Тематика НИР: математические методы анализа текстов и информационного поиска

Константин Вячеславович Воронцов

- Лаборатория машинного обучения и семантического анализа Института искусственного интеллекта МГУ •
- Кафедра математических методов прогнозирования ВМК МГУ •
- Кафедра машинного обучения и цифровой гуманитаристики, кафедра интеллектуальных систем МФТИ •
- ФИЦ ИУ РАН •

voron@mlsa-iai.ru
http://www.MachineLearning.ru/wiki?title=User:Vokov

AI Masters, ИИИ МГУ • 18 ноября 2023

Содержание

Концепция «мастерской знаний»

«Огромное и все возрастающее богатство знаний разбросано сегодня по всему миру. Этих знаний, вероятно, было бы достаточно для решения всего громадного количества трудностей наших дней, но они рассеяны и неорганизованы. Нам необходима очистка мышления в *своеобразной мастерской*, где можно получать, сортировать, суммировать, усваивать, разъяснять и сравнивать знания и идеи»
— *Герберт Уэллс, 1940*

"An immense and ever-increasing wealth of knowledge is scattered about the world today; knowledge that would probably suffice to solve all the mighty difficulties of our age, but it is dispersed and unorganized. We need a sort of mental clearing house for the mind: a depot where knowledge and ideas are received, sorted, summarized, digested, clarified and compared"
— *Herbert Wells, 1940*

Концепция сервисов «Мастерской знаний»

Подборка — долгосрочный поисковый интерес пользователя

Поисково-рекомендательные функции:

- поиск тематически близких документов по подборке
- мониторинг новых документов для подборки
- контекстные рекомендации по документу из подборки

Аналитические функции:

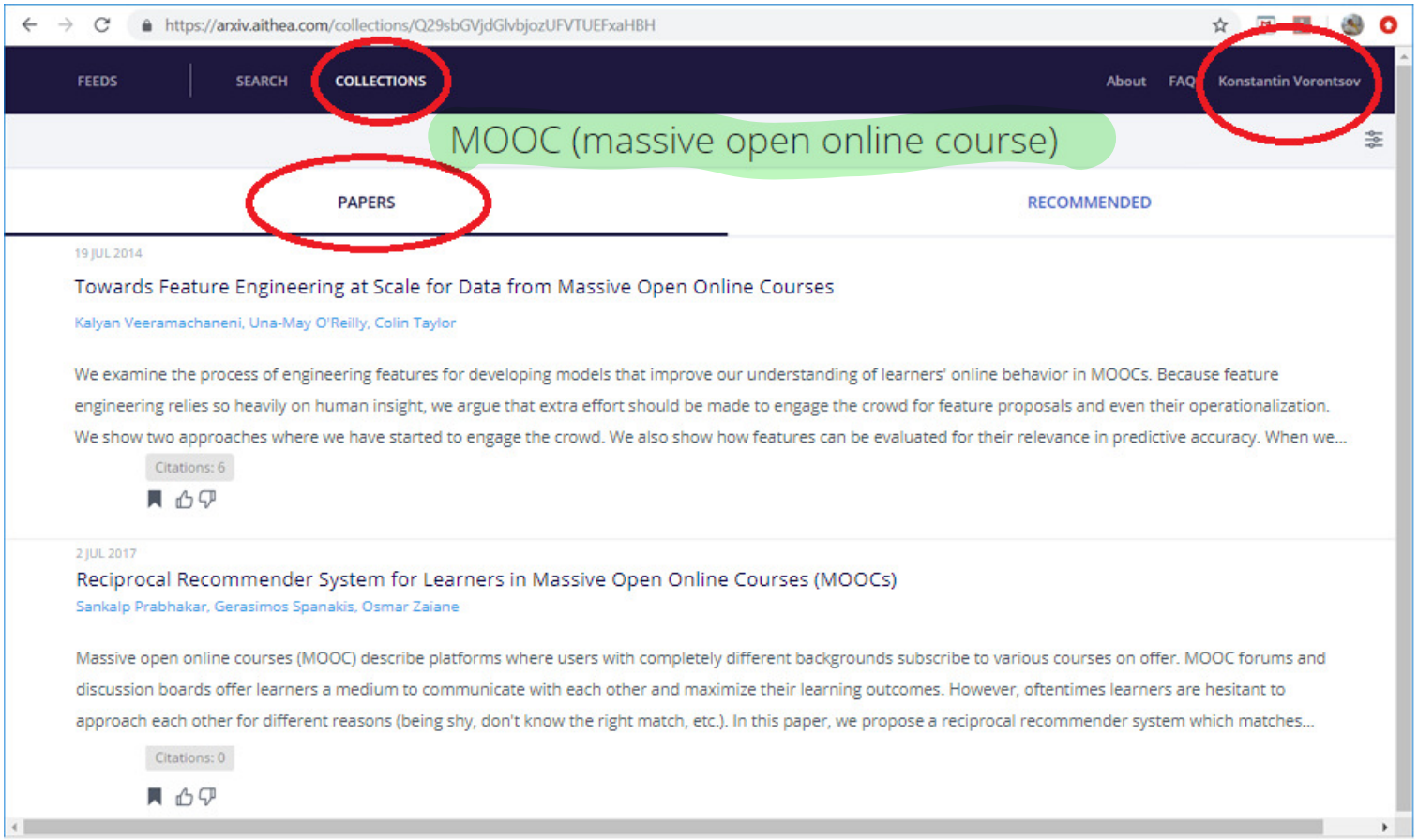
- автоматизация реферирования подборки
- кластеризация трендов, аспектов, отношений в подборке
- рекомендация порядка чтения внутри подборки
- выделение «важных мест» в документе из подборки

