

# Вероятностные тематические модели

## Лекция 2. Примеры прикладных задач и проект «Тематизатор»

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)»

МФТИ – ФИЦ ИУ РАН • 14 сентября 2023

- 1 Вероятностное тематическое моделирование**
  - Постановка задачи
  - Теория ARTM и библиотека BigARTM
  - Визуализация тематических моделей
- 2 Прикладные задачи тематического моделирования**
  - Поиск этно-релевантных тем в социальных сетях
  - Анализ программ развития российских вузов
  - Социо-гуманитарные исследования
- 3 Проект «Тематизатор»**
  - Мотивации и приложения
  - Анализ требований
  - MVP: минимально жизнеспособный Тематизатор

## Постановка задачи тематического моделирования

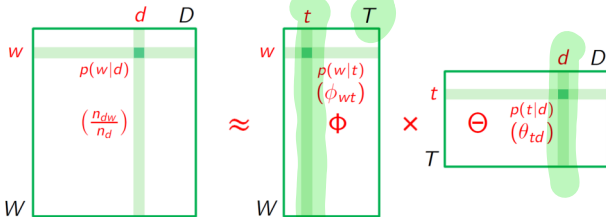
**Дано:** коллекция текстовых документов

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

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

- $\phi_{wt} = p(w|t)$  — вероятности термов  $w$  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

**Критерий:** максимум лог-правдоподобия с регуляризатором

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

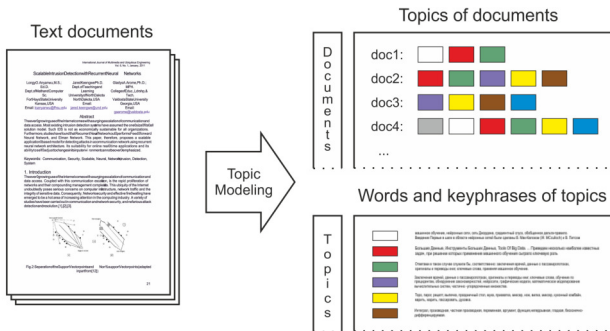
EM-алгоритм: метод простой итерации для системы уравнений

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

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

# Мультимодальная тематическая модель

Тема может порождать термины различных *модальностей*:  
 $p(\text{слово} | t)$ ,  $p(n\text{-грамма} | t)$ ,

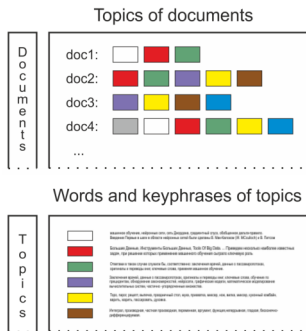


## Мультимодальная тематическая модель

Тема может порождать термины различных *модальностей*:

$p(\text{слово}|t)$ ,  $p(n\text{-грамма}|t)$ ,  $p(\text{автор}|t)$ ,  $p(\text{время}|t)$ ,  $p(\text{источник}|t)$ ,

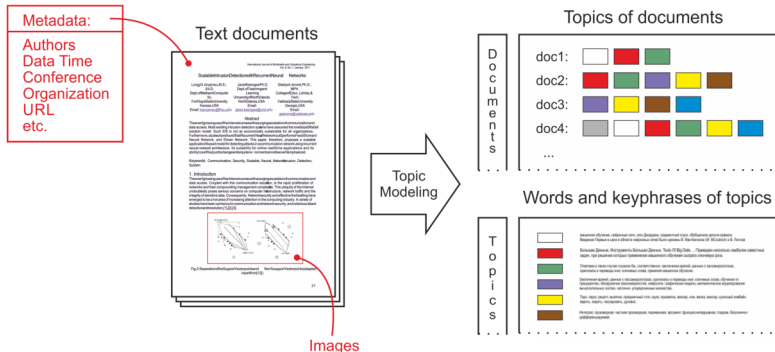
Metadata:  
 Authors  
 Data Time  
 Conference  
 Organization  
 URL  
 etc.



## Мультимодальная тематическая модель

Тема может порождать термины различных *модальностей*:

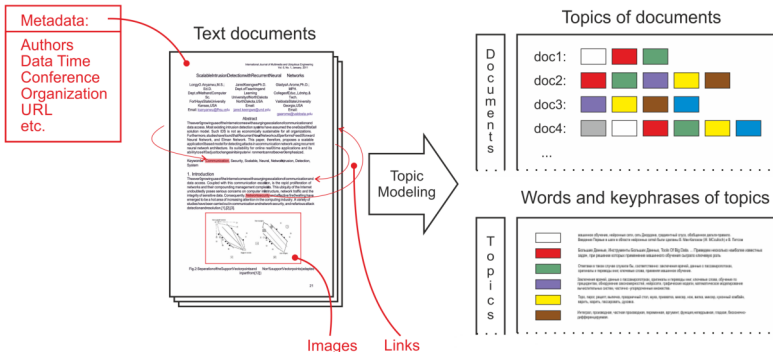
$p(\text{слово} | t)$ ,  $p(n\text{-грамма} | t)$ ,  $p(\text{автор} | t)$ ,  $p(\text{время} | t)$ ,  $p(\text{источник} | t)$ ,  
 $p(\text{объект} | t)$ ,



## Мультимодальная тематическая модель

Тема может порождать термины различных *модальностей*:

$p(\text{слово} | t)$ ,  $p(n\text{-грамма} | t)$ ,  $p(\text{автор} | t)$ ,  $p(\text{время} | t)$ ,  $p(\text{источник} | t)$ ,  
 $p(\text{объект} | t)$ ,  $p(\text{ссылка} | t)$ ,

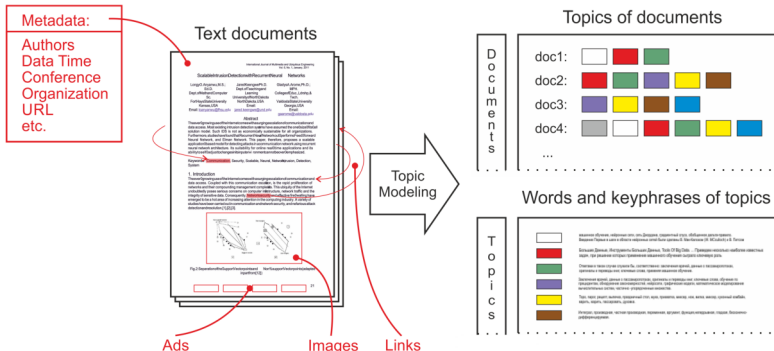




## Мультимодальная тематическая модель

Тема может порождать термины различных *модальностей*:

