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)»

ВМК МГУ • 21 февраля 2023

Содержание

Постановка задачи тематического моделирования

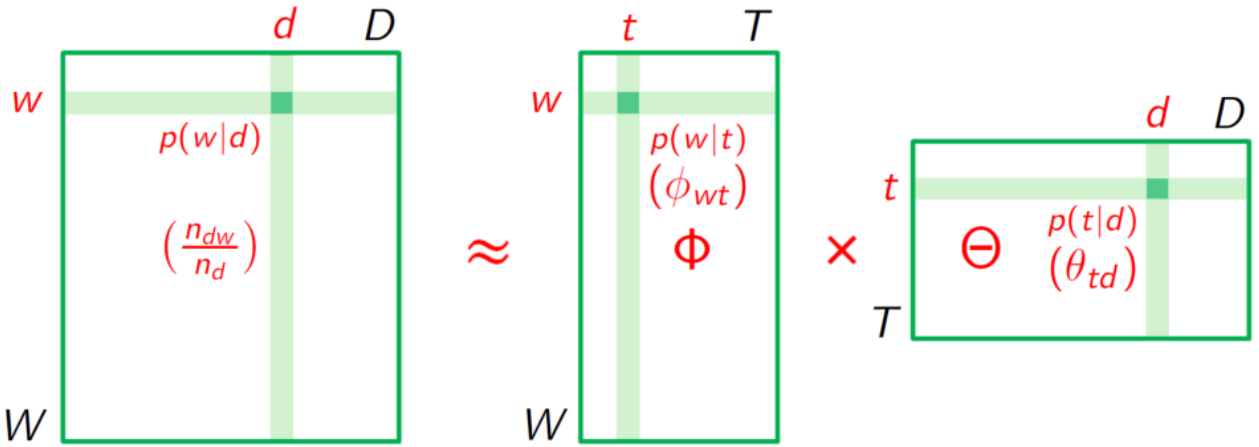
Дано: коллекция текстовых документов

- 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

Максимизация логарифма правдоподобия с регуляризатором:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

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

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

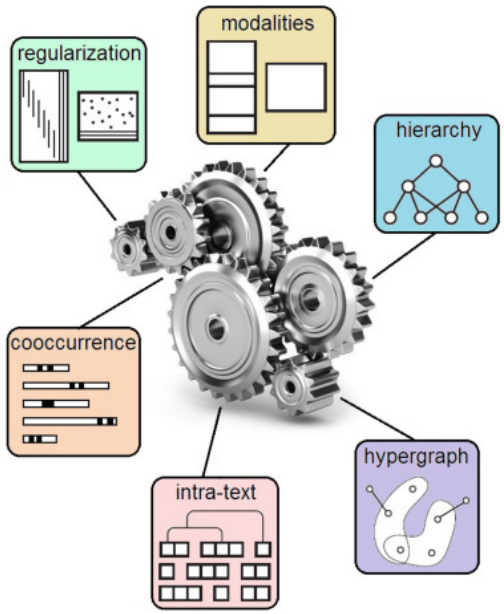
Лицензия и среда разработки:

- 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

VisARTM: визуализация для BigARTM

- 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 асессорских оценок тем

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

VisARTM: Визуализация документа

- метаданные документа
- распределение $p(t|d)$
- главная тема для каждого слова $\max_t p(t|d,w)$

Химические коммуникации планктона

Эколог Егор Задереев о типах химических сигналов, миграциях зоопланктона и образовании покоящихся яиц

Text | Bag of words

Что исследователи знают о химической коммуникации планктона в воде? Какими сигналами обменивается зоопланктон? Как размножается зоопланктон? Об этом рассказывает кандидат биологических наук Егор Задереев.

Планктон — это организмы, местоположение которых в водной толще в основном определяется течениями. То есть это что-то маленькое, то, что переносится течениями. Планктон делится на фитопланктон (это водоросли) и зоопланктон. Мы будем говорить про зоопланктон — это рачки. То, как водные объекты между собой коммуницируют с помощью химических сигналов, исследовано довольно плохо. В наземных экосистемах, мы знаем, есть феромоны, различные сигнальные системы, которые хорошо исследованы. Мы используем их для создания ловушек, например, для вредителей — феромонные ловушки. Вода — это среда, которая благоприятна для химической коммуникации.

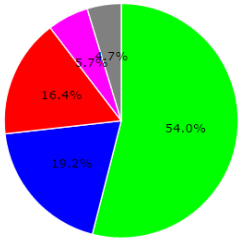
[post id="33793"]

Химические сигналы от хищников заставляют зоопланктон мигрировать. Это одно из самых масштабных на планете перемещений биомассы, которые ежесуточно происходят в океанах, морях и озерах. Зоопланктон ночью поднимается к поверхности, а днем уходит на глубину. Днем свет сверху помогает хищникам ловить животных, и животные уходят на глубину, а ночью поднимаются к поверхности, чтобы есть. Было показано, что эти вертикальные миграции регулируются двумя факторами. Первый — это освещенность. Очевидно, что, если не будет света, не будет сигнала. А второй — это химия, которую выделяют хищники.

В 2006 и 2009 годах выходили хорошие обзоры по химическим коммуникациям. То есть а) это очень маленькие молекулы, и б) они работают в очень низких концентрациях. Это до сих пор удивляет и поражает, потому что сообщества зоопланктона и вообще планктона в водных экосистемах — это сотни видов водорослей, рачков, которые живут в озерах, в морях, взаимодействуют между собой. А между ними есть очень сложная, судя по тому, что мы получаем в лаборатории, и разветвленная сеть химических сигналов и коммуникаций, которые влияют на разные поведенческие, физиологические и продуктивные функции. И эта сложная цепь, сеть взаимодействий до сих пор слабо исследована.