Коммуникативные функции:

- совместное составление и использование подборок
- интерактивная визуализация и инфографика по подборке

Поисково-рекомендательная система SciSearch.ai

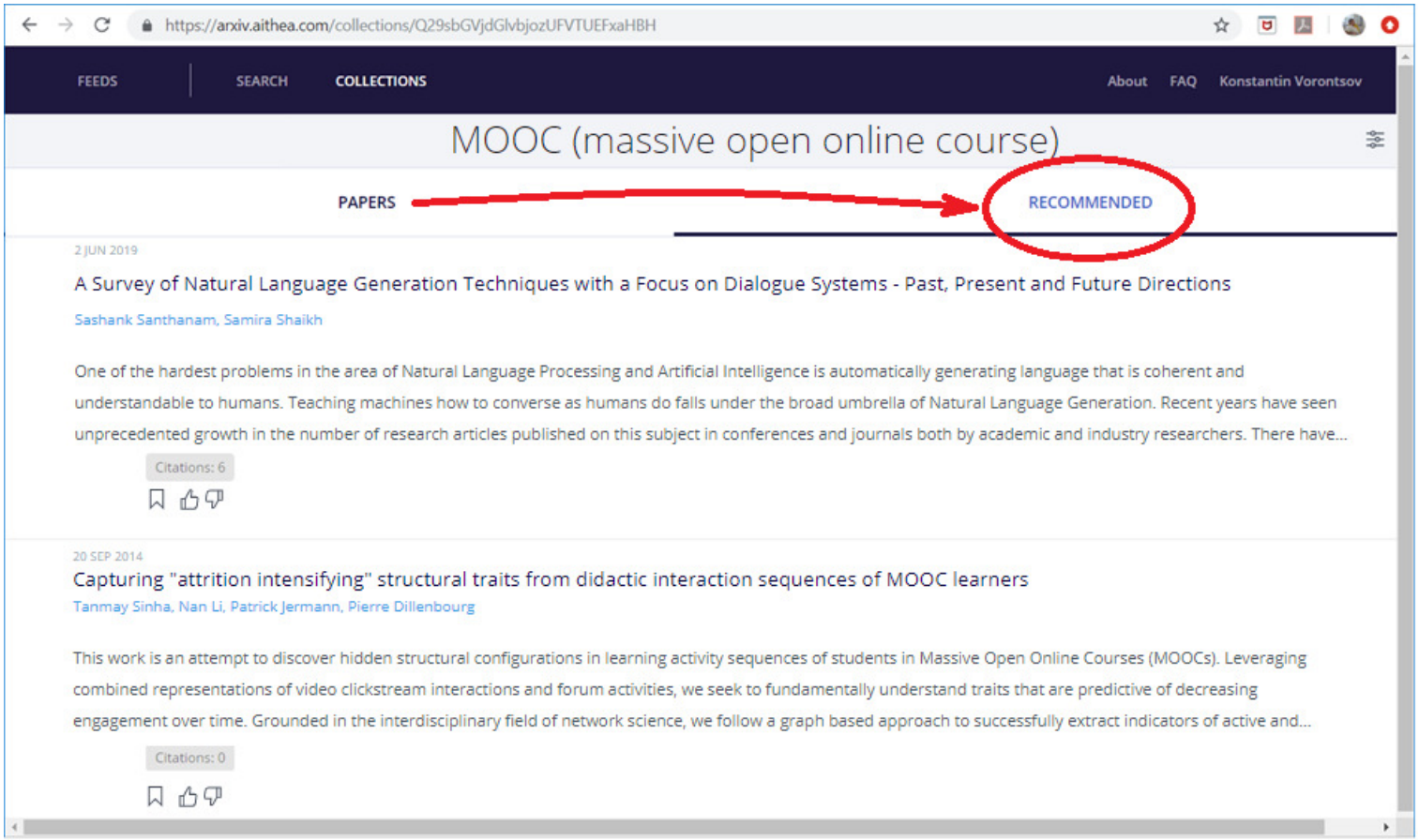
Тематическая подборка пользователя:



