

Вероятностные тематические модели

Лекция 10. Примеры прикладных задач и проект «Тематизатор»

К. В. Воронцов
`k.vorontsov@iai.msu.ru`

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

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

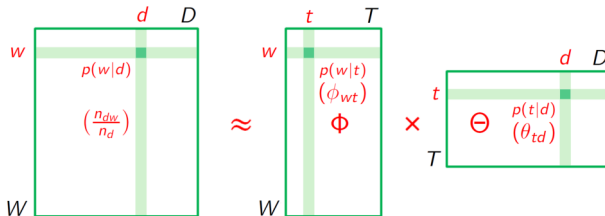
Дано: коллекция текстовых документов

- 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

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

ЕМ-алгоритм: метод простой итерации для системы уравнений

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

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

Мультимодальная тематическая модель

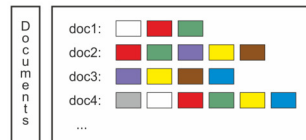
Тема может порождать термины различных *модальностей*:
 $p(\text{слово}|t)$, $p(n\text{-грамма}|t)$,

Text documents



Topic
Modeling

Topics of documents



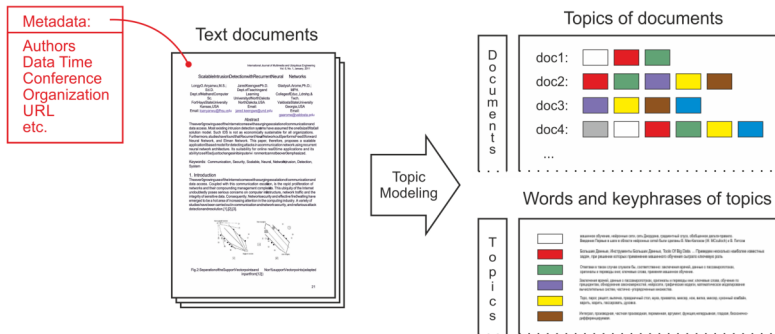
Words and keyphrases of topics



Мультимодальная тематическая модель

Тема может порождать термины различных *модальностей*:

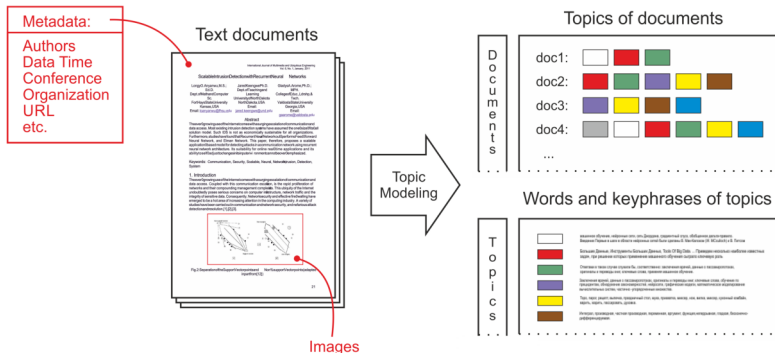
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,



Мультимодальная тематическая модель

Тема может порождать термины различных *модальностей*:

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



Тема может порождать термы различных *модальностей*:

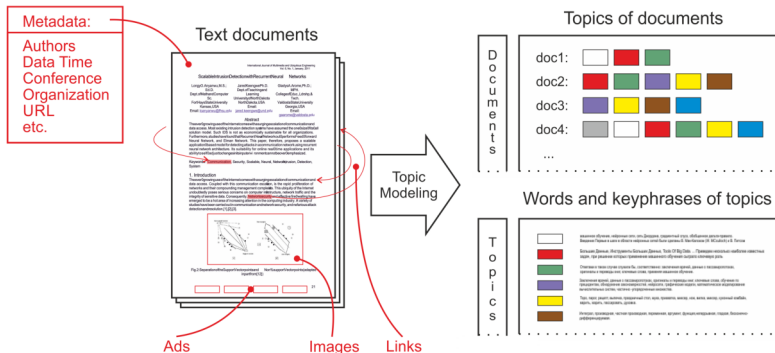
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,
 $p(\text{объект}|t)$, $p(\text{ссылка}|t)$,



Мультимодальная тематическая модель

Тема может порождать термины различных *модальностей*:

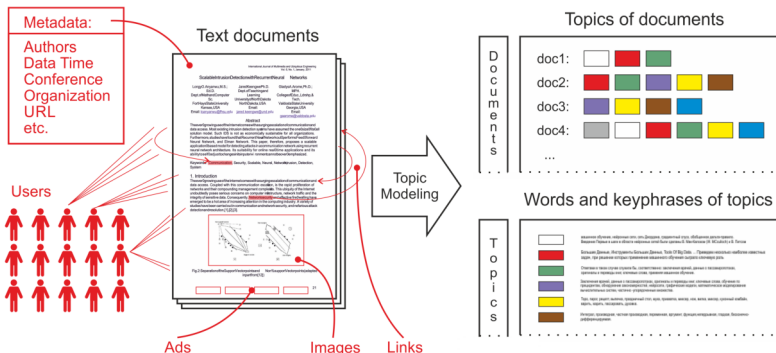
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,
 $p(\text{объект}|t)$, $p(\text{ссылка}|t)$, $p(\text{баннер}|t)$,



Мультимодальная тематическая модель

Тема может порождать термины различных *модальностей*:

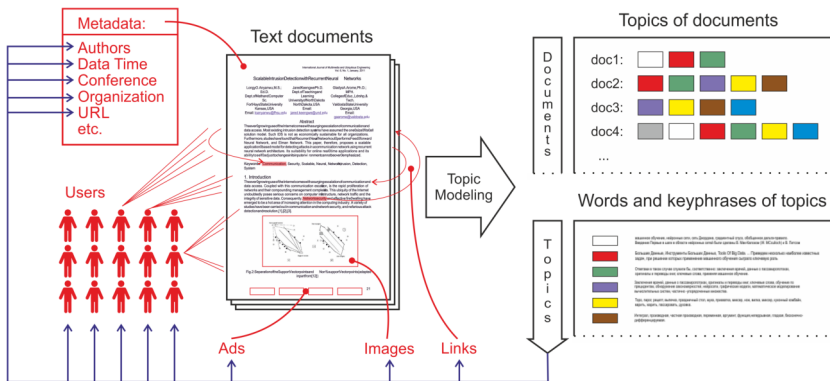
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,
 $p(\text{объект}|t)$, $p(\text{ссылка}|t)$, $p(\text{баннер}|t)$, $p(\text{пользователь}|t)$



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:

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

ЕМ-алгоритм: метод простой итерации для системы уравнений

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

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
- Онлайновый 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>



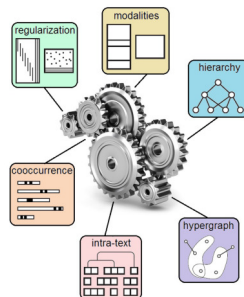
Лицензия и среда разработки:

- 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



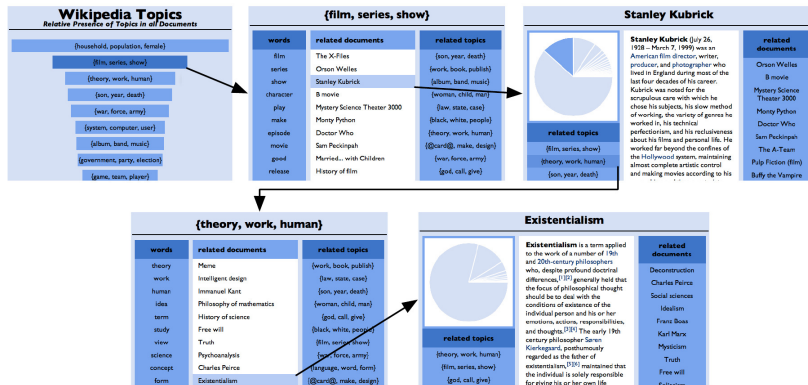
TopicNet

- 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)

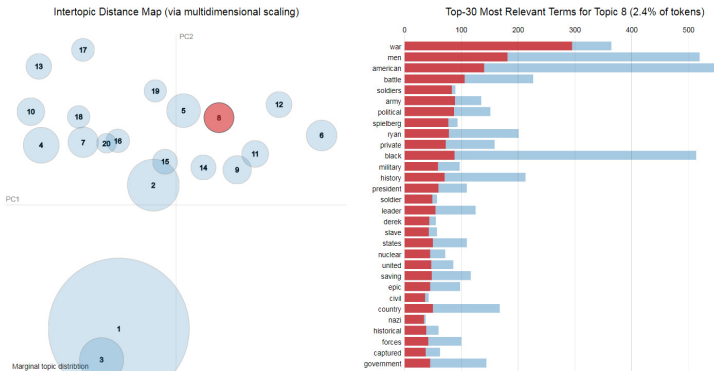
Тематический навигатор с веб-интерфейсом:



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

Система LDAvis

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

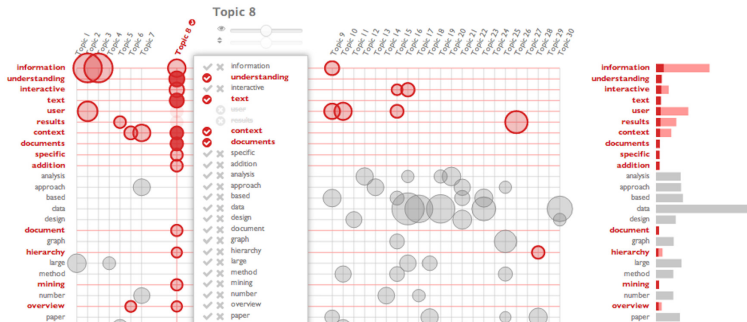


<https://github.com/cpsievert/LDAvis>

C.Sievert, K.Shirley. LDAvis: A method for visualizing and interpreting topics. 2014.

Система Termite

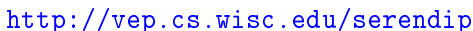
Интерактивная визуализация матрицы Φ и сравнение тем:



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

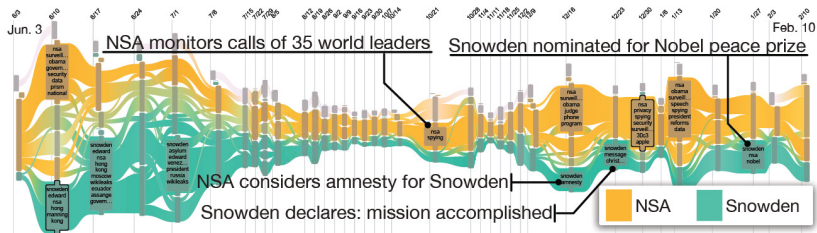
Chuang J., Manning C., Heer J. Termite: Visualization Techniques for Assessing Textual Topic Models. IWCAVI 2012.

Визуализация матриц Φ , Θ и тематики слов в текстах:



К. В. Воронцов (k.v.vorontsov@phystech.edu)

Динамика тем: эволюция предметной области



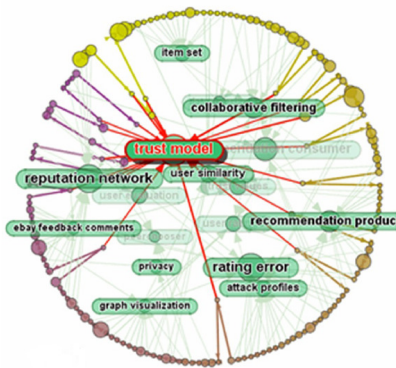
Эволюция выбранных тем иерархии. Данные Prism (2013/06/03–2014/02/09)

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.

© 2006 The Authors
Journal compilation © 2006 Blackwell Publishing Ltd



К. В. Воронцов (k.v.vorontsov@phystech.edu) Вероятностные тематические модели

Пример иерархической карты области *Data Mining*



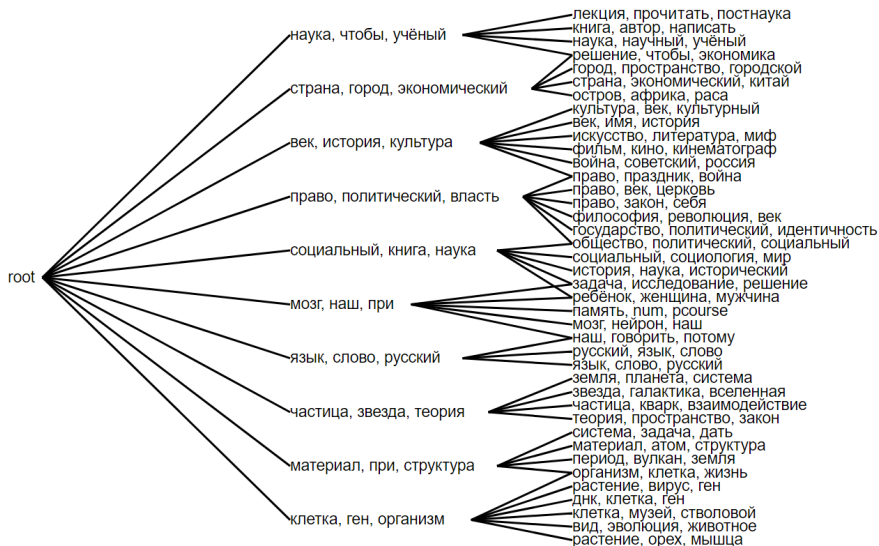
FoamTree: <https://carrotsearch.com/foamtree>