Dataset: postnauka
Time: Dec. 14, 2014, 3 p.m.
View original
index_id: 1866
text_id: 36719.txt
Terms count: 0
Unique terms count: 0
Model: flat-20
Highlighting: Words

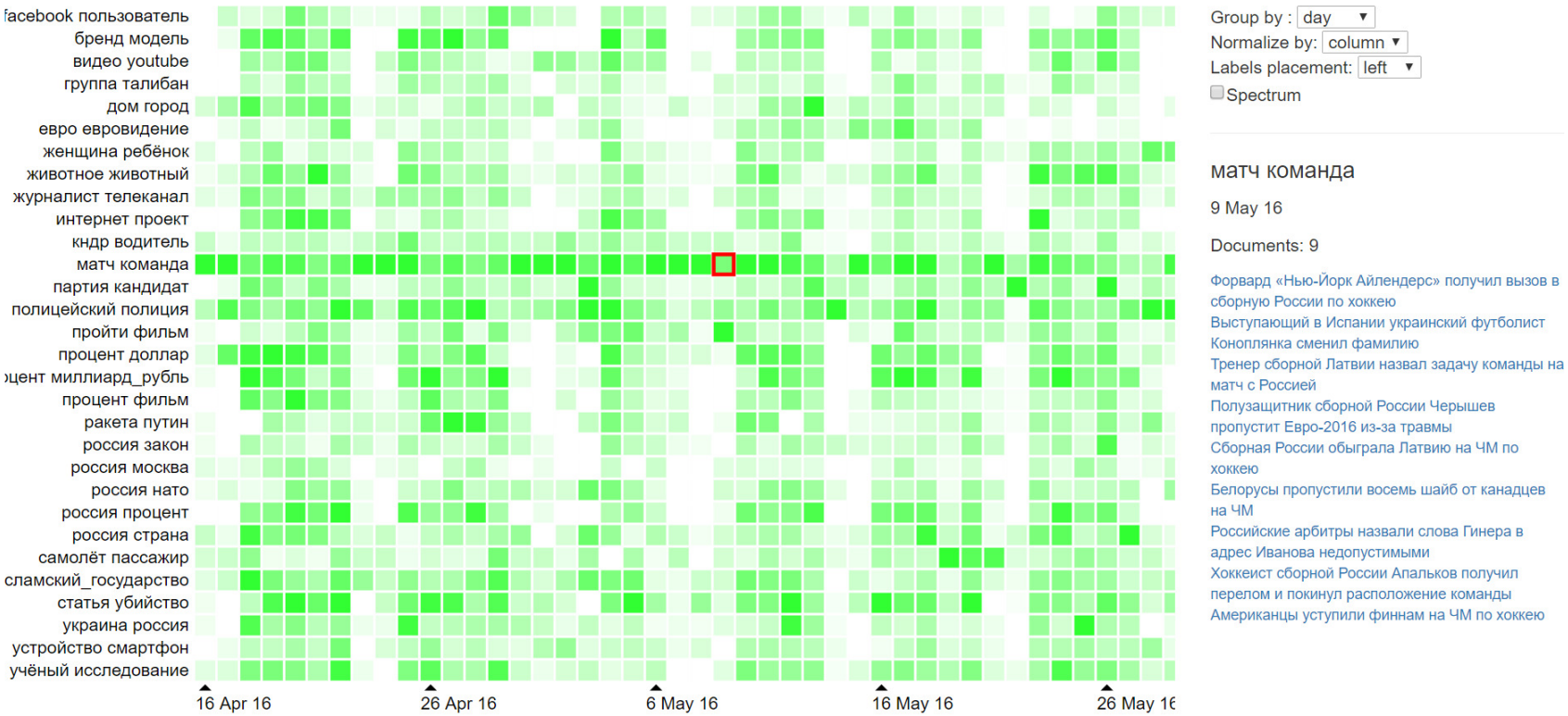
Topic distribution



- земля, микроорганизм, вид
- вид, эволюция, ген
- материал, квантовый, структура
- город, социальный, пространство
- Other

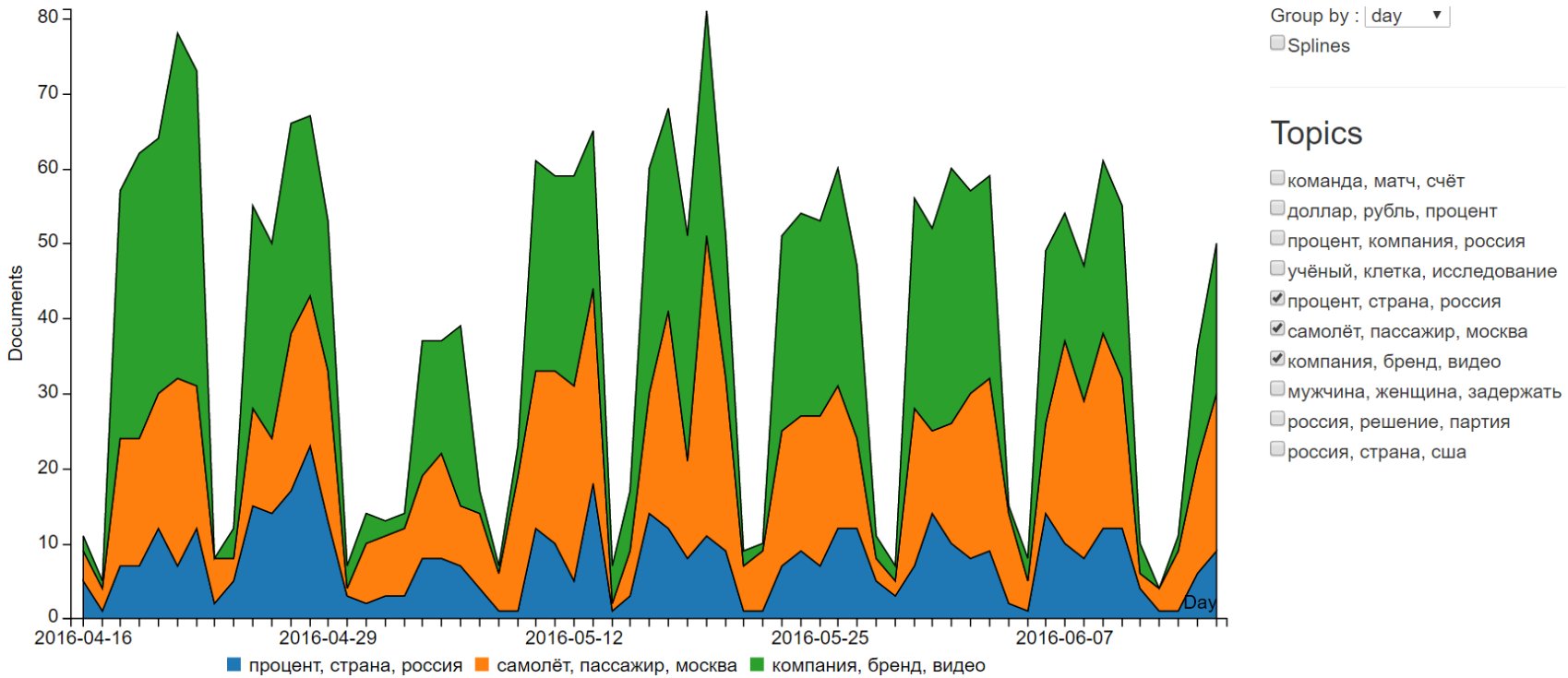
VisARTM: Визуализация темпоральной модели

