

Вероятностные тематические модели

Лекция 2. Обзор задач и моделей

К. В. Воронцов
vokov@forecsys.ru

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

Задача тематического моделирования текстовой коллекции

Дано: W — словарь терминов

D — коллекция текстовых документов $d = \{w_1 \dots w_{n_d}\}$

n_{dw} = сколько раз термин w встречается в документе d

Найти параметры модели $p(w|d) = \sum_t \phi_{wt}\theta_{td}$:

$\phi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t

$\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Задача максимизации логарифма правдоподобия:

$$\mathcal{L}(\Phi, \Theta) = \sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} \rightarrow \max_{\Phi, \Theta}$$

при ограничениях нормировки и неотрицательности

$$\phi_{wt} \geq 0; \quad \sum_w \phi_{wt} = 1; \quad \theta_{td} \geq 0; \quad \sum_t \theta_{td} = 1$$

1 Обзор тематических моделей

- Разновидности тематических моделей
- Средства визуализации
- Разведочный информационный поиск

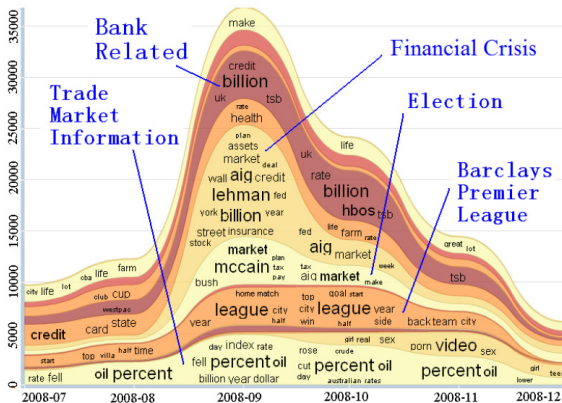
2 Проекты по тематизации текстовых коллекций

- Поиск этно-релевантных тем в социальных сетях
- Информационный анализ электрокардиосигналов
- Пресс-релизы, статьи, авторефераты и другие проекты

3 Задачи и открытые проблемы

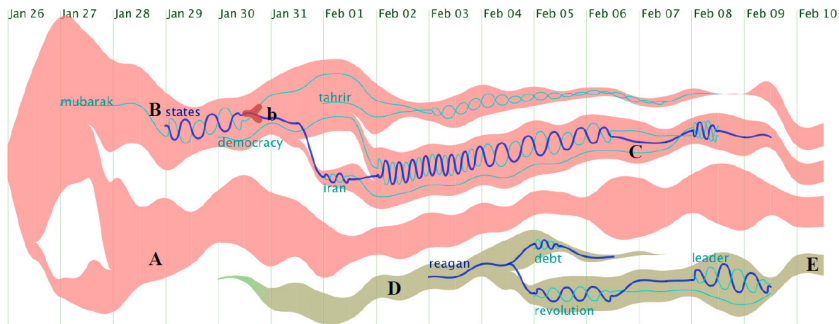
- Проект BigARTM
- Направления дальнейших исследований

Динамические модели, учитывающие время



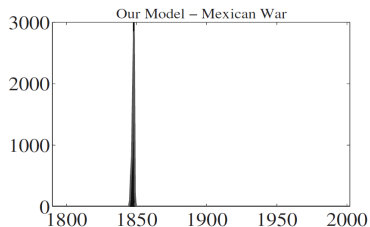
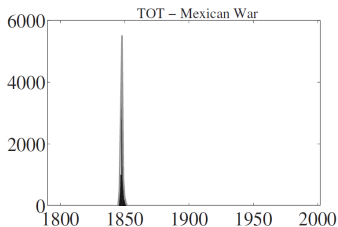
Jianwen Zhang, Yangqiu Song, Changshui Zhang, Shixia Liu Evolutionary Hierarchical Dirichlet Processes for Multiple Correlated Time-varying Corpora // KDD'10, July 25–28, 2010.

Динамические модели эволюции тем



Weiwei Cui, Shixia Liu, Li Tan, Conglei Shi, Yangqiu Song, Zekai J. Gao, Xin Tong, Huamin Qu TextFlow: Towards Better Understanding of Evolving Topics in Text // IEEE Transactions On Visualization And Computer Graphics, Vol. 17, No. 12, December 2011.

Совмещение динамической и n -граммной модели

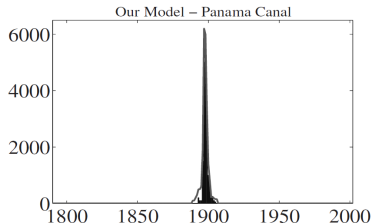
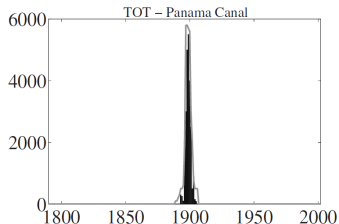


1. mexico	8. territory
2. texas	9. army
3. war	10. peace
4. mexican	11. act
5. united	12. policy
6. country	13. foreign
7. government	14. citizens

1. east bank	8. military
2. american coins	9. general herrera
3. mexican flag	10. foreign coin
4. separate independent	11. military usurper
5. american commonwealth	12. mexican treasury
6. mexican population	13. invaded texas
7. texan troops	14. veteran troops

Shoaib Jameel, Wai Lam. An N-Gram Topic Model for Time-Stamped Documents // 35'th ECIR 2013, Moscow, March 24–27. — pp. 292–304.

Совмещение динамической и n -граммной модели



1. government	8. spanish
2. cuba	9. island
3. islands	10. act
4. international	11. commission
5. powers	12. officers
6. gold	13. spain
7. action	14. rico

1. panama canal	8. united states senate
2. isthmian canal	9. french canal company
3. isthmus panama	10. caribbean sea
4. republic panama	11. panama canal bonds
5. united states government	12. panama
6. united states	13. american control
7. state panama	14. canal

Shoaib Jameel, Wai Lam. An N-Gram Topic Model for Time-Stamped Documents // 35'th ECIR 2013, Moscow, March 24–27. — pp. 292–304.

Биграммы радикально улучшают интерпретируемость тем

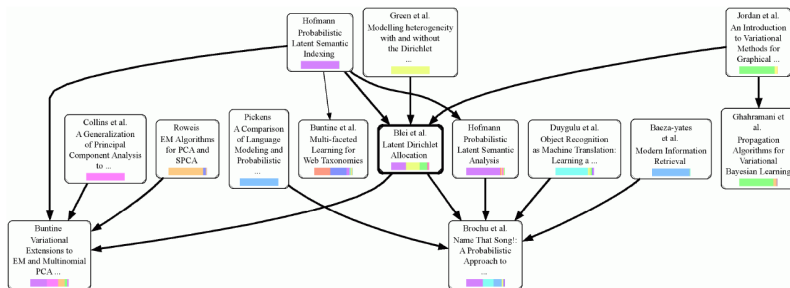
Коллекция 1000 статей конференций ММРО, ИОИ на русском

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

Модели, учитывающие цитирования или гиперссылки

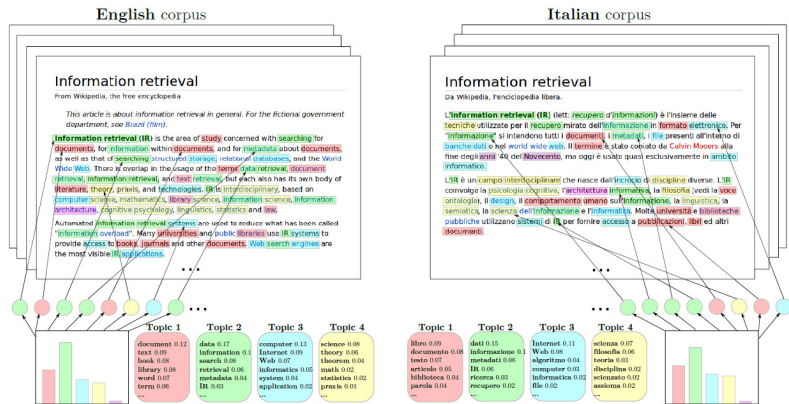
Учёт ссылок уточняет тематическую модель

Тематическая модель выявляет самые влиятельные ссылки



Laura Dietz, Steffen Bickel, Tobias Scheffer. Unsupervised prediction of citation influences // ICML-2007, Pp. 233–240.