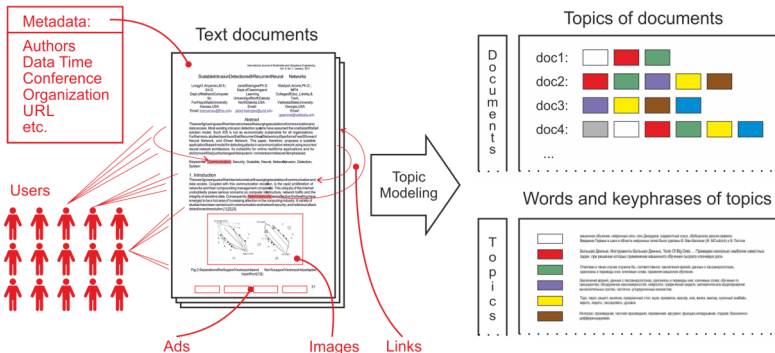
$p(\text{слово} | t)$ ,  $p(n\text{-грамма} | t)$ ,  $p(\text{автор} | t)$ ,  $p(\text{время} | t)$ ,  $p(\text{источник} | t)$ ,  
 $p(\text{объект} | t)$ ,  $p(\text{ссылка} | t)$ ,  $p(\text{баннер} | t)$ ,



## Мультимодальная тематическая модель

Тема может порождать термины различных *модальностей*:

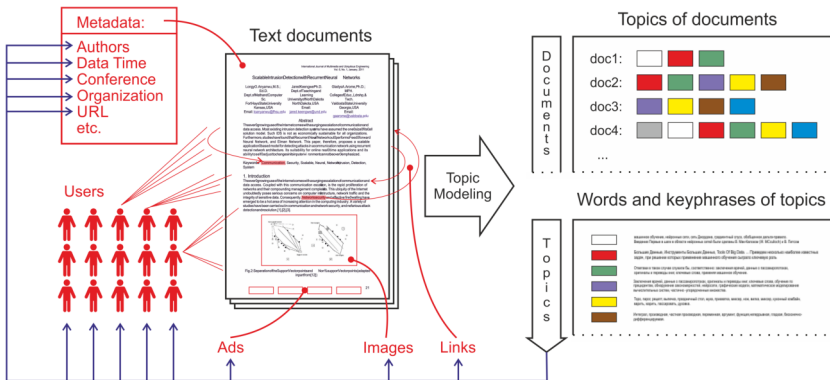
$p(\text{слово} | t)$ ,  $p(n\text{-грамма} | t)$ ,  $p(\text{автор} | t)$ ,  $p(\text{время} | t)$ ,  $p(\text{источник} | t)$ ,  
 $p(\text{объект} | t)$ ,  $p(\text{ссылка} | t)$ ,  $p(\text{баннер} | t)$ ,  $p(\text{пользователь} | t)$



## Мультимодальная тематическая модель

Тема может порождать термины различных *модальностей*:

$p(\text{слово} | t)$ ,  $p(n\text{-грамма} | t)$ ,  $p(\text{автор} | t)$ ,  $p(\text{время} | t)$ ,  $p(\text{источник} | t)$ ,  
 $p(\text{объект} | t)$ ,  $p(\text{ссылка} | t)$ ,  $p(\text{баннер} | t)$ ,  $p(\text{пользователь} | t)$



## Мультимодальная ARTM

$W_m$  — словарь термов  $m$ -й модальности,  $m \in M$

Максимизация суммы log-правдоподобий с регуляризацией:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

*K. Vorontsov, O. Freij, M. Apishev et al.* Non-Bayesian additive regularization for multimodal topic modeling of large collections. CIKM TM workshop, 2015.

## BigARTM: библиотека тематического моделирования

### Ключевые возможности:

- Большие данные: коллекция не хранится в памяти
- Онлайн-параллельный мультимодальный ARTM
- Встроенная библиотека регуляризаторов и мер качества

### Сообщество:

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



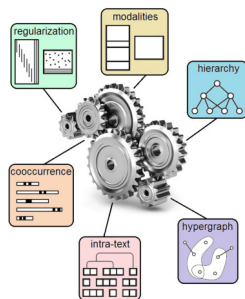
### Лицензия и среда разработки:

- Свободная коммерческая лицензия (BSD 3-Clause)
- Кросс-платформенность: Windows, Linux, MacOS (32/64 bit)
- Интерфейсы API: command-line, C++, and Python

## Ключевые возможности библиотек BigARTM и TopicNet

### BigARTM

- библиотека регуляризаторов
- мультимодальные модели
- иерархические модели
- гиперграфовые модели
- модели связности текста



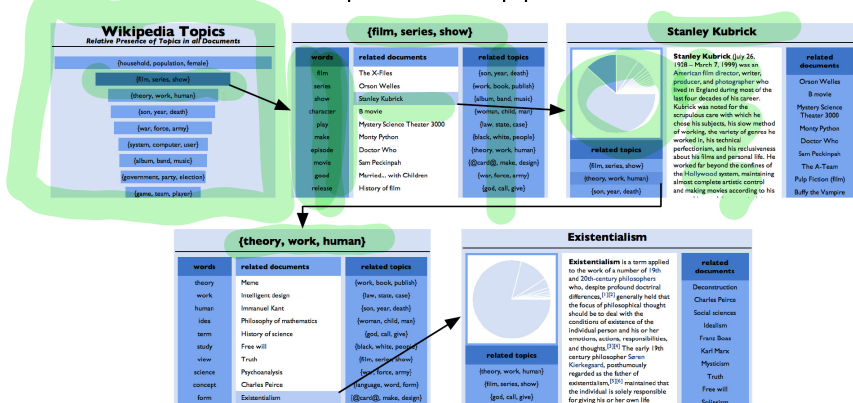
### TopicNet

- Перебор сценариев регуляризации для выбора моделей
- Автоматическое протоколирование экспериментов
- Построение «банка тем» из множества моделей
- Визуализация результатов тематического моделирования

V.Bulatov, E.Egorov, E.Veselova, D.Polyudova, V.Alekseev, A.Goncharov, K.Vorontsov.  
TopicNet: making additive regularisation for topic modelling accessible. LREC-2020

# Визуализация TMVE (Topic Model Visualization Engine)

Тематический навигатор с веб-интерфейсом:

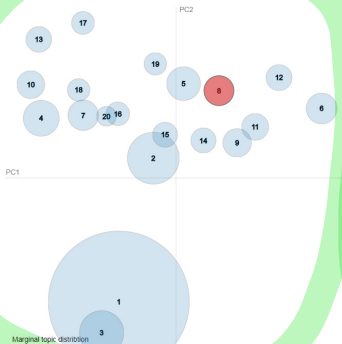


<https://github.com/ajbc/tmv>

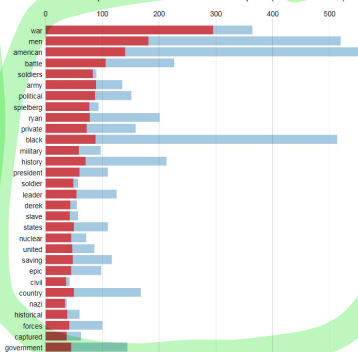
## Система LDAvis

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

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 8 (2.4% of tokens)

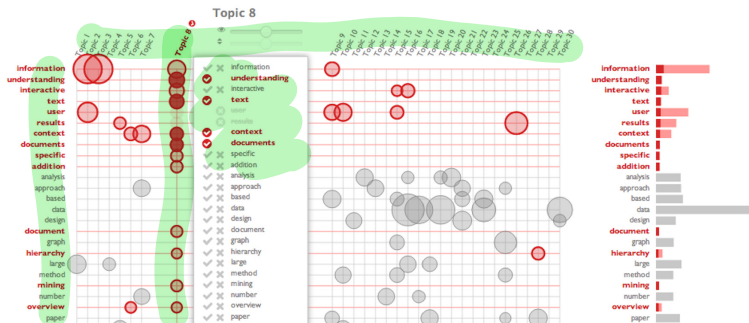


<https://github.com/cpsievert/LDAvis>



## Система Termite

Интерактивная визуализация матрицы  $\Phi$  и сравнение тем:

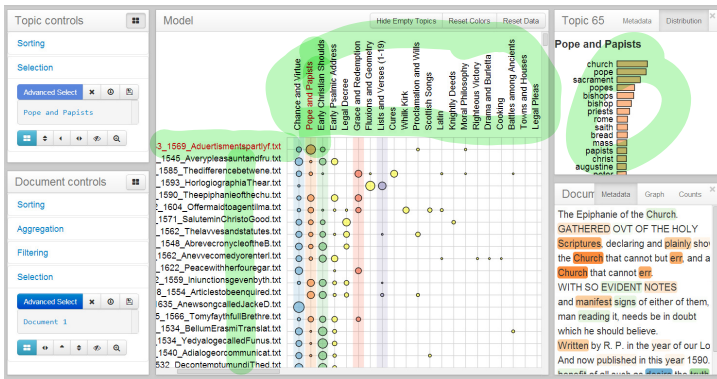


<https://github.com/uwdata/termite-visualizations>

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

## Система Serendip

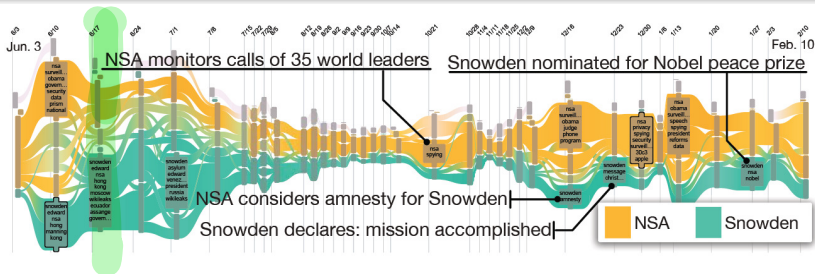
Визуализация матриц  $\Phi$ ,  $\Theta$  и тематики слов в текстах:



<http://vep.cs.wisc.edu/serendip>

E.Alexander, J.Kohlmann, R.Valenza, M.Witmore, M.Gleicher. Serendip: Topic Model-Driven Visual Exploration of Text Corpora. IEEE VAST 2014.

## Динамика тем: эволюция предметной области



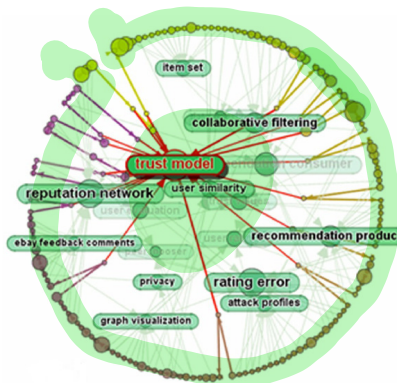
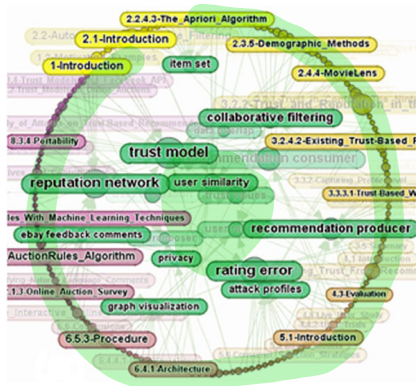
Эволюция выбранных тем иерархии. Данные Prism (2013/06/03–2014/02/09)

Стратегия визуализации в системах TextFlow и RoseRiver:

- эксперт задаёт сечение иерархии (дерева) тем,
- интерактивно выбирает подмножество тем и событий,
- получает сгенерированный отчёт с инфографикой.

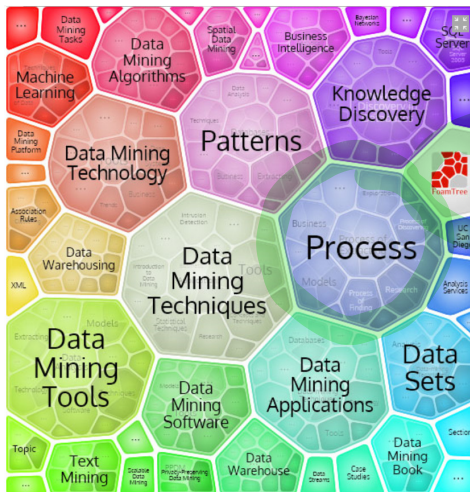
*Weiwei Cui, Shixia Liu, Zhuofeng Wu, Hao Wei.* How hierarchical topics evolve in large text corpora. 2014.

## Динамика тем внутри документа: тематическая сегментация



Gretarsson B., O'Donovan J., Bostandjiev S., Hollerer T., Asuncion A., Newman D., Smyth P. TopicNets: visual analysis of large text corpora with topic modeling. ACM Trans. on Intelligent Systems and Technology. 2012.