Разработка: http://aithea.com, http://ddecisions.ai, http://machine-intelligence.ru

Поисково-рекомендательная система SciSearch.ai

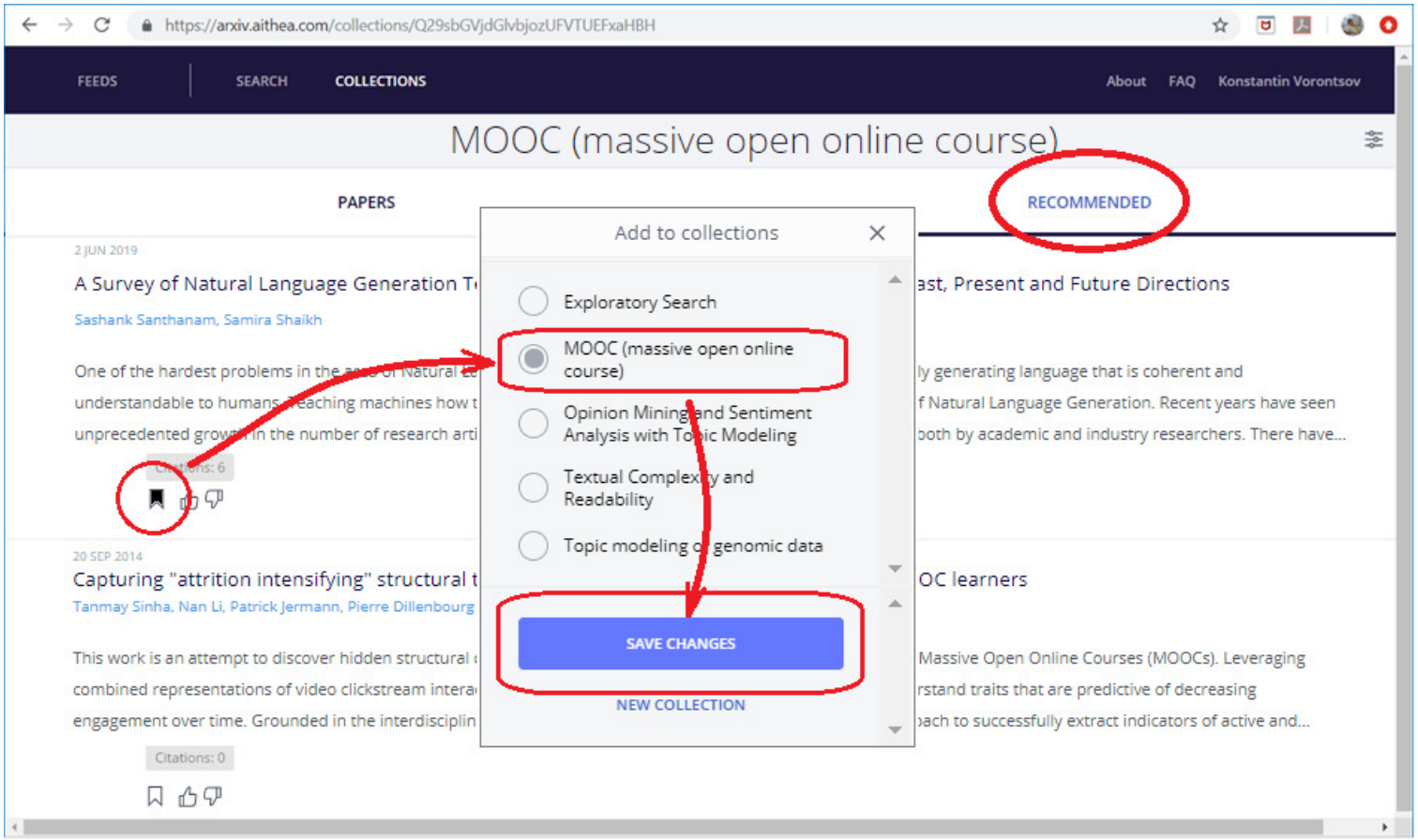
Список статей, рекомендуемых для добавления в подборку:



Разработка: http://aithea.com, http://ddecisions.ai, http://machine-intelligence.ru

Поисково-рекомендательная система SciSearch.ai

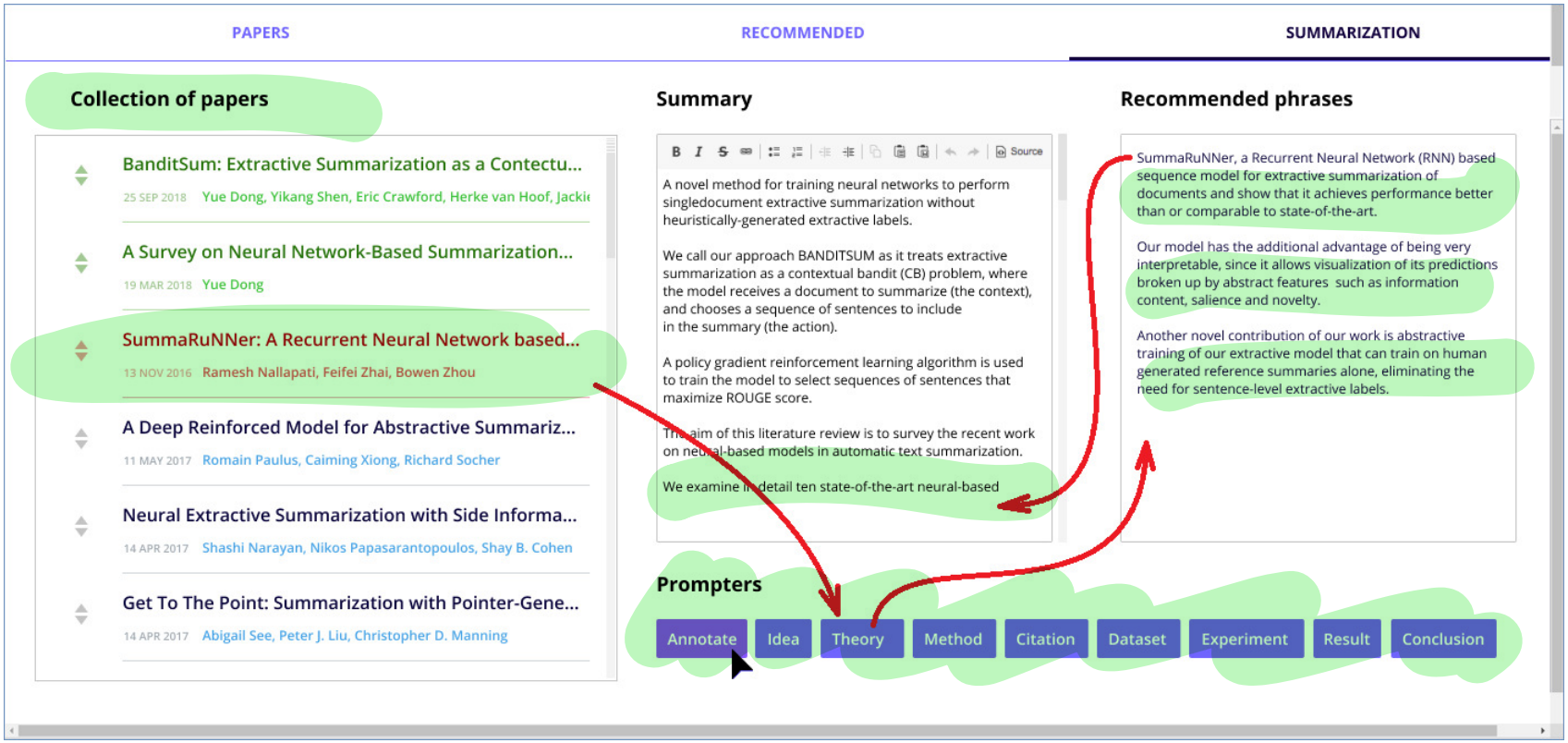
Добавление статьи из списка рекомендаций в подборку:



Разработка: http://aithea.com, http://ddecisions.ai, http://machine-intelligence.ru

Полуавтоматическое реферирование тематических подборок

Рекомендации фраз для реферата с помощью суфлёров:



А.Власов. Методы полуавтоматической суммаризации подборок научных статей. 2020. ФУПМ МФТИ.

С.Крыжановская. Технология полуавтоматической суммаризации тематических подборок научных статей. 2022. ВМК МГУ.