Многоязычные модели



I. Vulić, W. De Smet, J. Tang, M.-F. Moens. Probabilistic topic modeling in multilingual settings: a short overview of its methodology with applications // NIPS, 7–8 December 2012. — Pp. 1–11.

Мультиязычная модель Википедии

216 175 русско-английских пар статей Вики.

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

Topic 68				Topic 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.

Мультиязычная модель Википедии

216 175 русско-английских пар статей Вики.

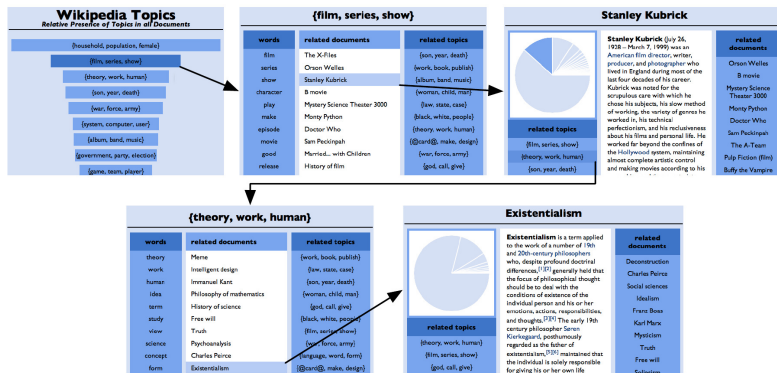
Первые 10 слов и их вероятностями $p(w|t)$ в %:

Topic 88				Topic 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

Независимый ассессор оценил 396 тем из $|T| = 400$ как хорошо интерпретируемые.

Система TMVE — Topic Model Visualization Engine

Тематический навигатор с веб-интерфейсом:

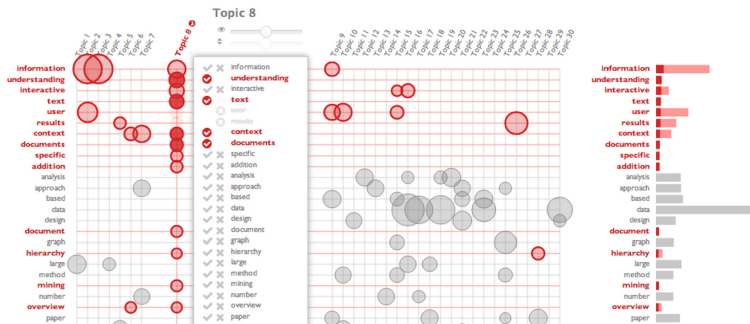


<https://github.com/ajbc/tmv>

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

Система Termite

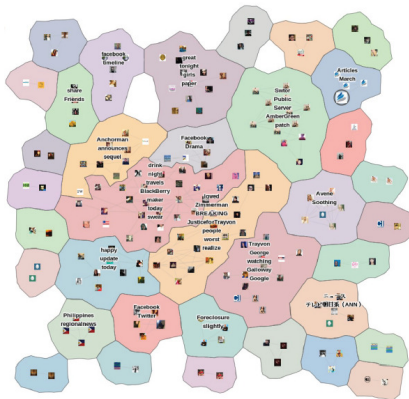
Интерактивная визуализация матрицы Φ и сравнение тем:



<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.

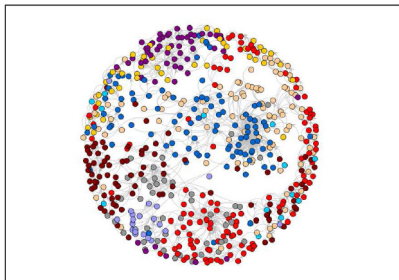
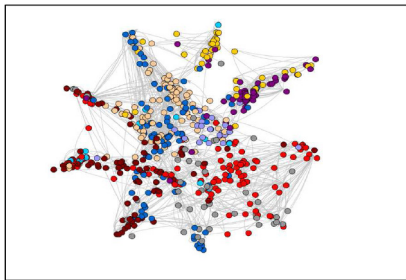
Дорожная карта: кластеризация релевантных документов



«A map metaphor visualization (left) seems more appealing than a plain graph layout (right), and clusters seem easier to identify.»

E.R.Gansner, Y.Hu, S.North. Visualizing Streaming Text Data with Dynamic Maps. 2012.

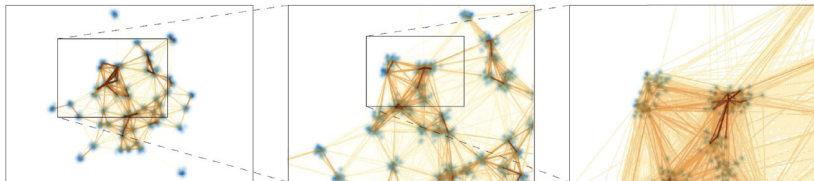
Дорожная карта: кластеризация релевантных документов



- Точки — это документы (или их фрагменты)
- Кластеры — это группы тематически схожих документов
- Форму облака точек можно настраивать

Tuan M. V. Le, Hady W. Lauw Probabilistic Latent Document Network Embedding. IEEE International Conference ICDM. 2014.

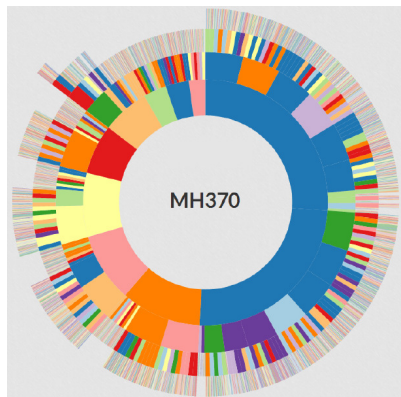
Дорожная карта: кластеризация релевантных документов



- Кластеры
 кластеров
 кластеров
 кластеров...

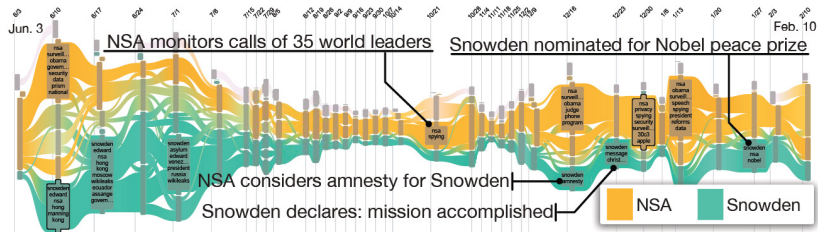
M.Zinsmaier, U.Brandes, O.Deussen, H.Strobelt. Interactive level-of-detail rendering of large graphs. IEEE Trans. Vis. Comput. Graph. 2012.

Тематическая иерархия: структура предметной области



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

Динамика тем: эволюция предметной области

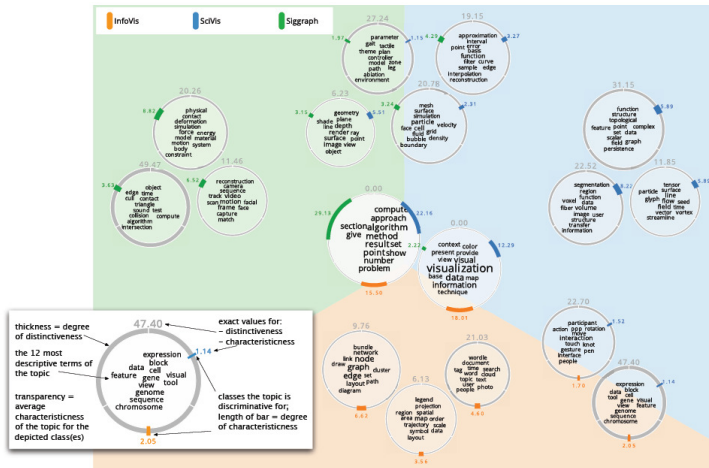


Эволюция иерархии тем. Коллекция Prism (2013/06/03 – 2014/02/09)

- эксперт выбирает сечение тематической иерархии,
- затем отмечает события в интерактивном режиме,
- и генерирует отчёт.

Weiwei Cui, Shixia Liu, Zhuofeng Wu, Hao Wei. How hierarchical topics evolve in large text corpora. IEEE Trans. Vis. Comput. Graph. 2014.