## Пример иерархической карты области *Data Mining*



FoamTree: <https://carrotsearch.com/foamtree>

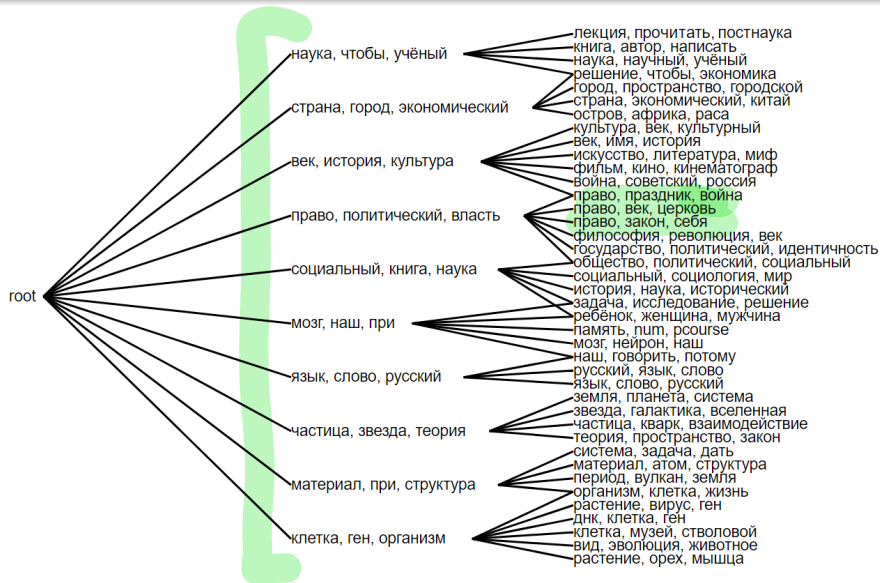
## Библиотека VisARTM для BigARTM (уже не поддерживается)

- Web-приложение для визуализации ARTM моделей
- Открытый код: <https://github.com/bigartm/visartm>
- Автоматическое перестроение моделей через BigARTM
- Текстовые интерактивные визуализации документов, тем, термов, модальностей
- Графическая визуализация иерархических моделей
- Графическая визуализация темпоральных моделей
- Тематические спектры
- Сбор ассессорских оценок тем

---

*Дмитрий Федоряка.* Технология интерактивной визуализации тематических моделей. Бакалаврская диссертация. МФТИ, 2017.

## Иерархический спектр тем (коллекция postnauka.ru)



## Построение спектра тем. Постановка задачи

*Тематический спектр* — такая перестановка тем  $t_1, \dots, t_{|T|}$ , что сумма расстояний между соседними темами минимальна:

$$\sum_{i=2}^{|T|} \rho(t_i, t_{i-1}) \rightarrow \min$$

*Функция расстояния*  $\rho(t, t')$  между темами, примеры:

- Манхэттенское:  $\rho(t, t') = \sum_{w \in W} |\phi_{wt} - \phi_{wt'}|$
- Хеллингера:  $\rho^2(t, t') = \frac{1}{2} \sum_{w \in W} (\sqrt{\phi_{wt}} - \sqrt{\phi_{wt'}})^2$
- Жаккара:  $\rho(t, t') = 1 - \frac{|W_t \cap W_{t'}|}{|W_t \cup W_{t'}|}$ ,  $W_t = \{w : \phi_{wt} > \frac{1}{|W|}\}$



## Построение спектра тем — это задача коммивояжёра

### Задача TSP (traveling salesman problem)

Найти путь минимальной суммарной стоимости, соединяющий  $T$  городов так, чтобы в каждом городе побывать один раз.

Алгоритм Лина–Кернигана в реализации Хельсгауна — лучший для решения задачи TSP, по данным *Encyclopedia of operations research* на 2013 год.

Вычислительная сложность  $T^{2.2}$ .

Другие алгоритмы оказались не только медленнее, но и хуже по качеству тематических спектров.

---

*Keld Helsgaun*. An effective implementation of the Lin–Kernighan traveling salesman heuristic. EJOR, 2000.

*Дмитрий Федоряка*. Технология интерактивной визуализации тематических моделей. Бакалаврская диссертация. МФТИ, 2017.

Источники вдохновения: <http://textvis.lnu.se>

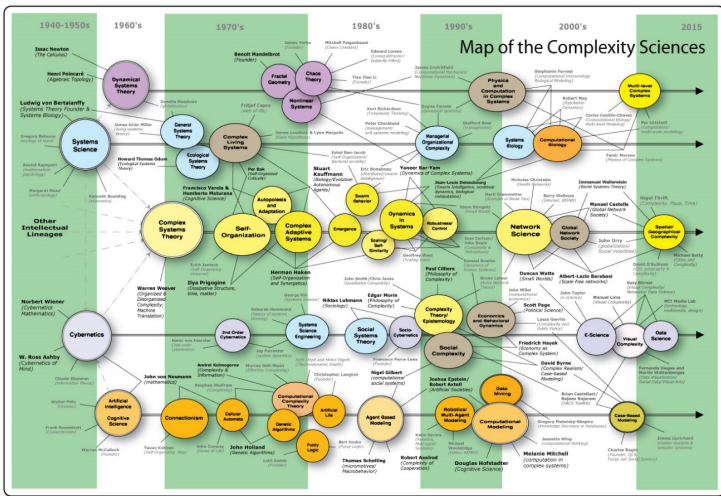
Интерактивный обзор 440 средств визуализации текстов



*Shixia Liu, Weiwei Cui, Yingcai Wu, Mengchen Liu. A survey on information visualization: recent advances and challenges. 2014.*

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

# Пример карты предметной области, построенной вручную



<http://www.theoryculturesociety.org/brian-castellani-on-the-complexity-sciences>

## Тематическая карта научных знаний (концепт)

- **Интерпретация осей:** темы/время/важность/сложность
- **Иерархичность:** темы делятся на подтемы
- **Спектр тем:** гуманитарные → естественные → точные
- **Интерактивность:** реализация мантры Шнейдермана
- **Суммаризация:** масштаб карты определяет объём текста



## Поиск этно-релевантных тем в социальных сетях

- **Дано:**

- 1) данные социальных медиа (ВК и др.)
- 2) словарь ~~этнонимов~~ (около 300)