Концепция MAHS (Machine Aided Human Summarization)

1. **Система** рекомендует *сценарий реферата* — список статей подборки, ранжированный в порядке упоминания
2. **Пользователь** может скорректировать сценарий в соответствии со своими целями и творческим замыслом
3. В цикле по ранжированному списку статей подборки:
   - **пользователь** запрашивает аспекты статьи у суфлёров: «как другие авторы ссылаются на эту статью», «цель», «идея», «подход», «достижение», «недостаток», «результат», «вывод» и т.д.
   - **суфлёр** выдаёт ранжированный список найденных фраз
   - **пользователь** добавляет фразу из поисковой выдачи и корректирует её в соответствии с целями и замыслом

А.Власов. Методы полуавтоматической суммаризации подборок научных статей. 2020. ФУПМ МФТИ.
С.Крыжановская. Технология полуавтоматической суммаризации тематических подборок научных статей. 2022. ВМК МГУ.

Полуавтоматическое реферирование тематических подборок

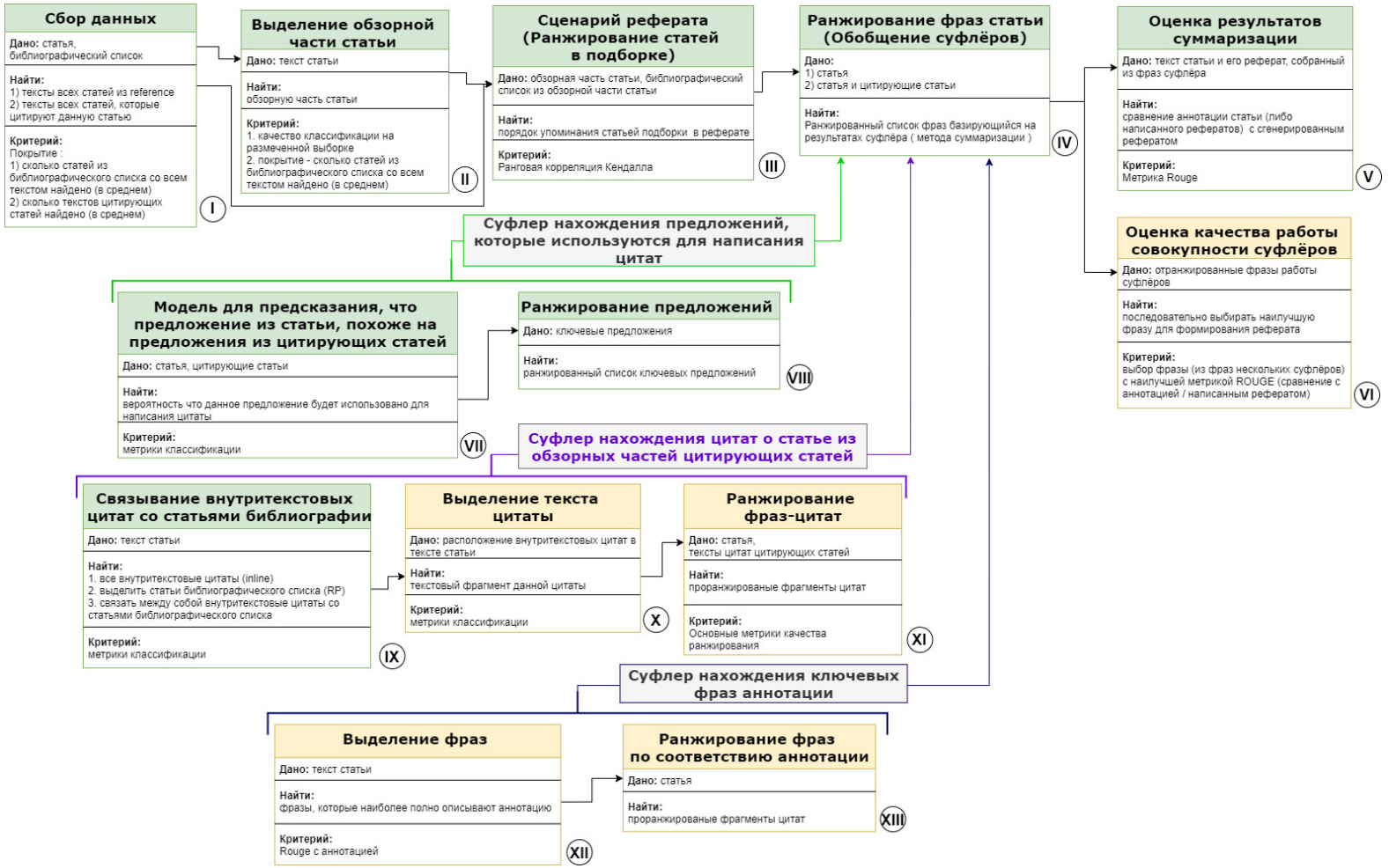
Задачи машинного обучения для MAHS:

1. Формирование обучающей выборки: paper $\rightarrow$ (refs, survey)
2. Ранжирование статей подборки для сценария реферата
3. Выбор релевантных фраз из текста статьи для суфлёра
4. Ранжирование выбранных фраз для каждого суфлёра
5. Выбор начала и конца контекста фразы, в частности, выбор релевантного контекста вокруг ссылки:

Few contextual citation graphs are publicly available. The ACL Anthology Network (AAN) (Radev et al., 2009) is one such contextual citation graph built from the ACL Anthology corpus (Bird et al., 2008), consisting of 24.6K papers manually augmented with citation information. CiteSeer (Giles et al., 1998) provides a large corpus consisting of 1.0M papers with full text and bibliography entries parsed from PDFs. Saier and Farber (2019) introduces a contextual citation graph of approximately 1.0M arXiv papers with full text LaTeX parses where citations are linked to papers in the Microsoft Academic Graph.

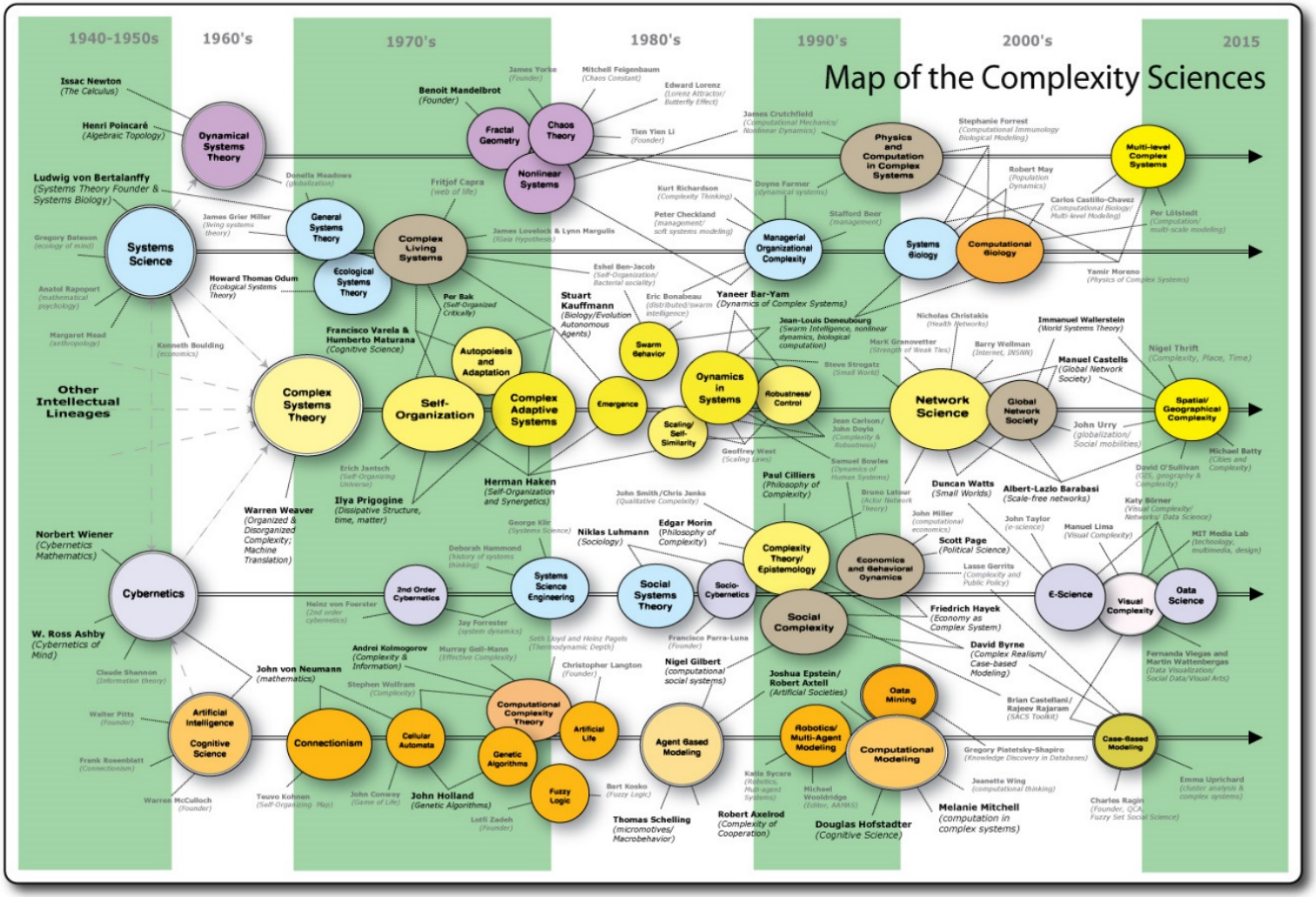
M. Yasunaga et al. ScisummNet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. 2019.