Тематический анализ источников



Oelke D., Strobel H., Rohrdantz C., Gurevych I., Deussen O. Comparative exploration of document collections: a visual analytics approach. EuroVis. 2014.

<http://textvis.lnu.se>

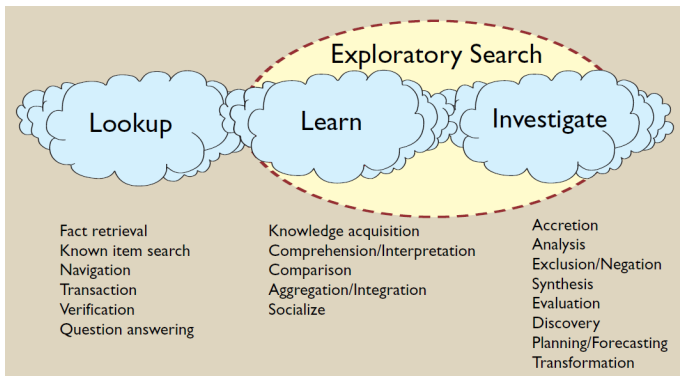
Интерактивный обзор 272 средств визуализации текстов



Айсина Р. М. Обзор средств визуализации тематических моделей коллекций текстовых документов // Машинное обучение и анализ данных (<http://jmla.org>). 2015. Т. 1, № 11. С. 1584-1618.

Разведочный поиск как инструмент самообразования

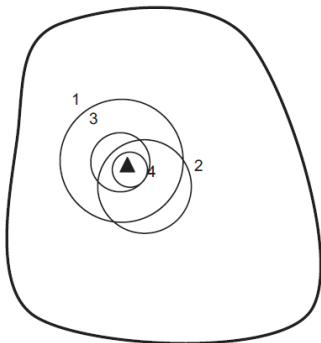
- пользователь может не знать ключевых терминов
- пользователя может интересовать множество ответов



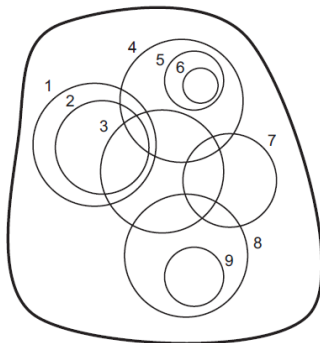
Gary Marchionini. Exploratory Search: from finding to understanding. Communications of the ACM. 2006, 49(4), p. 41–46.

От поиска “query-browse-refine” к разведочному поиску

Iterative Search



Exploratory Search



- ▲ Search target ◻ Information space
○ Result sets (larger = more results, intersection = overlap, # = iteration)

R.W.White, R.A.Roth. Exploratory Search: beyond the Query-Response paradigm. San Rafael, CA: Morgan and Claypool, 2009.

Возможный сценарий разведочного поиска

Поисковый запрос:

- документ любой длины или даже коллекция документов

Цели поиска:

- к каким темам относится мой запрос?
- что ещё известно по этим темам?
- какова тематическая структура этой предметной области?
- что ещё есть понятного, обзорного, важного, свежего?

Сценарий поиска:

- 1 имея любой текст под рукой, в любом приложении,
- 2 хотим получить картину содержащихся в нём тем-подтем,
- 3 и «дорожную карту» предметной области в целом

Разведочный поиск: прототип интерфейса

Радужная полоса напоминает, что знания всегда под рукой

BigARTM

BigARTM — открыта библиотека для тематического моделирования больших коллекций текстовых документов. Параллельная распределенная реализация методов вероятностного тематического моделирования на основе аддитивной регуляризации.

ARTM (англ.) — Additive Regularization for Topic Modeling.

ARTM (рус.) — Аддитивная Регуляризация Тематических Моделей.

Теоретическое введение

Основная статья: Тематическое моделирование

Вероятностное тематическое моделирование — это современный инструмент статистического анализа текстов, предназначенный для выявления тематик коллекций документов. Тематическая модель включает в себя следующие распределения на множестве терминов, каждый документ — дисперсионное распределение на множестве тем, тематические модели используются для информационного поиска, классификации, категоризации, аннотирования, семантизации текстов.

Тематическая модель — это представление наблюдаемого условного распределения $p(v|d)$ терминов (или их омонимов) v в документе d коллекции D :

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

где T — множество тем;

$\phi_{vt} = p(v|t)$ — неизвестное распределение терминов в теме t ;

$\theta_{td} = p(t|d)$ — неизвестное распределение тем в документе d .

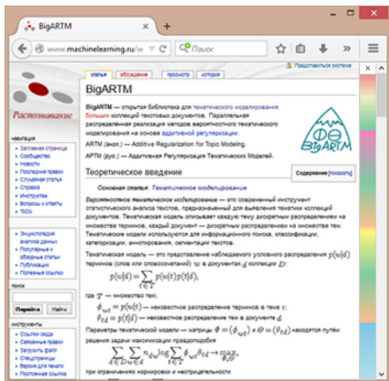
Параметры тематической модели — матрицы $\Phi = (\phi_{vt})$ и $\Theta = (\theta_{td})$ являются путями решения задачи минимизации правдоподобия

$$\sum_{d \in D} \sum_{v \in V} n_{dv} \log \sum_{t \in T} \phi_{vt} \theta_{td} \rightarrow \min_{\Phi, \Theta},$$

при ограничениях нормировки и неотрицательности

Разведочный поиск: прототип интерфейса

Клик по **радужной полосе** — тематический поисковый запрос



Разведочный поиск: прототип интерфейса

Темы-подтемы выбранного фрагмента текста

BigARTM — открытая библиотека для тематического моделирования больших коллекций текстовых документов. Параллельная распределенная реализация метода вероятностного тематического моделирования на основе аддитивной регуляризации.

ARTM (англ.) — Additive Regularization for Topic Modeling.

ARTM (рус.) — Аддитивная Регуляризация Тематических Моделей.

Теоретическое введение

Основная статья: Тематическое моделирование

Вероятностное тематическое моделирование — это современный инструмент статистического анализа текстов, предназначенный для выявления тематик коллекций документов. Тематическая модель включает в себя:

- тему: дискретное распределение на множестве термине, каждый документ — дискретное распределение на множестве тем;
- тематические модели используются для информационного поиска, классификации, категоризации, аннотирования, семантики текстов.

Тематическая модель — это представление наблюдаемого условного распределения $p(w|d)$ термине (или ее логосимilarity) θ в документе d коллекции D :

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

где T — множество тем;

$$\phi_{wt} = p(w|t) \text{ — неизвестное распределение термине в теме } t;$$

$$\theta_{dt} = p(t|d) \text{ — неизвестное распределение тем в документе } d.$$

Параметры тематической модели — матрицы $\Phi = (\phi_{wt})$ и $\Theta = (\theta_{dt})$ находят путь решения задачи максимизации правдоподобия

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

при ограничениях: неотрицательности

Topics in «BigARTM» [English] [Russian]

- Natural language processing
 - Statistical text analysis
 - Probabilistic topic modeling
- Probability theory
 - Likelihood maximization
- Mathematical programming
 - Nonconvex optimization
 - Constrained nonconvex optimization
- Machine Learning
 - Topic Modeling
 - Probabilistic Topic Modeling
- Matrix Factorization
 - Nonnegative Matrix Factorization
 - Probabilistic Topic Modeling
- Parallel computing
- Big Data

Разведочный поиск: прототип интерфейса

Документы и иные объекты, ранжированные по релевантности

BigARTM — открытая библиотека для тематического моделирования больших коллекций текстовых документов. Параллельная распределенная реализация метода вероятностного тематического моделирования на основе аддитивной регуляризации.

ARTM (англ.) — Additive Regularization for Topic Modeling.

ARTM (рус.) — Аддитивная Регуляризация Тематических Моделей.

Теоретическое введение

Основная статья: Тематическое моделирование

Вероятностное тематическое моделирование — это современный инструмент статистического анализа текстов, предназначенный для выявления тематик коллекций документов. Тематическая модель включает в себя: теги, дисперсионные на множестве термов, каждый документ — дисперсионное разложение на множество тем, Тематические модели используются для информационного поиска, классификации, категоризации, аннотирования, семантики текстов.