- XY-график $p(t|i)$ в осях интервалы–темы
- заголовки документов для каждой точки (i, t)



VisARTM: Визуализация темпоральной модели

- выбор тем для отображения временных рядов
- временные ряды: $p(t|i)$ или документная частота $N(t,i)$



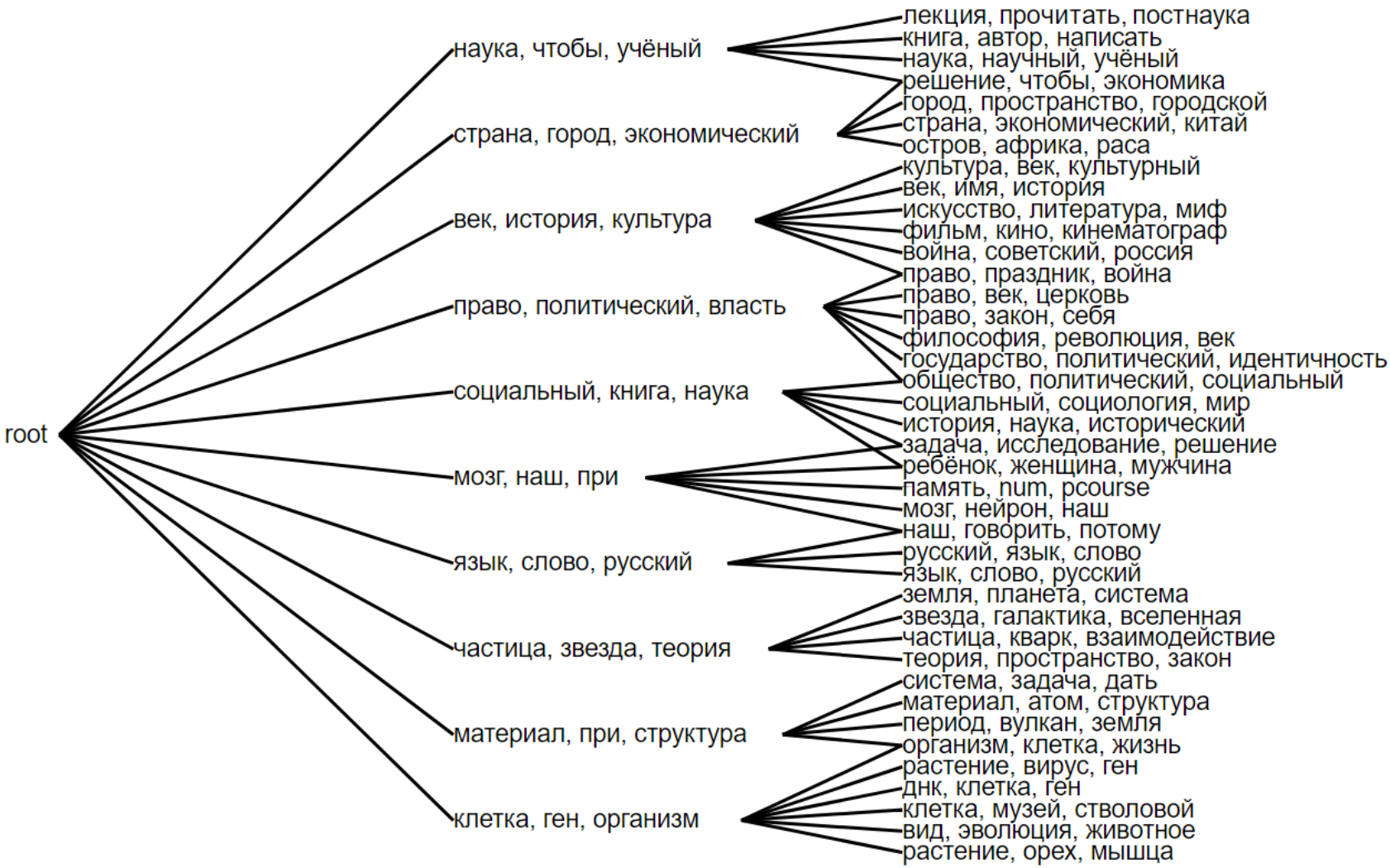
VisARTM: Визуализация иерархической модели

- Использовалась библиотека визуализации FoamTree

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

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

Иерархический спектр тем (коллекция postnauka.ru)



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

Задача 1. 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

Задача 1. 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

Задача 1. 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, алесио, 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,

Задача 1. 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, упс, 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,

Задача 1. 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, мадуро, боливия, фидель, глава, латинский, венесуэльский, лидер, боливарианской, президентский, альенде, гевару,
(китайцы): китайский, россия, производство, китай, 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,

Задача 1. 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.

Задача 1. 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.

