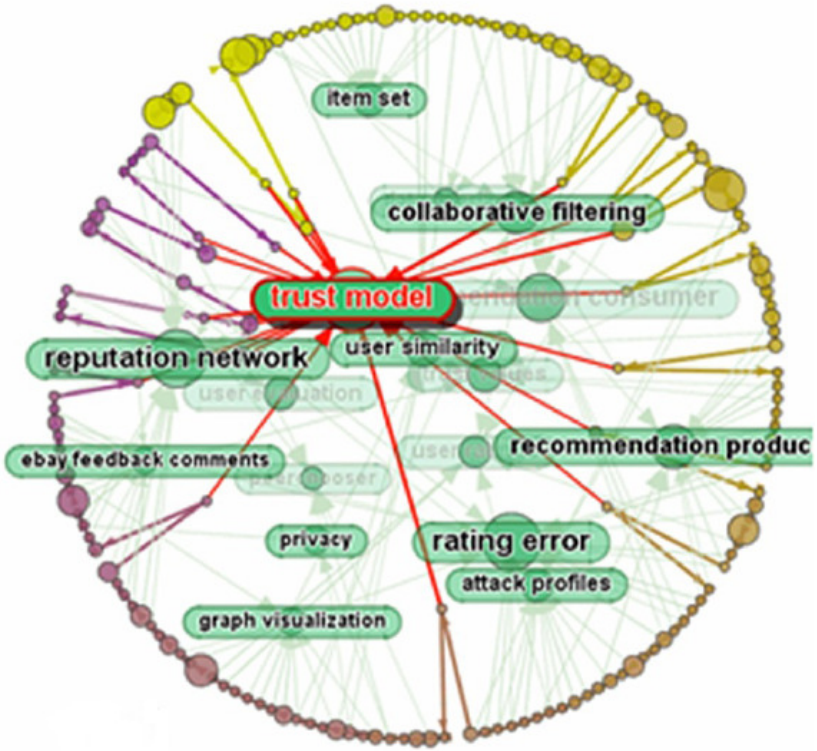


https://github.com/uwdata/termite-visualizations

Chuang J., Manning C., Heer J. Termite: Visualization Techniques for Assessing Textual Topic Models // International Working Conference on Advanced Visual Interfaces, 2012. ACM. pp. 74–77.

Тематическая сегментация документа запроса





Gretarsson, O'Donovan, Bostandjiev, Hollerer, Asuncion, Newman, Smyth. TopicNets: visual analysis of large text corpora with topic modeling. ACM Trans. on Intelligent Systems and Technology. 2012.

http://textvis.lnu.se

Интерактивный обзор 375 средств визуализации текстов

Айсина Р. М. Обзор средств визуализации тематических моделей коллекций текстовых документов // JMLDA, 2015.

Резюме

- Модальности
  - классы, категории
  - языки
  - биграммы
- Псевдо-документы
  - родительские темы в иерархиях
  - контексты слов (модель WNTM)
- Регуляризаторы
  - разреживание, сглаживание, частичное обучение
  - декоррелирование тем
  - битермы (для коротких текстов)
  - когерентность
- Метрики качества
  - внутренние (перплексия, когерентность, различность тем)
  - внешние (качество классификации, поиска, рекомендаций)