Тематическая модель — это представление наблюдаемого условного разложения $p(v|d)$ термов (или их эмбедингов) v по документам d коллекции D :

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

где T — множество тем;

$$\phi_{vt} = p(v|t) \text{ — неизвестное распределение термов в теме } t;$$

$$\theta_{dt} = p(t|d) \text{ — неизвестное распределение тем в документе } d.$$

Параметры тематической модели — матрицы $\Phi = (\phi_{vt})$ и $\Theta = (\theta_{dt})$ — вводятся нулевыми решения задачи максимизации правдоподобия

$$\sum_{d \in D} \sum_{v \in V} n_{dv} \log \sum_{t \in T} \phi_{vt} \theta_{dt} \rightarrow \max_{\Phi, \Theta}.$$

при ограничениях неотрицательности и нормированности

BigARTM - MachineLearning.ru
www.machinelearning.ru/wiki/index.php?title=BigARTM -
30 апр. 2015 г. - BigARTM — открытая библиотека для тематического моделирования больших коллекций текстовых документов. Параллельная ...
Теоретическое введение - Функциональные возможности BigARTM

Welcome to BigARTM's documentation! — BigARTM 1.0 ...
bigartm.readthedocs.org/ - Перейдите эту страницу
BigARTM FAQ - Can I use BigARTM from other programming languages (not Python)? How to retrieve Theta matrix from BigARTM BigARTM Developer's Guide.

Tutorial — BigARTM 1.0 documentation
bigartm.readthedocs.org/en/latest/tutorial.html - Перейдите эту страницу
Please refer to Basic BigARTM tutorial for Windows users or Basic BigARTM tutorial for Linux and Mac OS-X users depending on your operating system.

BigARTM FAQ — BigARTM 1.0 documentation
bigartm.readthedocs.org/en/latest/faq.html - Перейдите эту страницу
Can I use BigARTM from other programming languages (not Python)? - The following figure shows how to call BigARTM methods directly on arm.dll (Windows) ...

bigartm/bigartm - GitHub
https://github.com/bigartm/bigartm - Перейдите эту страницу
Contribute to bigartm development by creating an account on GitHub.

bigartm/tutorial.txt at master · bigartm/bigartm - GitHub
https://github.com/bigartm/bigartm/blob/master/tutorial.txt - Перейдите эту страницу
Contribute to bigartm development by creating an account on GitHub.

Releases · bigartm/bigartm - GitHub
https://github.com/bigartm/bigartm/releases - Перейдите эту страницу
Contribute to bigartm development by creating an account on GitHub.

Разведочный поиск: прототип интерфейса

Дорожная карта: кластеризация релевантных документов

BigARTM — открытая библиотека для тематического моделирования. Сопровождается коллекцией тестовых документов. Параллельная распределённая реализация метода вероятностного тематического моделирования на основе аддитивной регуляризации.

ARTM (англ.) — Additive Regularization for Topic Modeling.

ARTM (рус.) — Аддитивная Регуляризация Тематических Моделей.

Теоретическое введение

Основная идея: Тематическое моделирование

Вероятностное тематическое моделирование — это современный инструмент статистического анализа текстов, предназначенный для выявления тематик коллекций документов. Тематическая модель включает в себя: точечное распределение на множестве тем, тематические модели, каждый документ — линейное распределение на множестве тем, Тематические модели используются для информационного поиска, классификации, категоризации, аннотирования, семантики текстов.

Тематическая модель — это представление наблюдаемого условного распределения $p(v|d)$ термов (слов или эмблематичей) v в документе d :

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

где T — множество тем;

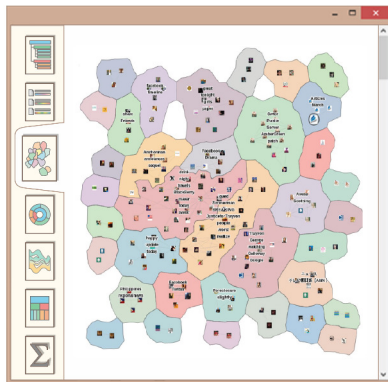
$\phi_{vt} = p(v|t)$ — неизвестное распределение термов в теме t ;

$\theta_{dt} = p(t|d)$ — неизвестное распределение тем в документе d .

Параметры тематической модели — матрицы $\Phi = (\phi_{vt})$ и $\Theta = (\theta_{dt})$ находят путь решения задачи максимизации правдоподобия:

$$\sum_{d \in D} \sum_{v \in V} \sum_{t \in T} \phi_{vt} \log \sum_{t \in T} \theta_{dt} \phi_{vt}$$

при ограничениях: нормировка и неотрицательность.



Разведочный поиск: прототип интерфейса

Динамика тем: эволюция предметной области

BigARTM

BigARTM — открытая библиотека для тематического моделирования больших коллекций текстовых документов. Параллельная распределенная реализация метода вероятностного тематического моделирования на основе аддитивной регуляризации.

ARTM (англ.) — Additive Regularization for Topic Modeling.

ARTM (рус.) — Аддитивная Регуляризация Тематических Моделей.

Теоретическое введение

Основная идея: Тематическое моделирование

Вероятностное тематическое моделирование — это современный инструмент статистического анализа текстов, предназначенный для выделения тематик коллекций документов. Тематическая модель включает в себя: точечные распределения на множестве терминов, каждый документ — дискретное распределение на множестве тем. Тематические модели используются для информационного поиска, классификации, категоризации, аннотирования, семантики текстов.

Тематическая модель — это представление наблюдаемого условного распределения $p(w|d)$ термов w (слов или эмблематик) d в документе d :

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

где T — множество тем;

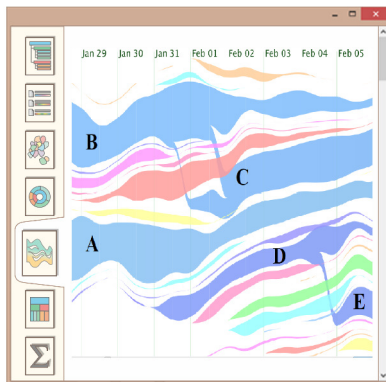
$\phi_{wt} = p(w|t)$ — известное распределение термов в теме t ;

$\theta_{dt} = p(t|d)$ — известное распределение тем в документе d .

Параметры тематической модели — матрицы $\Phi = (\phi_{wt})$ и $\Theta = (\theta_{dt})$ находят путем решения задачи максимизации правдоподобия:

$$\sum_{d \in D} \sum_{w \in V} \sum_{t \in T} \phi_{wt} \log \sum_{t \in T} \theta_{dt} \phi_{wt} \rightarrow \max_{\Phi, \Theta}.$$

при ограничениях: неотрицательности и нормированности.



Разведочный поиск: прототип интерфейса

Тематическая сегментация документа запроса

BigARTM — открытая библиотека для тематического моделирования больших коллекций текстовых документов. Параллельная распределенная реализация метода вероятностного тематического моделирования на основе аддитивной регуляризации.

ARTM (англ.) — Additive Regularization for Topic Modeling.

ARTM (рус.) — Аддитивная Регуляризация Тематических Моделей.

Теоретическое введение

Основная статья: Тематическое моделирование

Вероятностное тематическое моделирование — это современный инструмент статистического анализа текстов, предназначенный для выявления тематик коллекций документов. Тематическая модель включает в себя тему: дисперсионное разложение на множество термов, каждый документ — дисперсионное разложение на множество тем. Тематические модели используются для информационного поиска, классификации, категоризации, аннотирования, сегментации текстов.