Систематизация задач машинного обучения для MAHS



А.Власов. Методы полуавтоматической суммаризации подборок научных статей. 2020. ФУПМ МФТИ.

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

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.

Направления исследований

- Участие в разработке «Мастерской знаний»
- Эффективный векторный поиск в коллекциях 100M+
- Мультиязычные нейросетевые модели языка (ruSciBERT)
- Сервис полуавтоматической суммаризации
  — задача выделения обзорных разделов из публикации
  — задача ранжирования статей (сценарий реферата)
  — задачи ранжирования фраз для суфлёров
  — задача выделения цитирующего фрагмента
- Сервис выявления новых тем и трендов
- Сервис визуализации «карты знаний»
- Сервис контекстных подсказок

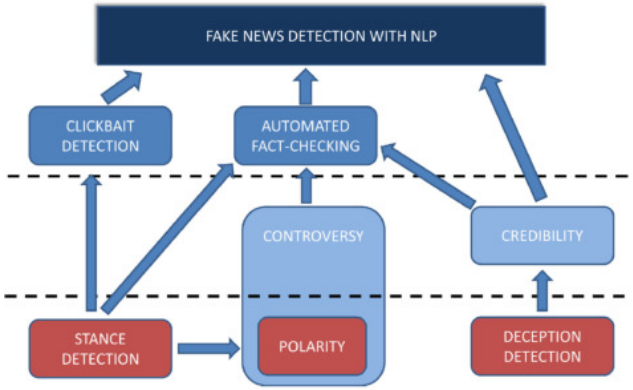
Концепция «новостного коллайдера»

Цель создания адронного коллайдера — сталкивая потоки частиц, узнать больше о строении материи

Цель создания новостного коллайдера — сталкивая потоки новостей, защитить общество от угроз эпохи постправды и информационных войн

Область исследований «Fake News Detection»

1. **Deception Detection**
   выявление обмана в тексте
2. **Automated Fact-Checking**
   автоматическая проверка фактов
3. **Stance Detection**
   выявление позиции за или против
4. **Controversy Detection**
   выявление и кластеризация разногласий
5. **Polarization Detection**
   выявление полярных позиций
6. **Clickbait Detection**
   противоречия заголовка и текста
7. **Credibility Scores**
   оценка достоверности источников



E.Saquete, D.Tomas, P.Moreda, P.Martinez-Barco, M.Palomar. Fighting post-truth using natural language processing: a review and open challenges. Expert Systems With Applications, Elsevier, 2020.

Задачи Propaganda/Manipulation/Persuasion Detection

Базовая разметка: «фрагмент, метка класса»

Gallia est omnis divisa in partes tres, quarum unam incolunt Belgae, aliam Aquitani, tertiam qui ipsorum lingua Celtae, nostra Galli appellantur. Hi omnes lingua, institutis, legibus inter se differunt. Gallos ab Aquitanis Garumna flumen, a Belgis Matrona et Sequana dividit. Horum omnium fortissimi sunt Belgae, propterea quod a cultu atque humanitate provinciae longissime absunt, minimeque ad eos mercatores saepe commeant atque ea quae ad effeminandos animos pertinent important, proximique sunt Germanis, qui trans Rhenum incolunt, quibuscum continenter bellum gerunt. Qua de causa Helvetii quoque reliquos Gallos virtute praecedunt, quod fere cotidianis proeliis cum Germanis contendunt, cum aut suis finibus eos prohibent aut ipsi in eorum finibus bellum gerunt. Eorum una pars, quam Gallos obtinere dictum est, initium capit a flumine Rhodano, continetur Garumna flumine, Oceano, finibus Belgarum, attingit etiam ab Sequanis et Helvetiis flumen Rhenum, vergit ad septentriones. Belgae ab extremis Galliae finibus oriuntur, pertinent

Manipulative Wording: Loaded Language
Attack on Reputation: Smears
Manipulative Wording: Exaggeration
Justification: Appeal to Values

Commissio
PopulusQue
Europaea

Упрощённая разметка: «предложение, метка класса»

Продвинутая разметка: «фрагмент, мишень, метка класса»

SemEval-2023 task 3. Detecting the genre, the framing, and the persuasion techniques in online news in a multi-lingual setup.
https://propaganda.math.unipd.it/semeval2023task3

G.Martino, P.Nakov et al. A survey on computational propaganda detection. 2020.