- **Найти:**

- 1) как можно больше тем про этничности
- 2) темы с сочетанием этничностей (возможные конфликты)

- **Критерий:**

- 1) интерпретируемость всех тем
- 2) точность и полнота поиска этно-релевантных тем

### Используемые регуляризаторы:

- сглаживание этно-релевантных тем по словарю этнонимов
- декоррелирование этно-релевантных тем
- модальность этнонимов

## Примеры этнонимов (всего около 300)

|                     |            |
|---------------------|------------|
| 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 этно-релевантных 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, уничтожать, 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, уничтожать, сирия, подразделение, квартал, армейский,

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

## Примеры этно-релевантных 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 этно-релевантных 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,

## Результат: модель ARTM находит больше этно-тем

Число этно-релевантных тем, найденных моделью:

| модель | этно-тем | фон.тем | ++        | +-        | -+        | всего      |
|--------|----------|---------|-----------|-----------|-----------|------------|
| PLSA   | 300      |         | 9         | 11        | 18        | 38         |
| PLSA   | 400      |         | 12        | 15        | 17        | 44         |
| ARTM-1 | 200      | 100     | 18        | 33        | 20        | 71         |
| ARTM-1 | 250      | 150     | 21        | 27        | 20        | 68         |
| ARTM-2 | 200      | 100     | 28        | 23        | 23        | 74         |
| ARTM-2 | 250      | 150     | <b>38</b> | <b>42</b> | <b>30</b> | <b>104</b> |

Регуляризаторы ARTM-1:

**этно темы:** разреживание, декоррелирование, сглаживание этнонимов

**фоновые темы:** сглаживание, разреживание этнонимов

Регуляризаторы ARTM-2:

ARTM-1 + **модальность этнонимов**

---

*M. Apishev, S. Koltcov, O. Koltsova, S. Nikolenko, K. Vorontsov. Additive regularization for topic modeling in sociological studies of user-generated text content. MICAI 2016.*

## Аналогичные по структуре исследования

### Задачи «поиска и классификации иголок в стоге сена»

- поиск и кластеризация новостей [1]
- поиск в социальных медиа информации, связанной с болезнями, симптомами и методами лечения [2]
- поиск чатов, связанных с преступностью и экстремизмом [3, 4]
- поиск выступлений о правах человека в ООН [5]

- 
1. *J.Jagarlamudi, H.Daumé III, R.Udupa*. Incorporating lexical priors into topic models. 2012.
  2. *M.Paul, M.Dredze*. Discovering health topics in social media using topic models. 2014.
  3. *M.A.Basher, A.Rahman, B.C.M.Fung*. Analyzing topics and authors in chat logs for crime investigation. 2014.
  4. *A.Sharma, M.Pawar*. Survey paper on topic modeling techniques to gain useful forecasting information on violant extremist activities over cyber space. 2015.
  5. *Kohei Watanabe, Yuan Zhou*. Theory-driven analysis of large corpora: semisupervised topic classification of the UN speeches. 2022.

## Анализ программ развития российских вузов

**Цель** — выявить закономерности в стратегиях развития вузов, не читая всех этих документов (Distant Reading)

- **Дано:**  
программам развития ВУЗов: 396 файлов, 284 вуза
- **Найти:**  
полный тематический спектр направлений развития
- **Критерий:**  
интерпретируемость тем;  
чёткого количественного критерия нет :(

## Пример интерпретации темы

**(слова):** инновационный исследование результат региональный предприятие проведение основа среда внедрение уровень рамка сфера исследовательский научно научно-исследовательский участие приоритетный специалист цель выполнение международный прикладной ведущий взаимодействие

**(биграммы):** научный \_ исследование инновационный \_ деятельность приоритетный \_ направление научно \_ исследовательский исследование \_ разработка развитие \_ инновационный фундаментальный \_ прикладной разработка \_ внедрение направление \_ развитие мировой \_ уровень научно \_ образовательный исследовательский \_ деятельность инновационный \_ развитие малое \_ инновационный инновационный \_ предприятие научный \_ инновационный модернизация \_ научно-исследовательский прикладной \_ исследование инновационный \_ проект развитие \_ научный инновационный \_ инфраструктура проведение \_ научный

**(ИНТЕРПРЕТАЦИЯ):** научные исследования и инновационное развитие

## Пример интерпретации темы

**(слова)**: международный число количество участие конференция  
зарубежный увеличение учёный академический мобильность конкурс  
сотрудничество грант иностранный аспирант совместный молодая  
ведущий специалист привлечение преподаватель исследование школа  
сотрудник семинар

**(биграммы)**: увеличение\_ количество академический\_ мобильность\_  
увеличение\_ число международный\_ деятельность\_  
международный\_ сотрудничество международный\_ научный  
развитие\_ международный принять\_ участие российский\_ международный  
научный\_ мероприятие международный\_ образовательный\_  
участие\_ международный иностранный\_ студент количество\_ студент  
научный\_ проект университет\_ международный международный\_ уровень  
международный\_ академический количество\_ участник  
научный\_ конференция программа\_ академический участие\_ студент

**(ИНТЕРПРЕТАЦИЯ)**: академическая мобильность и международное  
сотрудничество

## Пример интерпретации темы

**(слова)**: общежитие корпус здание ремонт площадь инфраструктура комплекс помещение строительство объект капитальный кампус имущественный спортивный реконструкция безопасность территория сооружение место оборудование современный замена учебно-лабораторный комфортный

**(биграммы)**: учебный\_корпус капитальный\_ремонт имущественный\_комплекс общий\_площадь здание\_сооружение студенческий\_общежитие корпус\_общежитие развитие\_имущественный инфраструктура\_университет создание\_комфортный развитие\_инфраструктура университетский\_кампус комплекс\_университет спортивный\_комплекс студент\_сотрудник объект\_университет земельный\_участок условие\_проживание территория\_университет объект\_инфраструктура социальный\_инфраструктура использование\_имущественный строительство\_новый ремонтный\_работа общежитие\_университет

**(ИНТЕРПРЕТАЦИЯ)**: инфраструктура, кампус, строительство

## Интерпретация всех 50 тем

- Для интерпретируемости тем важны биграммы
- Модель построили примерно с 10-й попытки (подбирали число тем, регуляризацию, добивались различности тем)
- Интерпретация 50 тем заняла примерно 20 минут работы
- Иногда выделялись темы исследований и разработок, но для этого нужна более гранулированная модель
- Темы были сгруппированы вручную по 5 категориям:
  - 1 16 тем про науку, инновации и сотрудничество
  - 2 14 тем про образование и кадровый потенциал
  - 3 11 тем про административное управление и хозяйство вуза
  - 4 3 темы «юридические», о самой стратегии развития
  - 5 6 тем «малые и мусорные», вместе не более 5% контента