Тематическая модель — это представление наблюдаемого условного разложения $p(v|d)$ термов (или их эмбедингов) t в документе d коллекции \mathcal{D} :

$$p(v|d) = \sum_{t \in \mathcal{T}} p(v|t)p(t|d),$$

где \mathcal{T} — множество тем;

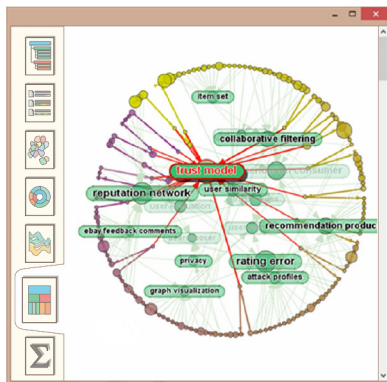
$$\phi_{vt} = p(v|t) \text{ — неизвестное распределение термов в теме } t;$$

$$\theta_{td} = p(t|d) \text{ — неизвестное распределение тем в документе } d.$$

Параметры тематической модели — матрицы $\Phi = (\phi_{vt})$ и $\Theta = (\theta_{td})$ называются путями решения задачи максимизации правдоподобия

$$\sum_{d \in \mathcal{D}} \sum_{v \in \mathcal{V}} n_{dv} \log \sum_{t \in \mathcal{T}} \phi_{vt} \theta_{td} \rightarrow \max_{\Phi, \Theta},$$

при ограничениях нормировки и неотрицательности



Разведочный поиск: прототип интерфейса

Суммаризация документа запроса

BigARTM — открыта библиотека для тематического моделирования больших коллекций текстовых документов. Параллельная распределенная реализация метода вероятностного тематического моделирования на основе аддитивной регуляризации.

ARTM (рус.) — Additive Regularization for Topic Modeling.

ARTM (рус.) — Additive Regularization for Topic Modeling.

Теоретическое введение

Основная статья: Тематическое моделирование

Вероятностное тематическое моделирование — это современный инструмент статистического анализа текстов, предназначенный для вычлечения тематик коллекций документов. Тематическая модель включает в себя: тему, дисперсионное на множестве термов, каждый документ — дисперсионное разложение на множество тем. Тематические модели используются для информационного поиска, классификации, категоризации, аннотирования, сегментации текстов.

Тематическая модель — это представление наблюдаемого условного распределения $p(\mathbf{y}|\mathbf{d})$ термов (слов или словосочетаний) \mathbf{y} в документе \mathbf{d} :

$$p(\mathbf{y}|\mathbf{d}) = \prod_{t \in \mathbf{y}} p_t(y_t|\mathbf{d}),$$

где \mathbf{y} — множество термов.

$\phi_{wt} = p_t(\mathbf{y}|\mathbf{d})$ — неизвестное распределение термов в теме w ;

$\theta_{td} = p(\mathbf{d}|\mathbf{t})$ — неизвестное распределение тем в документе \mathbf{d} .

Параметры тематической модели — матрицы $\Phi = (\phi_{wt})$ и $\Theta = (\theta_{td})$ — находят путем решения задачи максимизации правдоподобия:

$$\sum_{t \in \mathbf{y}} \sum_{w \in \mathcal{W}} \phi_{wt} \log \sum_{t \in \mathbf{y}} \phi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta}.$$

при ограничениях: нормировка и неотрицательности.

Суммаризация «BigARTM»

Тематическое моделирование — одно из современных направлений статистического анализа текста, активно развивающееся последние 10–15 лет. Тематические модели выявляют латентные темы в коллекциях текстовых документов и используются для создания систем семантического поиска, категоризации, суммаризации, сегментации текстов. Основные требования к тематическим моделям: они должны быть хорошо интерпретируемыми (автоматически строить темы, понятные конечным пользователям), мультимодальными (учитывать разнородные метаданные документов), динамическими (выявлять динамику тем во времени), иерархическими (автоматически разделять темы на подтемы), мультиязычными (использовать не только отдельные слова, но и ключевые фразы), и т.д. Библиотека с открытым кодом BigARTM предназначена для построения регуляризованных мультимодальных тематических моделей больших текстовых коллекций.

Технологические элементы разведочного поиска

- 1 Интернет-краулинг имеются готовые решения
- 2 Фильтрация контента имеются готовые решения
- 3 Тематическое моделирование **ведутся исследования**
- 4 Инвертированный индекс имеются готовые решения
- 5 Ранжирование имеются готовые решения
- 6 Визуализация имеются готовые решения

Тематическая модель для разведочного поиска должна быть...

- 1 **Интерпретируемая:** каждая тема понятна людям
- 2 **Мультиграммная:** термины-словосочетания неразрывны
- 3 **Мультимодальная:** авторы, связи, тэги, пользователи, ...
- 4 **Мультиязычная:** для кросс- и много-языкового поиска
- 5 **Динамическая:** выявление истории развития тем
- 6 **Иерархическая:** выявление иерархических связей тем
- 7 **Сегментирующая:** выделение тем внутри документа
- 8 **Обучаемая** по оценкам ассессоров и логам пользователей
- 9 **Определяющая** число тем автоматически
- 10 **Создающая и именующая** новые темы автоматически
- 11 **Онлайновая:** обрабатывающая коллекцию за 1 проход
- 12 **Параллельная, распределённая** для больших коллекций

Поиск этно-релевантных тем в социальных сетях

Основные задачи проекта:

- Разведочный поиск этнических тем в социальных медиа
- Мониторинг этих тем во времени и по регионам
- Оценивание враждебности, конфликтности
- Поддержка социологических исследований

Вспомогательные задачи:

- Фильтрация (обогащение) потока данных
- Обеспечение полноты поиска этнических тем
- Выявление тематических сообществ
- Выделение событийных и региональных тем
- Решение проблемы коротких сообщений

Примеры этнонимов

османский	русич
восточноевропейский	сингапурец
эвенк	перуанский
швейцарская	словенский
аланский	вепсский
саамский	ниггер
латыш	адыги
литовец	сомалиец
цыганка	абхаз
ханты-мансийский	темнокожий
карачаевский	нигериец
кубинка	лягушатник
гагаузский	камбоджиец

Примеры этнических тем

(русские): русский, князь, россия, татарин, великий, царить, царь, иван, император, империя, грозить, государь, век, московская, екатерина, москва,

(русские): акция, организация, митинг, движение, активный, мероприятие, совет, русский, участник, москва, оппозиция, россия, пикет, протест, проведение, националист, поддержка, общественный, проводить, участие,

(славяне, византийцы): славянский, святослав, жрец, древние, письменность, рюрик, летопись, византия, мефодий, хазарский, русский, азбука,

(сирийцы): сирийский, асад, боевик, район, террорист, уничтожить, группировка, дамаск, оружие, алесию, оппозиция, операция, селение, сша, нусра, турция,

(турки): турция, турецкий, курдский, эрдоган, стамбул, страна, кавказ, горин, полиция, премьер-министр, регион, курдистан, ататюрк, партия,

(иранцы): иран, иранский, сша, россия, ядерный, президент, тегеран, сирия, оон, израиль, переговоры, обама, санкция, исламский,

(палестинцы): террорист, израиль, терять, палестинский, палестинец, террористический, палестина, взрыв, территория, страна, государство, безопасность, арабский, организация, иерусалим, военный, полиция, газ,

(ливанцы): ливанский, боевик, район, ливан, армия, террорист, али, военный, хизбалла, раненый, уничтожить, сирия, подразделение, квартал, армейский,

(ливийцы): ливан, демократия, страна, ливийский, каддафи, государство, алжир, война, правительство, сша, арабский, али, муаммар, сирия,

Примеры этнических тем

(евреи): израиль, израильский, страна, война, нетаньяху, тель-авив, время, сша, сирия, египет, случай, самолет, еврейский, военный, ближний,

(американцы): американский, американка, война, россия, военный, страна, вашингтон, америка, армия, конгресс, сирия, союзный, российский, обама, войска, русский, оружие, операция,

(немцы): армия, война, войска, советский, военный, дивизия, немец, фронт, немецкий, генерал, борт, операция, оборона, русский, бог, победа,

(немцы): германий, немец, германский, ссср, немецкий, война, старое, советский, россия, береза, русский, правительство, территория, полный, документ, вопрос, сорт, договор, отношение, франция,

(евреи, немцы): еврей, еврейский, холодный, германий, антисемитизм, гетра, немец, синагога, сша, израиль, малиновского, комиссия, нацбол, документ, война, еврейка, миллион, украина,

(украинцы, немцы): украинский, унс, оун, немец, немецкий, ковальков, хохол, волынский, бандера, организация, россиянин, советский, русский, польский, армия, шухевича, ровенский,

(таджики, узбеки): мигрант, страна, россия, миграция, азия, нелегальный, миграционный, таджикистан, гастарбайтер, гражданка, трудовой, рабочий, фмс, коренево, среднее, узбекистан, таджик, проблема, русский, население,