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: $\rho(t, t') = 1 - \frac{|W_t \cap W_{t'}|}{|W_t \cup W_{t'}|}, \quad W_t = \{w: \phi_{wt} > \frac{1}{|W|}\}$

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>

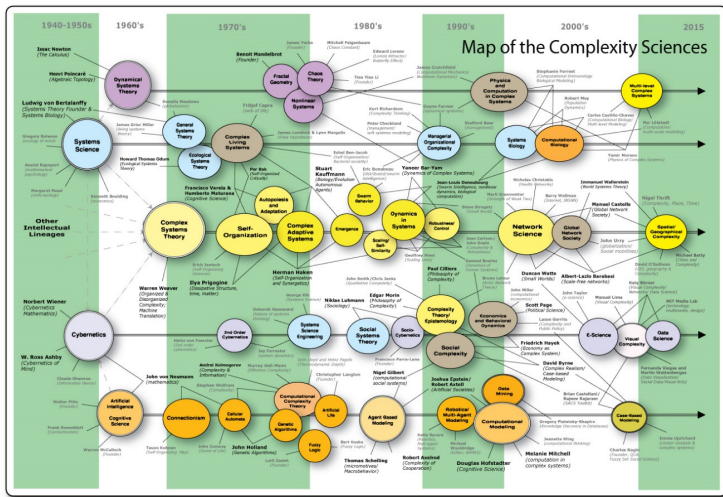
Интерактивный обзор 440 средств визуализации текстов



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, алеппо, 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, сообщать, 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, деньги, 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	38	42	30	104

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



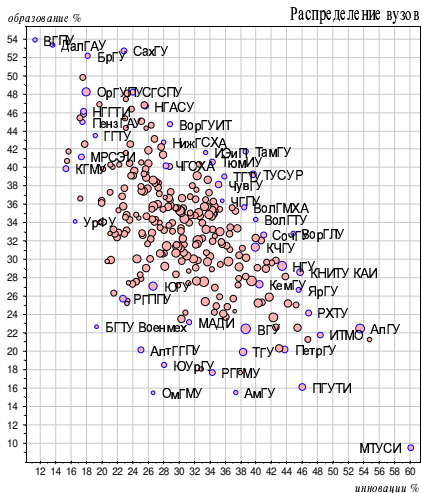
Тематическая карта вузов

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] Двухязычный 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 2MB;
- 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.

ТМ в политологии: анализ публичных выступлений

- [1] Выступления (210K) в Европарламенте, 1999–2014:
 - выявление событийных тем и эволюции перманентных тем;
 - как члены и комитеты ЕП влияют на формирование тем
- [2] Модель контрастных мнений (Contrastive Opinion Modeling)
 - выступления в Сенате США (www.votesmart.org);
 - СМИ: New York Times, Xinhua News, The Hindu, 2009–2010
- [3] Выступления в Совбезе ООН по Афганистану, 2001–2017:
 - динамика отношения разных стран к проблемам Афганистана

[1] *D. Greene, J.P. Cross*. Unveiling the political agenda of the European Parliament plenary: a topical analysis. 2015.

[2] *Fang, Y., et al*. Mining contrastive opinions on political texts using cross-perspective topic model. 2012.

[3] *M. Schönfeld*. Discursive landscapes and unsupervised topic modeling in IR: a validation of text-as-data approaches through a new corpus of UN Security Council speeches on Afghanistan. 2018.

ТМ в политологии: анализ СМИ и социальных медиа

- [1] Тематика изменения климата в СМИ Пакистана, 2010–2021
— выявление, группирование и динамика тем
- [2] Выявление поляризации новостей (AYLIEN COVID-19)
— 1,5М новостей, 440 источников СМИ, 11.2019–07.2020
- [3] Выявление политических взглядов пользователей Twitter
- [4] Что пишет NYT о ядерных технологиях с 1945 по н/в

-
- [1] *W.Ejaz et al.* Politics triumphs: A topic modeling approach for analyzing news media coverage of climate change in Pakistan. 2023
 - [2] *Zihao He.* Detecting polarized topics using partisanship-aware contextualized topic embeddings. 2021
 - [3] *R.Cohen, D.Ruths.* Classifying Political Orientation on Twitter: It's Not Easy! 2013.
 - [4] *C.Jacobi.* Quantitative analysis of large amounts of journalistic texts using topic modelling. 2015.
- H.Jelodar et al.* Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. 2019.