Задача 2. 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 :(

Задача 2. 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

Задача 2. 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

Задача 2. 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): инфрастуктура, кампус, строительство

Задача 2. Интерпретация всех 50 тем

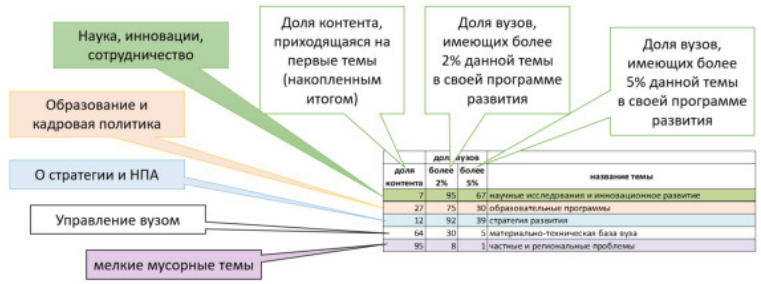
- Для интерпретируемости тем важны биграммы
- Модель построили примерно с 10-й попытки (подбирали число тем, регуляризацию, добивались различности тем)
- Интерпретация 50 тем заняла примерно 20 минут работы
- Иногда выделялись темы исследований и разработок, но для этого нужна более гранулированная модель
- 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

Задача 2. Интерпретация всех 50 тем

	доля вузов		
доля контента	более 2%	более 5%	название темы
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	инфрастуктура, 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 сх, лес, химия, ит
69	29	1	публикационная активность и защиты диссертаций
71	29	2	взаимодействие с региональной властью

	доля вузов		
доля контента	более 2%	более 5%	название темы
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



Задача 2. Тематическая карта вузов

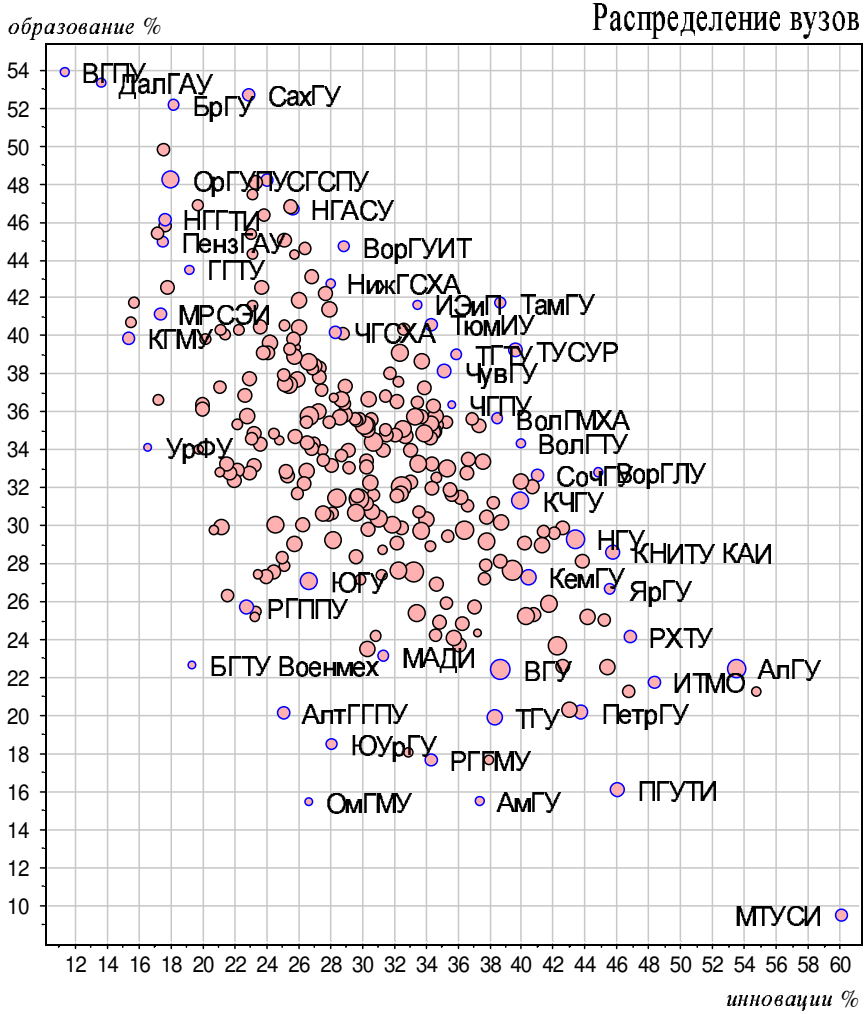
По осям:

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, 25M 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] Двуязычный 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.

Приложения и исследования, взятые для анализа требований

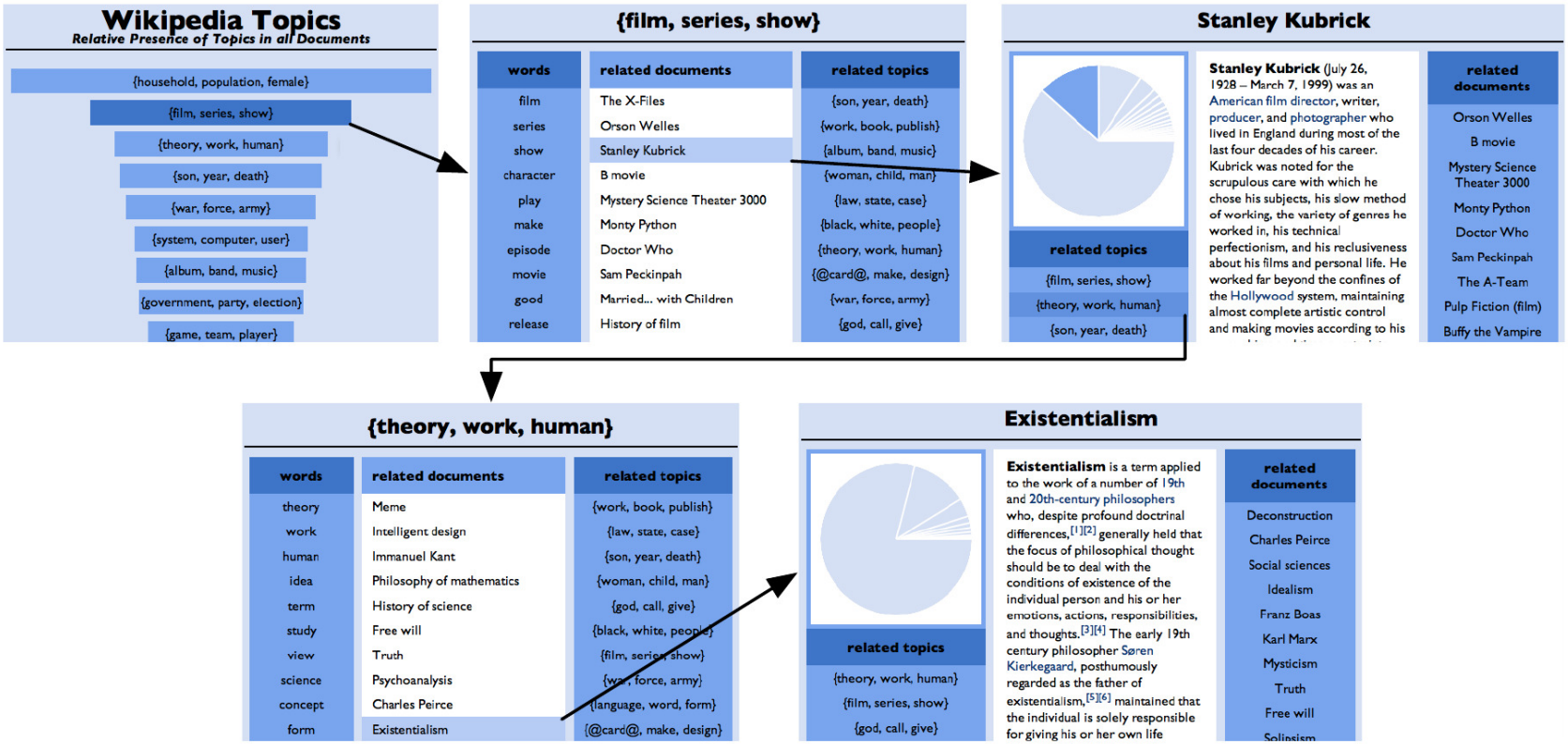
1. Поиск этно-релевантных тем в социальных медиа
2. Анализ программ развития российских вузов
3. Проекты Школы Прикладного Анализа Данных
4. Тематический поиск по длинному текстовому запросу
5. Составление тематических подборок
6. Поиск и рубрикация научных статей на 100 языках
7. Выявление трендов в 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