## Интерпретация всех 50 тем

| доля контента | доля вузов |          | название темы  |
|---------------|------------|----------|--|
|               | более 2%   | более 5% |  |
| 7             | 95         | 67       | научные исследования и инновационное развитие            |
| 12            | 92         | 39       | стратегия развития                                       |
| 15            | 84         | 23       | академическая мобильность и международное сотрудничество |
| 19            | 82         | 17       | кадровой потенциал и кадровая политика                   |
| 22            | 80         | 14       | иностранные студенты                                     |
| 27            | 75         | 30       | образовательные программы                                |
| 30            | 75         | 13       | повышение квалификации и переподготовка кадров           |
| 33            | 70         | 10       | система управления вузом                                 |
| 36            | 68         | 16       | учебный процесс  |
| 39            | 62         | 15       | финансы и бюджет   |
| 43            | 62         | 21       | бюрократия   |
| 45            | 56         | 3        | подготовка высококвалифицированных кадров                |
| 48            | 47         | 9        | инфраструктура, кампус, строительство                    |
| 50            | 44         | 4        | меры повышения качества образования                      |
| 52            | 42         | 4        | влияние на экономику региона                             |
| 54            | 41         | 8        | молодежная политика                                      |
| 56            | 41         | 6        | центры компетенций и технологического превосходства      |
| 58            | 40         | 6        | отсылки к стратегическим документам и НПА                |
| 60            | 36         | 1        | работа со школьниками и талантливой молодежью            |
| 62            | 34         | 7        | ректорат и органы управления вузом                       |
| 64            | 30         | 5        | материально-техническая база вуза                        |
| 65            | 29         | 2        | связь с общественностью, имидж вуза                      |
| 67            | 29         | 8        | исследования с/х, лес, химия, ит                         |
| 69            | 29         | 1        | публикационная активность и защиты диссертаций           |
| 71            | 29         | 2        | взаимодействие с региональной властью                    |

| доля контента | доля вузов |          | название темы  |
|---------------|------------|----------|--|
|               | более 2%   | более 5% |  |
| 72            | 27         | 1        | образовательные программы, аккредитация, профстандарты       |
| 74            | 25         | 3        | спортивная и культурная жизнь вуза                           |
| 75            | 21         | 5        | стратегия развития и региональная среда                      |
| 77            | 20         | 1        | образовательный процесс и образовательные технологии         |
| 78            | 19         | 1        | международное сотрудничество и договорные отношения          |
| 79            | 19         | 2        | цифровизация и цифровые технологии                           |
| 81            | 18         | 2        | медицинское обеспечение, обучение инвалидов                  |
| 82            | 18         | 5        | блоки мероприятий и показатели результативности              |
| 84            | 18         | 5        | работа структурных подразделений вуза                        |
| 85            | 17         | 2        | выход в мировые рейтинги университетов                       |
| 86            | 14         | 1        | технологии транспорта и искусственного интеллекта            |
| 87            | 13         | 1        | публикационная и издательская деятельность                   |
| 88            | 12         | 1        | финансовое и ресурсное обеспечение программы развития        |
| 89            | 11         | 1        | мониторинг показателей эффективности                         |
| 90            | 11         | 0        | сетевые образовательные программы, ворлдскиллс               |
| 92            | 11         | 1        | региональные особенности приёма и рынка труда                |
| 93            | 10         | 1        | приём абитуриентов   |
| 93            | 10         | 0        | исследования в экологии и медицине                           |
| 94            | 9          | 1        | образовательные программы (частные вопросы)                  |
| 95            | 8          | 1        | частные и региональные проблемы                              |
| 96            | 8          | 2        | авиационные технологии                                       |
| 97            | 8          | 0        | смесь тем  |
| 98            | 7          | 0        | образовательные программы & урбанистика и туризм (смесь тем) |
| 99            | 7          | 1        | смесь тем  |
| 100           | 7          | 1        | частные юридические вопросы                                  |

- 16 тем — наука и инновации
- 14 тем — образование и кадры
- 11 тем — управление и хозяйство
- 3 темы — о стратегии развития
- 6 тем — мелкие мусорные



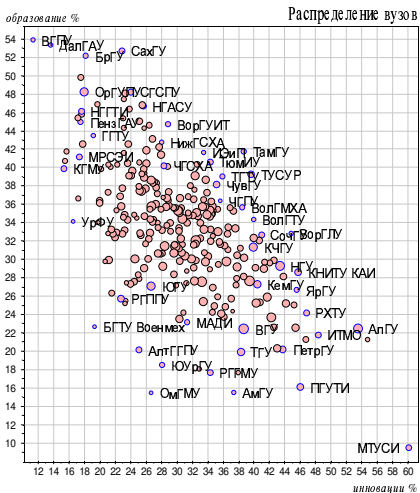
## Тематическая карта вузов

### По осям:

- объёмная доля тем
- про инновации
- про образование

### Вывод:

объёмные доли тем, возможно, показывают баланс приоритетов развития ...хотя... это похоже на оценивание научного отчёта толщиной в сантиметрах :)



## Проекты Школы Прикладного Анализа Данных (ноябрь, 2022)

### Исходные данные ВКонтакте (через сервисы Крибрум)

- Анализ социального влияния на формирование образа правильного питания у студентов г. Томска
- Анализ научной и публикационной активности сотрудников университета или научной организации
- Анализ практик участия в читательских сообществах, формирующихся вокруг авторов или жанров
- Анализ социального и политического взаимодействия сетевых сообществ в регионах ресурсного типа
- Анализ туристической активности и оценка портрета потенциального туриста — путешественника по Камчатке
- Анализ корпуса текстов образовательных дисциплин: программа курса + материалы курса + отчёты студентов
- Анализ научной педагогической литературы для построения карт компетенций

## Исторические исследования: газетные архивы

- [1] Корпус *Pennsylvania Gazette* 1728–1800, 25М слов:
  - выделение последовательности событийных тем;
  - изучение синхронности событий;
  - комбинирование автоматического анализа и ручного.
- [2] Газеты *Техаса* от гражданской войны до наших дней:
  - выделение всех тем, связанных с хлопком;
  - построение серии моделей в скользящих окнах;
  - важность качественной предобработки текстов.
- [3] Газеты и периодика Финляндии (1854–1917):
  - выделение тем о церкви, религии, образовании;
  - тренды модернизации и секуляризации финского общества.