(канадцы): команда, игра, игрок, канадский, сезон, хоккей, сборная, играть, болельщик, победа, кубок, счет, забирать, хоккейный, выигрывать, хоккеист, чемпионат, шайба,

Примеры этнических тем

(японцы): японский, япония, корея, китайский, жилища, авария, фукусиму, цунами, общаться, океан, станция, хатико, район, правительство, атомный,

(норвежцы): дитя, ребенок, родиться, детский, семья, воспитанный, право, возраст, отец, воспитание, норвежский, родительский, родить, мальчик, взрослый, опека, сын,

(венесуэльцы): куба, кастро, венесуэла, чавес, президент, уго, мадура, боливия, фидель, глава, латинский, венесуэльский, лидер, боливарианской, президентский, альенде, гевару,

(китайцы): китайский, россия, производство, китаи, продукция, страна, предприятие, компания, технология, военный, регион, производить, производственный, промышленность, российский, экономический, кнр,

(азербайджанцы): русский, азербайджан, азербайджанец, россия, азербайджанский, таксист, диаспора, анапа, народ, москва, страна, армянин, слово, рынок,

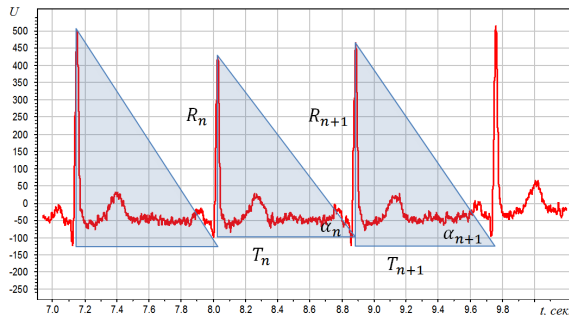
(грузины): грузинский, спецназ, военный, август, баташева, российский, спецназовец, миротворец, операция, румын, бригада, миротворческий, абхазия, группа, войска, русский, цхинвале,

(осетины): конституция, осетия, аминат, русский, осетинский, южный, северный, россия, война, республика, вопрос, алахай, российский, население, конфликт,

(цыгане): наркотик, цыган, цыганка, хороший, место, страна, деньга, время, работать, жизнь, жить, рука, дом, цыганский, наркоманка,

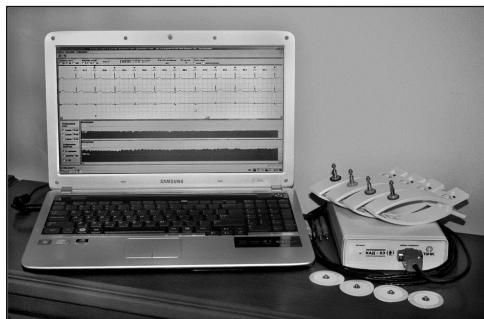
Информационный анализ ЭКГ-сигналов

Теория информационной функции сердца В.М.Успенского:
амплитуды R_n и интервалы T_n кардиоциклов несут
информацию о многих заболеваниях внутренних органов.



$$\alpha_n = \arctg \frac{R_n}{T_n}$$

Диагностическая система «Скринфакс»



- более 15 лет опытной эксплуатации
- более 20 тысяч прецедентов (кардиограмма + диагноз)
- более 30 заболеваний

Объём исходных данных (по заболеваниям)

абсолютно здоровые	A3	193
аденома простаты	ДГПЖ	260
аднексит хронический	АХ	276
анемия железодефицитная	ЖДА	260
асептический некроз головки бедренной кости	НГБК	324
вегетососудистая дистония	ВСД	694
гипертоническая болезнь	ГБ	1894
дискинезия желчевыводящих путей	ДЖВП	717
желчнокаменная болезнь	ЖКБ	278
ишемическая болезнь сердца	ИБС	1265
миома матки	ММ	781
мочекаменная болезнь	МКБ	654
рак общий (онкопатология различной локализации)	РО	530
сахарный диабет (СД1 и СД2)	СД	871
узловой (диффузный) зоб щитовидной железы	УЩ	748
холецистит хронический	ХХ	340
хронический гастрит (гастродуоденит) гиперацидный	ХГ1	324
хронический гастрит (гастродуоденит) гипоацидный	ХГ2	700
язвенная болезнь	ЯБ	785

Этап 1: предобработка. Дискретизация ЭКГ-сигнала

Вход: последовательность интервалов и амплитуд $(T_n, R_n)_{n=1}^N$

Правила кодирования:

если $R_n < R_{n+1}$, $T_n < T_{n+1}$, $\alpha_n < \alpha_{n+1}$ то $S_n = A$

если $R_n \geq R_{n+1}$, $T_n \geq T_{n+1}$, $\alpha_n < \alpha_{n+1}$ то $S_n = B$

если $R_n < R_{n+1}$, $T_n \geq T_{n+1}$, $\alpha_n < \alpha_{n+1}$ то $S_n = C$

если $R_n \geq R_{n+1}$, $T_n < T_{n+1}$, $\alpha_n \geq \alpha_{n+1}$ то $S_n = D$

если $R_n < R_{n+1}$, $T_n < T_{n+1}$, $\alpha_n \geq \alpha_{n+1}$ то $S_n = E$

если $R_n \geq R_{n+1}$, $T_n \geq T_{n+1}$, $\alpha_n \geq \alpha_{n+1}$ то $S_n = F$

Выход: кодограмма $d = (s_n)_{n=1}^{N-1}$ — последовательность символов алфавита $\mathcal{A} = \{A, B, C, D, E, F\}$:

DBEACFDAAFBABDDAADF AAFEEACFEACFB AEFFAABFFA AFFFAAFFA AFFFAEBFBEBFAEFCAFFFAAD
 FCAFFAADFCADFC CDFDACFFACDFAEFFACFFEADFC AFBCADFFECFFA AFFFAAFFFAEFCAFCFAEFFCAD
 DAADBF AAFFAEBFAABFACDFFAAFBAADF AADFDAFFCEFCEDFC EEFCAEFBECBBBAADBAACFFA AFFFA
 CFFCECFDAABDAEFFFAAFFCEDBFAAFFAEFFAEFBACFB AEDFEAFFFCAFFDAAFFAEBDAADBBADFAFF
 EABFCCAFDEEBDECFFACFFAABFAADFBAAFFACFFFAEFFACFFACFFCECFBAAFFFAAFFFAAFFAADF
 AABFACDFDAEFFAADBAAEFFFAFBCECFDECCFBAAFFAADF DACDFAAFFAADFCAADF AEFBAAFFCADFE
 AFFCECFCECFFAAFFABCFDAAFFAADFCAEFFAABFACBFAEBFAEBFAEFCAFFBAAFFA AFFDACFDAAFBF
 CAFFAECFFACFFACDFCADFDAABFAEDDABBFACDDBA AFFFAFFCADFAADF DACFFAEDFCACFAEBCE

Этап 1: предобработка. Векторизация кодограммы

Вход: кодограмма $d = (s_1, \dots, s_{N-1})$ как текстовая строка

DBEACFDAAFBABDDAADF AAFEEACFEACFBREFFAABFFAFAFFAFAFFAEBFBREBF EAFFCAFFAAD
 FCFAFAADF CADFCCDF DACFACDFAEFFACFFEADFCAFBCADFFECFFAFAFFAFAFFAEFFCACFCAEFFCAD
 DAADBF AAF EAEBFAABFACDFFAAFBAADF AADF ADFCECFCEDFCEEFCAEFBECBBBAADBAACFFAFAFFA
 CFFCECFDAABDAEFFAFAFFCEDBFAAFFAEFFAEFBACFBADFEAFAFFCAFFDAAFFAEBDAADBBADF ADF
 EABFCCAFDEEBDECFACFFAABFAADFBAAFFACFFFAEFFACFFACFFCECFBAAFFFAAFFFAADFBA
 AABFACDFDAEFFAABBAEFFEAFBCECFDECCFBAFFFAADF DACDFAAFFAADFCAADF AEFBAFFCADFE
 AFFCECFCECFFAAFFABCFDAAFFAADFCAEFFAABFACBFAEBFAEBFAFFBAFFFAFFDACFDAAABFB
 CAFFAECFFACFFACDFCADFDAABFAEEDABBFACDDBAFAFFCADFAADF DACFFAEDFCACFCAEBCE