Система TMVE — Topic Model Visualization Engine

Тематический навигатор с веб-интерфейсом:



https://github.com/ajbc/tmv

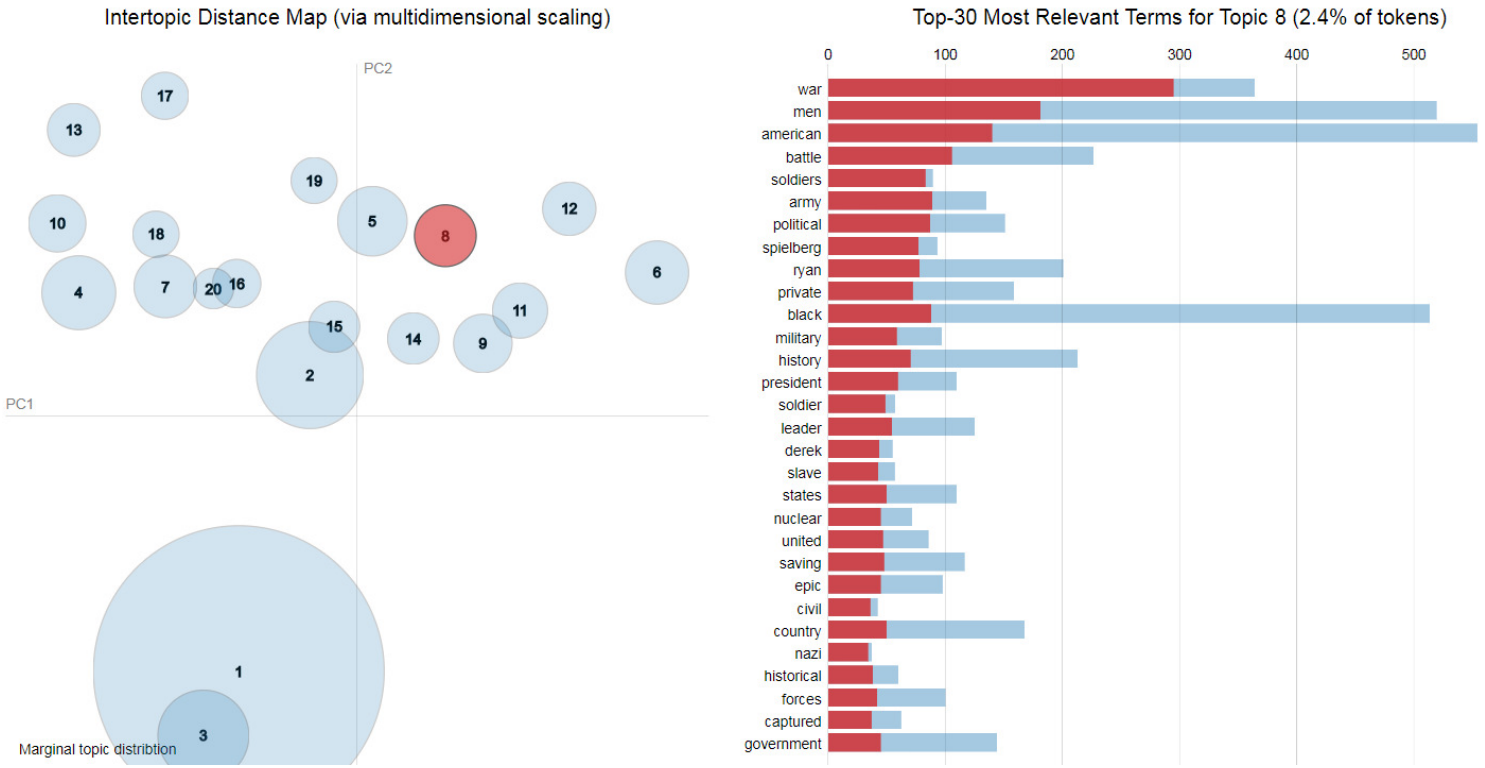
Chaney A., Blei D. Visualizing Topic Models, 2012.