Проект «Тематизатор»: общие требования

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

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

Приложения и исследования, взятые для анализа требований

- ❶ Поиск этно-релевантных тем в социальных медиа
- ❷ Анализ программ развития российских вузов
- ❸ Проекты Школы Прикладного Анализа Данных
- ❹ Тематический поиск по длинному текстовому запросу
- ❺ Составление тематических подборок
- ❻ Поиск и рубрикация научных статей на 100 языках
- ❼ Выявление трендов в коллекции научных публикаций
- ❽ Тематизация научно-просветительского онлайн-журнала
- ❾ Поиск похожих дел в актах арбитражных судов
- ❿ Тематизация пресс-релизов внешнеполитических ведомств
- ⓫ Тематизация twitter о российско-украинских отношениях
- ⓫ 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 именуются 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

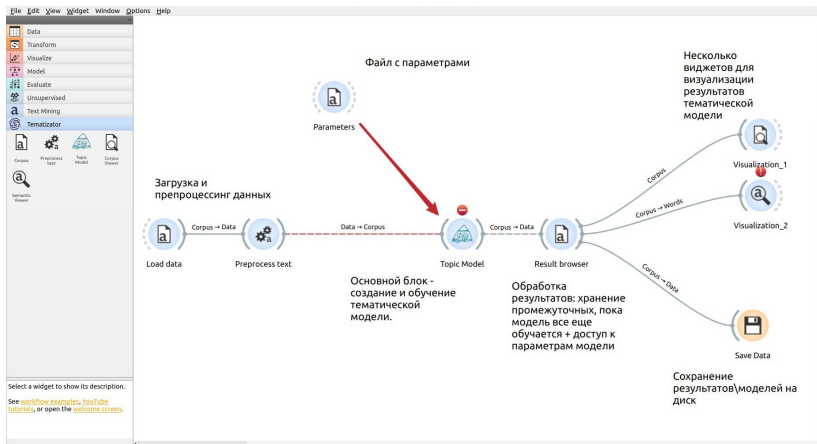
4 Визуализация

- 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

Встраивание модуля BigARTM в Orange



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

- ❶ Автоматическая токенизация и лемматизация
- ❷ Автоматическое исправление опечаток (соцсети)
- ❸ Автоматическое выделение терминов n -грамм
- ❹ Метаданные: авторы, время, категории, заголовки и др.
- ❺ Модальности: онимы, теги, ссылки, пользователи и др.
- ❻ Настройка шаблонов для выделения модальностей
- ❼ Сортировка по времени и нарезка по пакетам
- ❽ Автоматическое определение коротких текстов
- ❾ Автоматическая редукция словарей (по необходимости)
- ❿ Автоматическое определение языков
- ⓫ Машинный перевод для получения параллельных текстов
- ⓫ Предобработка не должна идти дольше 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

- ❶ Визуальная навигация по темам, документам, терминам
- ❷ XY-график тем в осях свойств тем
- ❸ XY-график документов/объектов в осях объёмов тем/групп
- ❹ Построение спектра тем по семантической близости
- ❺ XY-график документов в осях «время–спектр тем»
- ❻ Визуализация связей между словами и понятиями темы
- ❼ Визуализация динамики тем в осях «время–объём темы»
- ❽ Визуализация иерархии тем
- ❾ Визуализация связей тем по их сочетаемости в документах
- ❿ Визуализация тематической структуры документа
- ⓫ Выбор характеристик тем для осей XY-графиков
- ⓫ Выбор характеристик объектов и документов для осей

Требования к функциям Коррекции

- ❶ Разметка тем на релевантные, нерелевантные, мусорные
- ❷ Разметка релевантных термов, документов в темах
- ❸ Термы-затравки для «классификации иголок в стоге сена»
- ❹ Обнаружение и расщепление неоднородных тем
- ❺ Автоматический переход к тематической иерархии
- ❻ Детекция новых событийных тем в темпоральных моделях
- ❼ Накопление «банка тем» по множеству моделей
- ❽ Многокритериальное оценивание качества моделей
- ❾ Планирование экспериментов по улучшению моделей
- ❿ Тематическая фильтрация коллекции и потока
- ⓫ Создание пользовательских тем — подборок документов
- ⓬ 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>