Выход: частоты триграмм n_{dw} — сколько раз триграмма w появилась в кодограмме d , $w \in W$, словарь $|W| = 6^3 = 216$

1. FFA - 42	17. EFF - 10	33. CEC - 6	49. EAC - 3
2. FAA - 33	18. DAA - 10	34. ADB - 5	50. DDA - 3
3. AFF - 32	19. ECF - 9	35. FFE - 5	51. CAC - 3
4. AAF - 30	20. FFC - 9	36. EBF - 5	52. EBF - 3
5. ADF - 18	21. FEA - 9	37. CFD - 5	53. EFB - 3
6. FCA - 18	22. DFC - 8	38. AFB - 4	54. DBA - 3
7. ACF - 17	23. ABF - 8	39. AAE - 4	55. FCC - 2
8. AAD - 15	24. AAB - 8	40. CFC - 4	56. AFC - 2
9. CFF - 14	25. FCE - 8	41. CAE - 4	57. EAA - 2
10. AEF - 13	26. AEB - 7	42. DAC - 4	58. CED - 2
11. FDA - 13	27. DFD - 7	43. DBF - 4	59. CAA - 2
12. FAE - 12	28. ACD - 6	44. BFC - 4	60. BCA - 2
13. FAC - 12	29. CDF - 6	45. CFB - 4	61. BBA - 2
14. FBA - 11	30. DFA - 6	46. AED - 3	62. DFF - 2
15. BFA - 11	31. CAF - 6	47. FFF - 3	63. BDA - 2
16. BAA - 11	32. CAD - 6	48. FBC - 3	64. DAE - 2

Этап 2: машинное обучение

Мультимодальная тематическая модель классификации:

- документ \leftrightarrow кодограмма ЭКГ
- модальность №1: слово \leftrightarrow триграмма
- модальность №2: класс \leftrightarrow заболевание
- тема \leftrightarrow диагностический эталон заболевания

диагностические эталоны **состояния нормы**:

topic 1: AED, BCE, CED, DBD, DDC, EDF, EFC, FCA, FCE

topic 2: BCE, CAD, DBD, DDC, EDB, EDF, FCA, FCE

topic 3: AED, CED, DBD, DFC, EDB, EFC, FCE

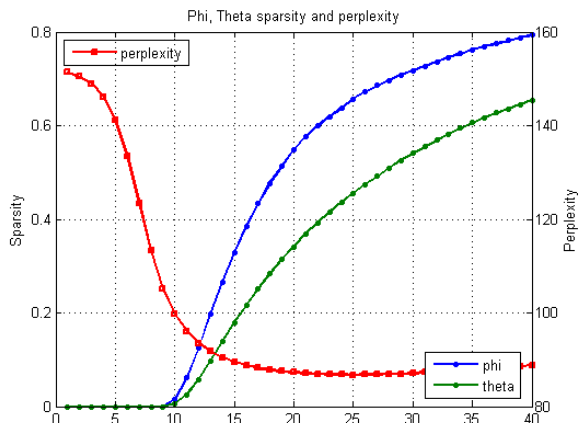
диагностические эталоны **болезни (диабет)**:

topic 1: AFC, CAF, AFA, FAE, AFB, BAF, BAD, EFC, EFA, CFC

topic 2: AFC, CAF, AFA, FAB, ABB, BAF, BCD, EFF

Мониторинг качества модели в EM-алгоритме

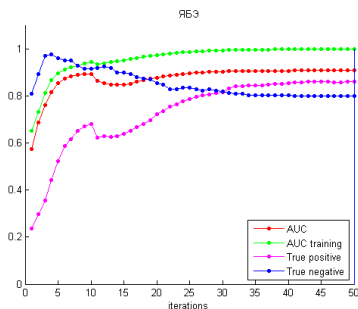
Разреженность матриц Φ , Θ и перплексия модели, при разреживании с 10-й итерации (язвенная болезнь)



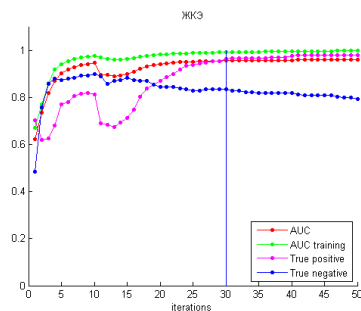
Мониторинг качества модели в EM-алгоритме

Включение разреживающего регуляризатора с 10-й итерации

Язвенная болезнь



Желчнокаменная болезнь

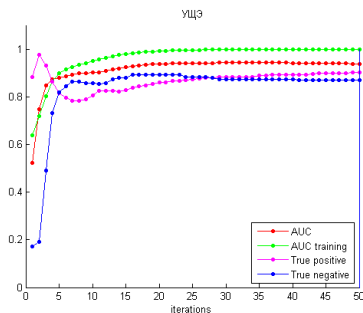
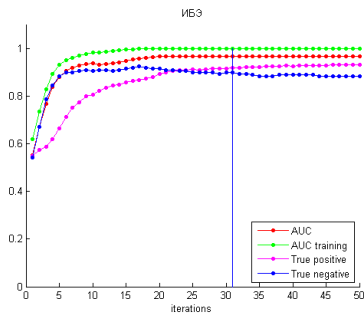


Мониторинг качества модели в EM-алгоритме

Включение разреживающего регуляризатора с 10-й итерации

Ишемическая болезнь сердца

Узловой зоб щитовидной железы



Андрей Шапулин. Регуляризация вероятностных тематических моделей для классификации символьных последовательностей // ВКР бакалавра, 2015. ВМК МГУ.

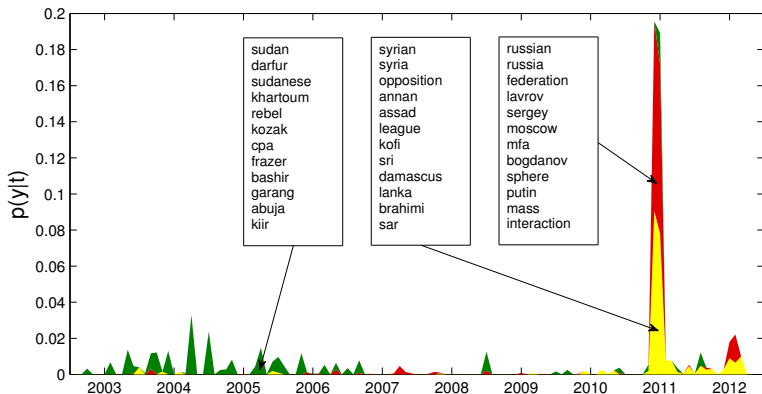
Результаты кросс-валидации

Обучающая выборка — для оптимизации параметров модели
Тестовая выборка — для оценивания чувс., спец., AUC
40×10-fold cross-validation — для доверительного оценивания

болезнь	выборка	AUC, %	C% при Ч=95%
некроз головки бедренной кости	327	99.19 ± 0.10	96.6 ± 1.76
желчнокаменная болезнь	277	98.98 ± 0.23	94.4 ± 1.54
ишемическая болезнь сердца	1262	97.98 ± 0.14	91.1 ± 1.86
гастрит	321	97.76 ± 0.11	88.3 ± 2.64
гипертоническая болезнь	1891	96.76 ± 0.09	84.7 ± 1.99
сахарный диабет	868	96.75 ± 0.19	85.3 ± 2.18
аденома простаты	257	96.49 ± 0.13	80.1 ± 3.19
рак	525	96.49 ± 0.28	82.2 ± 2.38
узловой зоб щитовидной железы	750	95.57 ± 0.16	73.5 ± 3.41
холецистит хронический	336	95.35 ± 0.12	74.8 ± 2.46
дискинезия ЖВП	714	94.99 ± 0.16	70.3 ± 4.67
мочекаменная болезнь	649	94.99 ± 0.11	69.3 ± 2.14
язвенная болезнь	779	94.62 ± 0.10	63.6 ± 2.55

Задача анализа потока пресс-релизов

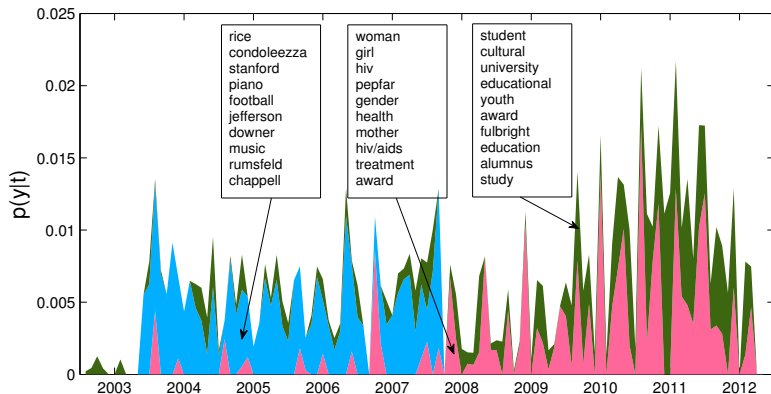
Примеры событийных тем и момента их совместного всплеска



Никита Дойков. Адаптивная регуляризация вероятностных тематических моделей // ВКР бакалавра, 2015. ВМК МГУ.