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

Система LDAvis

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

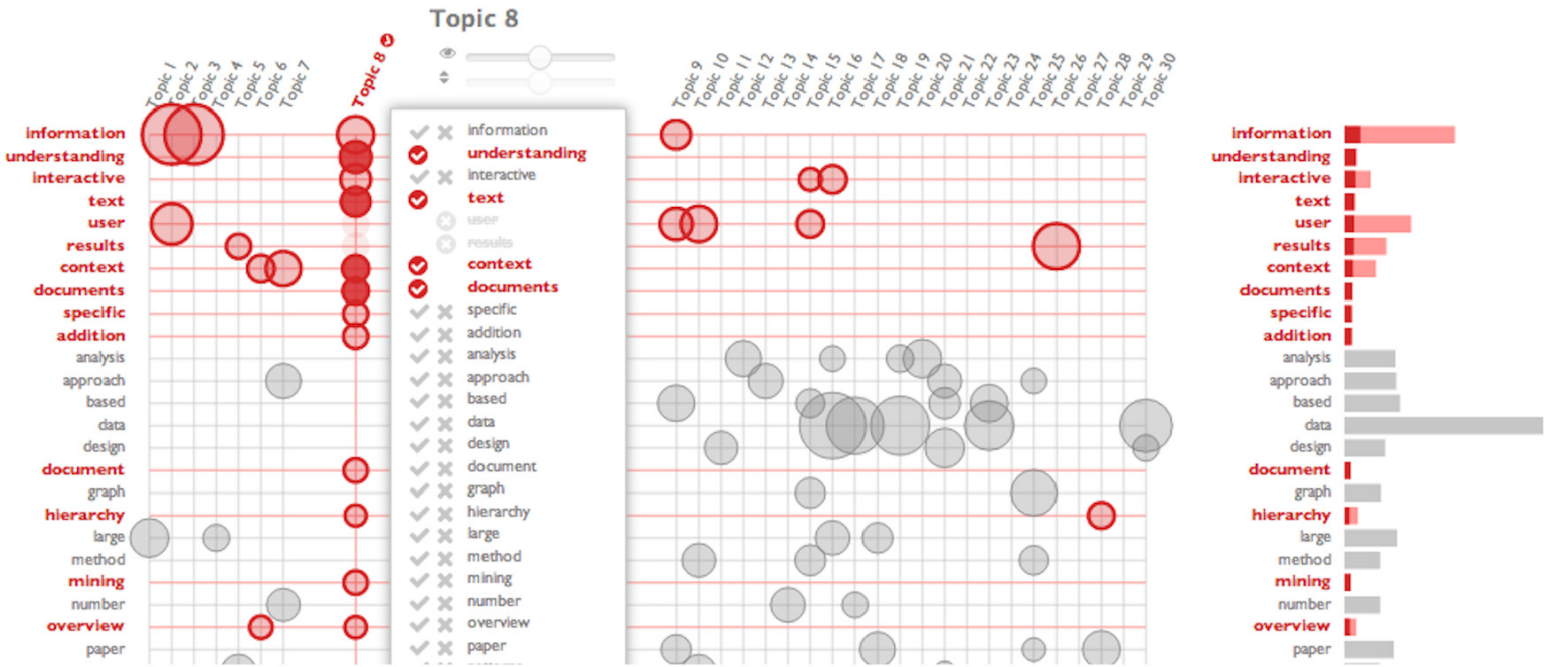


https://github.com/cpsievert/LDAvis

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.

Система Serendip

Визуализация матриц Φ, Θ 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.