- 
1. *D.Newman, S.Block*. Probabilistic topic decomposition of an eighteenth-century American newspaper. 2006.
  2. *Tze-I Yang, A.J.Torget, R.Mihalcea*. Topic modeling on historical newspapers. 2011.
  3. *J.Marjanen et al*. Topic modelling discourse dynamics in historical newspapers. 2021.

## Исторические исследования: летописи и дневники

- [1] Двужычный корпус книг на английском и немецком:
  - все темы, связанные с эпистемологией
  
- [2] Корпус текстов на китайском языке (1644–1912):
  - все темы, связанные с бандитизмом, преступлениями;
  - необходим контекст для установления типа преступления;
  - важность правильной токенизации для китайского языка.
  
- [3] Дневник Martha Ballard (1735–1812), охватывает 27 лет:
  - выделение событийных и перманентных тем;
  - выделение персональных и исторических тем;
  - специфичный английский XVIII века.

- 
1. *M. Erlin*. Topic modeling, epistemology, and the English and German novel. 2017.
  2. *Ian Matthew Miller*. Rebellion, crime and violence in Qing China, 2013.
  3. *Cameron Blevins*.  
<http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary>.

## Исторические исследования: научная и литературная периодика

Статьи коллекции JSTOR доступны в виде «мешков слов».

[1] Научные журналы XX века:

- различия тематики на английском и немецком языках;
- особенно исследовались различия, связанные со 2МВ;
- для объединения тем использовались интервики Википедии.

[2] Более 100 лет литературно-художественной периодики:

- как менялись темы;
- как менялись значения слов внутри каждой темы;
- как менялась тема насилия (violence, power, fear, blood, death, murder, act, guilt).

---

1. *D.Mimno*. Computational historiography: Data mining in a century of classics journals. 2012.

2. *A.Goldstone, T.Underwood*. The quiet transformations of literary studies: What thirteen thousand scholars could tell us. 2014.

## Проект «Тематизатор»: общие требования

Переход от библиотек (BigARTM, VisARTM, TopicNet) к приложению «Тематизатор» для конечного пользователя — аналитика в области цифровых гуманитарных исследований

- 1 Цели пользователя — разведочный анализ, понимание тематической структуры данных, «о чём эта коллекция»
- 2 Пользователь не обязан знать
  - форматы исходных данных и способы их предобработки
  - теорию TM и ARTM, виды регуляризаторов
  - методики подбора гиперпараметров
  - критерии качества моделей
  - библиотеку BigARTM
- 3 Интуитивная визуальная среда, веб-интерфейс
- 4 Пользователю должны быть доступны настройки
- 5 Дефолтные настройки должны работать на любых данных

## Приложения и исследования, взятые для анализа требований

- 1 Поиск этно-релевантных тем в социальных медиа
- 2 Анализ программ развития российских вузов
- 3 Проекты Школы Прикладного Анализа Данных
- 4 Тематический поиск по длинному текстовому запросу
- 5 Составление тематических подборок
- 6 Поиск и рубрикация научных статей на 100 языках
- 7 Выявление трендов в коллекции научных публикаций
- 8 Тематизация научно-просветительского онлайн-журнала
- 9 Поиск похожих дел в актах арбитражных судов
- 10 Тематизация пресс-релизов внешнеполитических ведомств
- 11 Тематизация twitter о российско-украинских отношениях
- 12 Выявление событийных тем в новостных потоках



## Функциональные требования (по приоритетности)

- 1 Визуализация множества всех тем и их характеристик
- 2 Визуализация каждой темы с её «рассказом о себе»
- 3 Возможность задавать словари затравок для (групп) тем
- 4 Определение динамики тем во времени
- 5 Выявление коротких тем-событий и долгих тем-трендов
- 6 Разбиение тем на подтемы иерархически
- 7 Возможность группировки тем вручную
- 8 Выявление связей тем по сочетаемости в документах
- 9 Возможность отбора и накопления «банка тем»
- 10 Тематическая фильтрация коллекции
- 11 Тематический поиск по документу или фрагменту
- 12 Рекомендательный поиск и построение подборок

## Требования к интерпретируемости (по приоритетности)

- 1 Доля интерпретируемых тем близка к 100%
- 2 Темы строятся более на терминах, чем на словах
- 3 Общая лексика выводится в отдельные фоновые темы
- 4 Нет мусорных тем, нет тем-дубликатов (декорреляция)
- 5 Решена проблема несбалансированности тем
- 6 Темы способны рассказать о себе словами и фразами
- 7 Нетекстовые термы способны рассказать о себе словами
- 8 Темы именуется автоматически
- 9 В иерархии имена дочерних тем уточняют родительские
- 10 Тематика слов согласуется с их локальными контекстами
- 11 Короткие тексты объяснимо наследуют тематику их слов
- 12 Длинные тексты разбиваются на тематические сегменты

## Основной пользовательский сценарий (без детализации)

### 1 Загрузка

- данные в различных «сырых» форматах
- возможна дозагрузка данных порциями

### 2 Предобработка

- автоматический выбор обработчиков на основании данных
- выделение модальностей: языков, времени, терминов и т.д.

### 3 Моделирование

- визуализация метрик качества в процессе обучения модели
- возможность перехода к анализу, не прерывая обучения

### 4 Визуализация

- каждая тема должна уметь «рассказать о себе»
- много разных графиков (distant reading)

### 5 Коррекция

- перебор моделей и накопление «банка тем»
- пользовательские темы как подборки с рекомендациями

## Требования к функциям Загрузки

- 1 Загрузка коллекций из различных сырых форматов
- 2 — txt, json, docx, odt, pdf и др.
- 3 — СМИ, соцмедиа, Википедия, статьи, патенты и др.
- 4 Представление метаданных и модальностей
- 5 Возможность загрузки как локально, так и из облака
- 6 Возможность дозагрузки данных из источника порциями
- 7 Текст как последовательность или как «мешок слов»
- 8 В одном файле один документ или много документов

## Требования к функциям Предобработки

- 1 Автоматическая токенизация и лемматизация
- 2 Автоматическое исправление опечаток (соцсети)
- 3 Автоматическое выделение терминов  $n$ -грамм
- 4 Метаданные: авторы, время, категории, заголовки и др.
- 5 Модальности: онимы, теги, ссылки, пользователи и др.
- 6 Настройка шаблонов для выделения модальностей
- 7 Сортировка по времени и нарезка по пакетам
- 8 Автоматическое определение коротких текстов
- 9 Автоматическая редукция словарей (по необходимости)
- 10 Автоматическое определение языков
- 11 Машинный перевод для получения параллельных текстов
- 12 Предобработка не должна идти дальше тематизации