Задача анализа потока пресс-релизов

Примеры перманентных тем



Никита Дойков. Адаптивная регуляризация вероятностных тематических моделей // ВКР бакалавра, 2015. ВМК МГУ.

Другие коллекции и проекты

- Разведочный поиск на habrahabr.ru и его оценивание
- Кросс-язычный разведочный поиск на arXiv.org и Википедии
- Тематизация коллекции научных статей ММРО/ИОИ
- Совместная тематизация текстов и изображений из соцсетей
- Тематизация картин британского музея и их описаний
- Классификация авторефератов по областям знаний
- Конкурс Kaggle «The Allen AI Science Challenge»
- Поиск мотивов в задачах биоинформатики

BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Онлайновая параллельная мультимодальная ARTM
- Большие данные: коллекция не хранится в памяти
- Встроенная библиотека регуляризаторов

Сообщество:

- Открытый код <https://github.com/bigartm>
(discussion group, issue tracker, pull requests)
- Документация <http://bigartm.org>



Лицензия и среда разработки:

- Freely available for commercial usage (BSD 3-Clause license)
- Cross-platform — Windows, Linux, Mac OS X (32 bit, 64 bit)
- Programming APIs: command-line, C++, and Python

От теории регуляризации к технологии BigARTM


Технология BigARTM упрощает процесс разработки тематической модели с требуемыми свойствами:


Этапы моделирования

Bayesian TM

ARTM

	Bayesian TM	ARTM	
	Анализ требований	Анализ требований	
Формализация:	Вероятностная порождающая модель данных	Стандартные критерии	Свои критерии
Алгоритмизация:	Байесовский вывод для данной порождающей модели (VI, GS, EP)	Общий регуляризованный EM-алгоритм для любых моделей	
Реализация:	Исследовательский код (Matlab, Python, R)	Промышленный код BigARTM (C++, Python API)	
Оценивание:	Исследовательские метрики, исследовательский код	Стандартные метрики	Свои метрики
	Внедрение	Внедрение	

 -- нестандартизируемые этапы, уникальная разработка для каждой задачи

 -- стандартизуемые этапы

Разработка тематических моделей в среде IPython Notebook

<http://nbviewer.ipython.org/github/bigartm/bigartm-book/tree/master/>

Коллекция:

Используем небольшую коллекцию 'kos', доступную в репозитории UCI
<https://archive.ics.uci.edu/ml/machine-learning-databases/bag-of-words/>. Параметры коллекции следующие:

- 3430 документов;
- 6906 слов в словаре;
- 46714 слов в коллекции.

Для начала подключим все необходимые модули (убедитесь, что путь к Python API BigARTM находится в вашей переменной PATH):

```
In [1]: %matplotlib inline
import glob
import matplotlib.pyplot as plt
import artm
```

Прежде всего необходимо подготовить входные данные. BigARTM имеет собственный формат документов для обработки, называемый батчами. В библиотеке присутствуют средства по созданию батчей из файлов Bag-Of-Words в форматах UCI и Wowpal Wabbit (подробности можно найти в <http://docs.bigartm.org/en/latest/formats.html>).

В Python API, по аналогии с алгоритмами из scikit-learn, входные данные представлены одним классом BatchVectorizer. Объект этого класса принимает на вход батчи или файлы с Bag-Of-Words и подается на вход всем методам. В случае, если входные данные не являются батчами, он создаст их и сохранит на диск для последующего быстрого использования.

Итак, создадим объект BatchVectorizer:

```
In [2]: batch_vectorizer = None
if len(glob.glob('kos' + '/*.*.batch')) < 1:
    batch_vectorizer = artm.BatchVectorizer(data_path='', data_format='bow
_uci', collection_name='kos', target_folder='kos')
else:
    batch_vectorizer = artm.BatchVectorizer(data_path='kos', data_format='
batches')
```

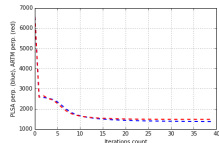
ARTM — это класс, представляющий собой Python API BigARTM, и позволяющий использовать практически все возможности библиотеки в стиле scikit-learn. Создадим две тематические модели для нашего эксперимента. Наиболее важным параметром модели является число тем. Опционально можно указать список регуляризаторов и функционалов качества, которые следует использовать для данной модели. Если этого не сделать, то регуляризаторы и функционалы всегда можно добавить позднее. Обратите внимание, что каждая модель задаёт

Продолжим обучение моделей, инициализовав 25 проходов по коллекции, после чего снова посмотрим на значения функционалов качества:

```
In [11]: model_plsa.fit_offline(batch_vectorizer=batch_vectorizer, num_collection_p
asses=25, num_document_passes=1)
model_artm.fit_offline(batch_vectorizer=batch_vectorizer, num_collection_p
asses=25, num_document_passes=1)
```

```
In [12]: print_measures(model_plsa, model_artm)
```

```
Sparsity Phi: 0.332 (FLSA) vs. 0.740 (ARTM)
Sparsity Theta: 0.082 (FLSA) vs. 0.602 (ARTM)
Kernel contrast: 0.530 (FLSA) vs. 0.548 (ARTM)
Kernel purity: 0.396 (FLSA) vs. 0.531 (ARTM)
Perplexity: 1362.804 (FLSA) vs. 1475.455 (ARTM)
```



Кроме того, для наглядности построим графики изменения разреженностей матриц по итерациям:

```
In [13]: plt.plot(xrange(model_plsa.num_phi_updates), model_plsa.score_tracker['Spa
rsityPhiScore'].value, 'b--',
               xrange(model_artm.num_phi_updates), model_artm.score_trac
ker['SparsityPhiScore'].value, 'r--', linewidth=2)
plt.xlabel('Iterations count')
plt.ylabel('FLSA Phi sp. (blue), ARTM Phi sp. (red)')
plt.grid(True)
plt.show()
```

```
plt.plot(xrange(model_plsa.num_phi_updates), model_plsa.score_tracker['Spa
rsityThetaScore'].value, 'b--',
         xrange(model_artm.num_phi_updates), model_artm.score_trac
ker['SparsityThetaScore'].value, 'r--', linewidth=2)
```

Эксперимент. Обгоняем конкурентов по скорости

- 3.7M статей английской Вики, 100K уникальных слов

	procs	train	inference	perplexity
BigARTM	1	35 min	72 sec	4000
Gensim.LdaModel	1	369 min	395 sec	4161
VowpalWabbit.LDA	1	73 min	120 sec	4108
BigARTM	4	9 min	20 sec	4061
Gensim.LdaMulticore	4	60 min	222 sec	4111
BigARTM	8	4.5 min	14 sec	4304
Gensim.LdaMulticore	8	57 min	224 sec	4455

- *procs* = число параллельных потоков
- *inference* = время тематизации 100K тестовых документов
- *perplexity* вычислена на тестовой выборке документов

Направления дальнейших исследований

- научиться строить 50 тысяч хорошо интерпретируемых тем
- научиться автоматически создавать и именовать темы
- обеспечивать интерпретируемость, устойчивость и полноту
- соединить лингвистическую регуляризацию и word2vec
- применять гиперграфовые модели к данным соцсетей
- разработать визуальные средства систематизации знаний
- создать систему тематического разведочного поиска

Задачи на синтетических данных

- тематическая сегментация
- выделение фоновых слов