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 дольше тематизации

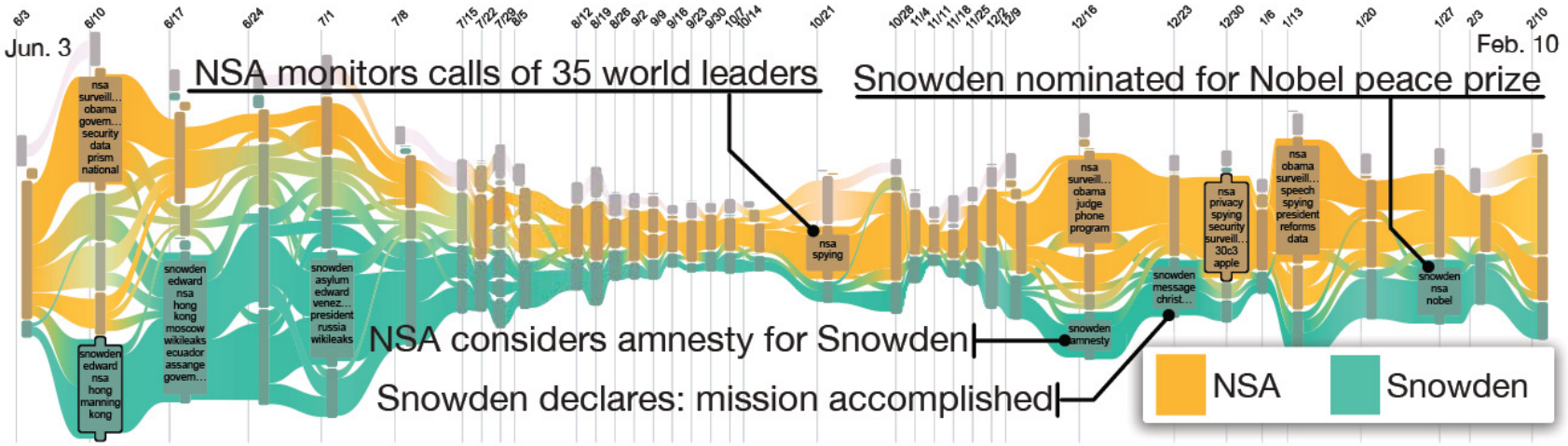
Требования к функциям Моделирования

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

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

Динамика тем: эволюция предметной области



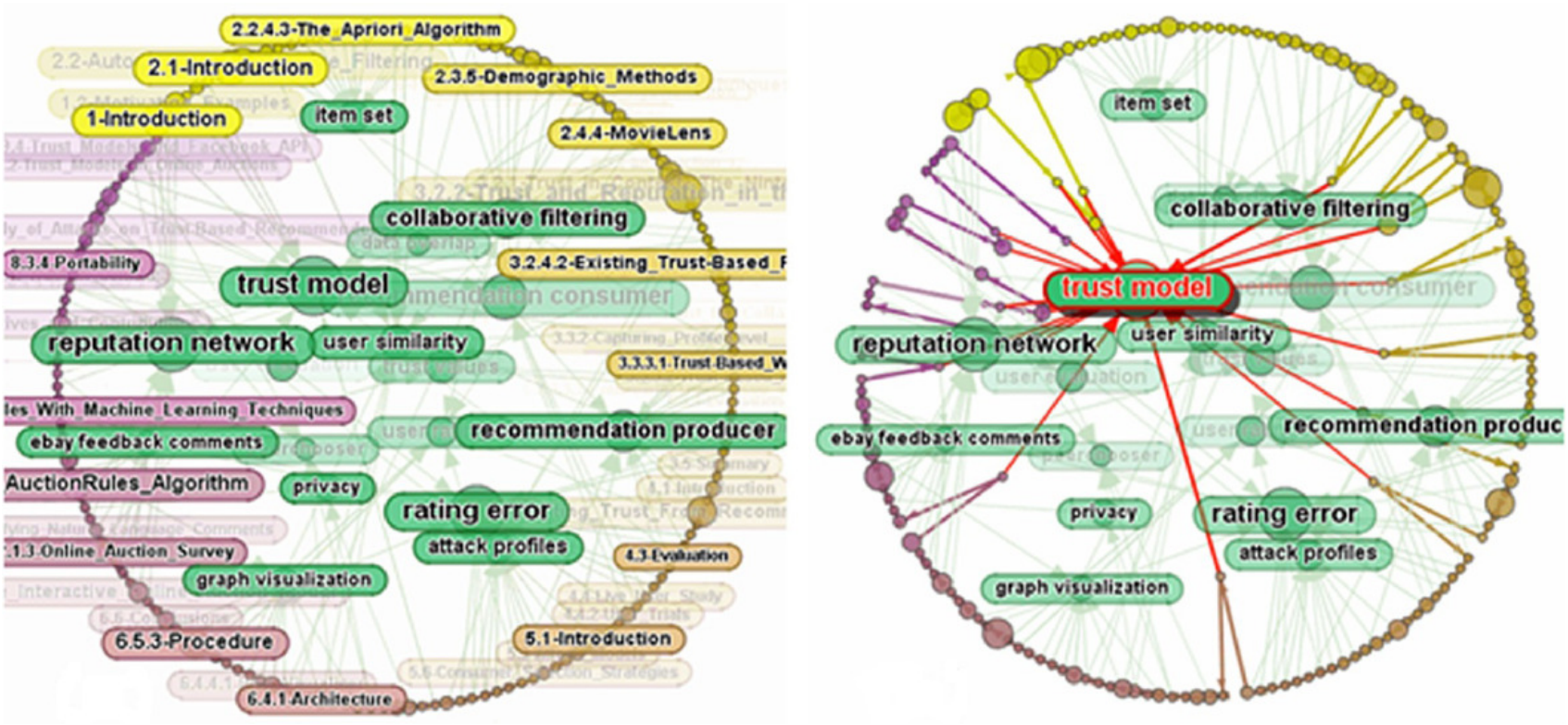
Эволюция выбранных тем иерархии. Данные Prism (2013/06/03–2014/02/09)

Стратегия визуализации в системах TextFlow и RoseRiver:

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

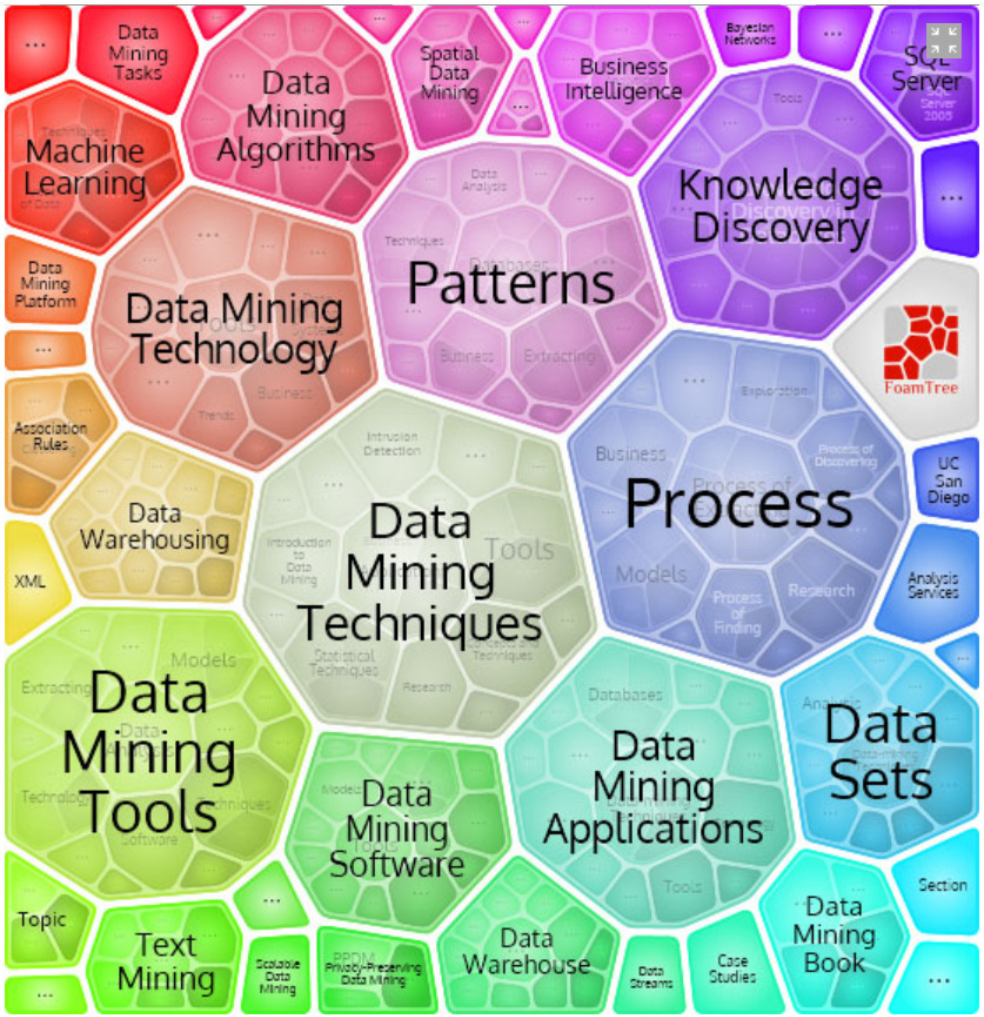
Weiwei Cui, Shixia Liu, Zhuofeng Wu, Hao Wei. How hierarchical topics evolve in large text corpora. 2014.