## Требования к функциям Моделирования

- 1 Визуализация процесса обучения модели
- 2 Вывод метрик на графиках от #итерации, #пакета
- 3 Метрики перплексии, разреженности, вырожденности и др.
- 4 Автоматическая подстройка под короткие тексты
- 5 Автоматическая подстройка под длинные тексты
- 6 Темпоральная модель, если есть модальность времени
- 7 Подбор числа тем или построение иерархии тем
- 8 Автоматический подбор гиперпараметров, AutoML
- 9 Логирование информации о найденных аномалиях
- 10 Логирование данных о моделях, журнал экспериментов
- 11 Возможность перехода к анализу, не прерывая обучения
- 12 Возможность замены BigARTM на альтернативы

## Требования к функциям Визуализации

- 1 Визуальная навигация по темам, документам, терминам
- 2 XY-график тем в осях свойств тем
- 3 XY-график документов/объектов в осях объёмов тем/групп
- 4 Построение спектра тем по семантической близости
- 5 XY-график документов в осях «время–спектр тем»
- 6 Визуализация связей между словами и понятиями темы
- 7 Визуализация динамики тем в осях «время–объём темы»
- 8 Визуализация иерархии тем
- 9 Визуализация связей тем по их сочетаемости в документах
- 10 Визуализация тематической структуры документа
- 11 Выбор характеристик тем для осей XY-графиков
- 12 Выбор характеристик объектов и документов для осей

## Требования к функциям Коррекции

- 1 Разметка тем на релевантные, нерелевантные, мусорные
- 2 Разметка релевантных термов, документов в темах
- 3 Термы-затравки для «классификации иголок в стоге сена»
- 4 Обнаружение и расщепление неоднородных тем
- 5 Автоматический переход к тематической иерархии
- 6 Детекция новых событийных тем в темпоральных моделях
- 7 Накопление «банка тем» по множеству моделей
- 8 Многокритериальное оценивание качества моделей
- 9 Планирование экспериментов по улучшению моделей
- 10 Тематическая фильтрация коллекции и потока
- 11 Создание пользовательских тем — подборок документов
- 12 Ранжирование рекомендаций для пользовательских тем



## Требования к рабочему пространству проекта пользователя

- 1 Настройки входных данных — коллекций и потоков
- 2 Настройки модулей предобработки
- 3 Структура и гиперпараметры сравниваемых моделей
- 4 Структура и гиперпараметры финальной модели
- 5 Визуализации процесса обучения модели
- 6 Визуализации количественных результатов моделирования
- 7 Визуализации качественных результатов (аннотации тем)
- 8 Банк тем — множество тем, отобранных из моделей
- 9 Пользовательские темы — подборки документов
- 10 Настройка подробности отчёта по проекту
- 11 Настройка комментариев к пунктам отчёта по проекту
- 12 Сгенерированный отчёт по проекту

## Жадная минимизация: что точно войдёт в MVP

Интеграция с системами Orange, PolyAnalyst и др.

### 1 Загрузка

- несколько коллекций для тестирования
- «мешок слов» в формате Vowpal Wabbit (BigARTM)
- модальности: языки, время,  $n$ -граммы и т.д.

### 2 Моделирование

- отображение статуса обработки пакетов и текущих метрик
- возможность прервать обучение и перейти к анализу
- несколько полезных регуляризаторов — встроены

### 3 Визуализация

- навигация по темам в духе TMVE
- спектр тем по семантической близости и релевантности

### 4 Коррекция

- разметка тем на релевантные, нерелевантные, мусорные
- перестроение модели с сохранением релевантных тем

## Организация командной работы в проекте «Тематизатор»

- Доработка модуля BigARTM для Orange
- Начало разработки модуля BigARTM для PolyAnalyst
- Собеседование, определение роли в команде
- Формирование рабочих групп по ролям, например: макетирование, моделирование, разработка, тестирование
- Консультации с сотрудниками лаборатории Машинного обучения и семантического анализа Института ИИ МГУ
- Использование GitLab для командной работы.
- Документирование каждого этапа разработки.
- Регулярное планирование и проверка выполнения этапа.

---

*К.Воронцов.* Вероятностное тематическое моделирование: теория регуляризации ARTM и библиотека с открытым кодом BigARTM. 2023. (для изд-ва URSS)  
<http://www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf>

## Два упражнения на принцип максимума правдоподобия:

- Униграммная модель документов:  $p(w|d) = \xi_{dw}$   
Найти параметры модели  $\xi_{dw}$ .
- Униграммная модель коллекции:  $p(w|d) = \xi_w$  для всех  $d$   
Найти параметры модели  $\xi_w$ .

Подсказка: применить условия ККТ.

## Третье упражнение в продолжение — более творческое:

- Предложите модель, определяющую роли слов в текстах:
  - тематические слова
  - специфичные слова документа (шум)
  - слова общей лексики (фон)

Подсказка 1: искать распределение ролей слов  $p(r|w)$ ,  $r \in \{т, ш, ф\}$ .

Подсказка 2: можно разреживать  $p(r|w)$  для жёсткого определения ролей.

Подсказка 3: можно использовать документную частоту слов.

1. Заменяем  $\log$  другой монотонно возрастающей функцией  $\mu$ :

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \mu \left( \sum_{t \in T} \phi_{wt} \theta_{td} \right) + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

Как изменится EM-алгоритм? Возможно ли подобрать функцию  $\mu$  так, чтобы сократился объём вычислений?

2. Заменяем  $\log$  монотонно возрастающей функцией  $\mu$  в регуляризаторе сглаживания–разреживания (модель LDA):

$$R(\Phi, \Theta) = \sum_{t \in T} \sum_{w \in W} \beta_w \mu(\phi_{wt}) + \sum_{d \in D} \sum_{t \in T} \alpha_t \mu(\theta_{td}).$$

Как изменится M-шаг и воздействие регуляризатора на модель?

3\*. Какому регуляризатору соответствует формула M-шага

$$\phi_{wt} = \underset{w}{\text{norm}} \left( n_{wt} [n_{wt} > \gamma n_t] \right)$$

**Подсказка:** см. слайд 12 следующей лекции.