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

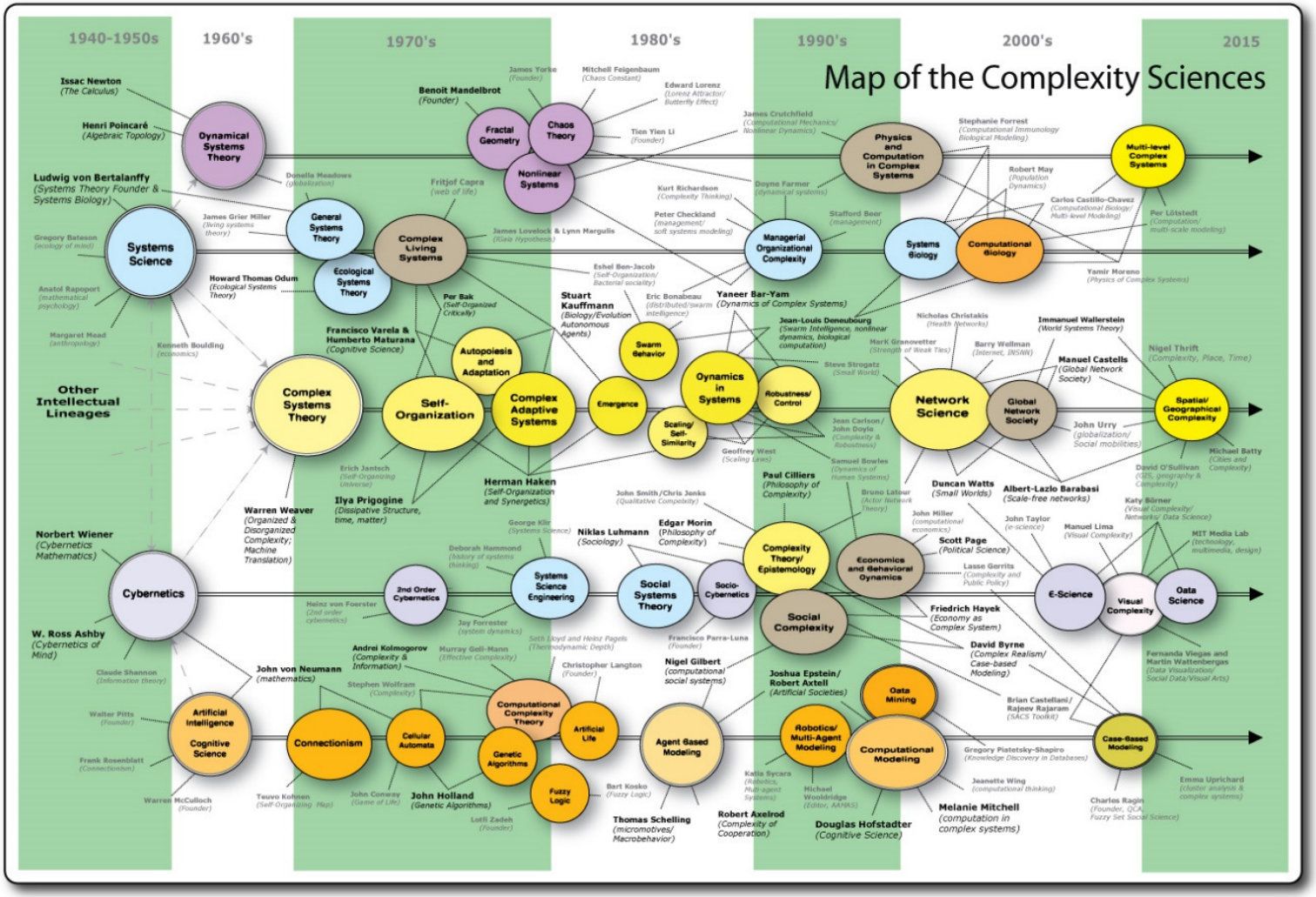
Источники вдохновения: http://textvis.lnu.se

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 $\rightarrow$ естественные $\rightarrow$ точные
- **Интерактивность:** реализация мантры Шнейдермана
- **Суммаризация:** масштаб карты определяет объём текста

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

Что точно войдёт в MVP

1. **Загрузка**
   - «мешок слов» в формате Vowpal Wabbit (BigARTM)
2. **Моделирование**
   - отображение статуса обработки пакетов и текущих метрик
   - возможность прервать обучение и перейти к анализу
   - регуляризатор декоррелирования — спрятан
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

Что точно не войдёт в MVP: открытые научные проблемы

1. Доля интерпретируемых тем близка к 100%
2. Проблема несбалансированности тем
3. Единая модель для коротких и длинных текстов
4. Корректное обнаружение новых тем в пакетах
5. Обеспечение полноты и устойчивости множества тем
6. Автоматический подбор гиперпараметров, AutoML
7. Оптимизация гиперпараметров в потоковом режиме
8. Именование и аннотирование тем
9. Бережное слияние моделей нескольких коллекций
10. Применение гиперграфовых тематических моделей
11. Создание тематических моделей внимания
12. Симбиоз тематических моделей с нейронными сетями

Этапизация работ

- Макетирование web-интерфейса
- Разработка сценариев использования и тестов
- Декомпозиция задачи (Test-Driven Development, TDD)
- Параллельно: решение тестовых референтных задач
- Подготовка инфраструктуры для разворачивания проекта
- Разработка демо-версии программы, включающей:
  - модуль загрузки данных в одном из форматов
  - базовую версию предварительной обработки
  - базовую версию обучения тематической модели
  - несколько модулей представления результатов
- Разработка и тестирование MVP
- Получение денег от венчурных инвесторов =)

Организация командной работы в проекте «Тематизатор»

- Собеседование, определение роли в команде
- Формирование рабочих групп по ролям, например: макетирование, проектирование, кодирование, тестирование
- Изучение аналогов: PolyAnalyst, Orange и др.
- 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

Теоретическое задание №2

1. В функционале log-правдоподобия логарифм заменили другой монотонно возрастающей функцией μ:

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

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

2. В регуляризаторе сглаживания–разреживания (модель LDA) заменили логарифм монотонно возрастающей функцией μ:

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

Как изменится M-шаг? Как это повоздействует на модель?

3*. Какому регуляризатору соответствует формула M-шага

$$\phi_{wt} = \operatorname*{norm}_{w}\big(n_{wt}[n_{wt} < \gamma n_t]\big)$$