Типология угроз и задачи их автоматической детекции

воздействия → фейки → пропаганда → инф.война

1. детекция приёмов манипулирования
2. детекция замалчивания
3. детекция обмана (deception detection), слухов (rumors d.), мистификаций (hoaxes d.)
4. детекция кликбэйта (clickbait detection)
5. автоматическая проверка фактов (auto fact-checking)
6. детекция позиции (stance d.), противоречий (controversy d.), поляризации (polarization d.)
7. выявление конструктов картины мира: идеологем, мифологем
8. оценивание возможных психо-эмоциональных реакций
9. выявление целевых аудиторий воздействия
10. оценивание и предсказание скорости распространения (virality prediction)
11. оценивание достоверности источников (credibility scores)
12. детекция прямой агрессии (угрозы, призывы, провокации, вербовка, экстремизм)

E.Saquete, D.Tomas, P.Moreda, P.Martinez-Barco, M.Palomar. Fighting post-truth using natural language processing: A review and open challenges // Expert Systems With Applications, Elsevier, 2020.

Типы задач ML/NLU для мониторинга медиа-пространства

1. **Классификация текста (сообщения/предложения) целиком**
   - deception detection, fact-checking, text credibility
2. **Классификация пары текстов**
   - stance, controversy, polarization, clickbait detection
   - выявление противоречий, разногласий, замалчивания
3. **Разметка текста (выделение и классификация фрагментов)**
   - поиск лингвистических маркеров (linguistic-based cues) в тексте
   - детекция приёмов манипулирования
   - выявление конструктов картины мира: мифологем, идеологем
   - выявление психо-эмоциональных реакций и целевых аудиторий
4. **Кластеризация или тематическое моделирование**
   - кластеризация мнений по заданной теме (controversy detection)
   - выявление поляризованных мнений (polarization detection)
   - выявление мнений как сочетаний слов, семантических ролей и тональностей
   - выявление «картин мира» – устойчивых сочетаний суждений и идеологем

ПРО//ЧТЕНИЕ — технологический конкурс Up Great

Задача: поиск смысловых ошибок в сочинениях ЕГЭ по русскому, литературе, истории, обществознанию, английскому

Период: декабрь 2019 — декабрь 2022

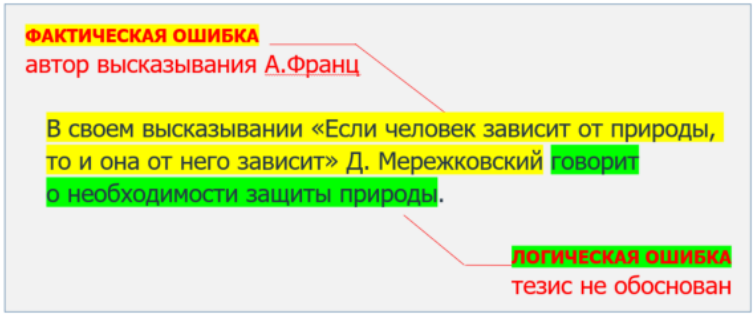
Призовой фонд:
— 100M руб. русский язык
— 100M руб. английский язык

Up Great technology contest

READ // ABLE
comprehension technology

Типов ошибок: 152
(р:70 л:16 о:23 и:20 а:23)
Подтипов ошибок: 236
(р:112 л:19 о:29 и:26 а:50)

Алгоритм должен выделять ошибки и давать их объяснения.



Официальный сайт конкурса: http://ai.upgreat.one

Сравнение двух разметок (алгоритма и эксперта)

1. насколько точно предсказана оценка за сочинение
2. насколько точно предсказаны фрагменты ошибок и блоков
3. насколько точно совпадают границы фрагментов
4. совпадают ли типы и подтипы ошибок
5. насколько содержательны сгенерированные пояснения

Модели разметки для формализации гуманитарных знаний

Цель — автоматизация обработки текстовых источников (контент-анализа и др.) в социогуманитарных исследованиях

Гипотеза: достаточно четырёх базовых операций разметки:

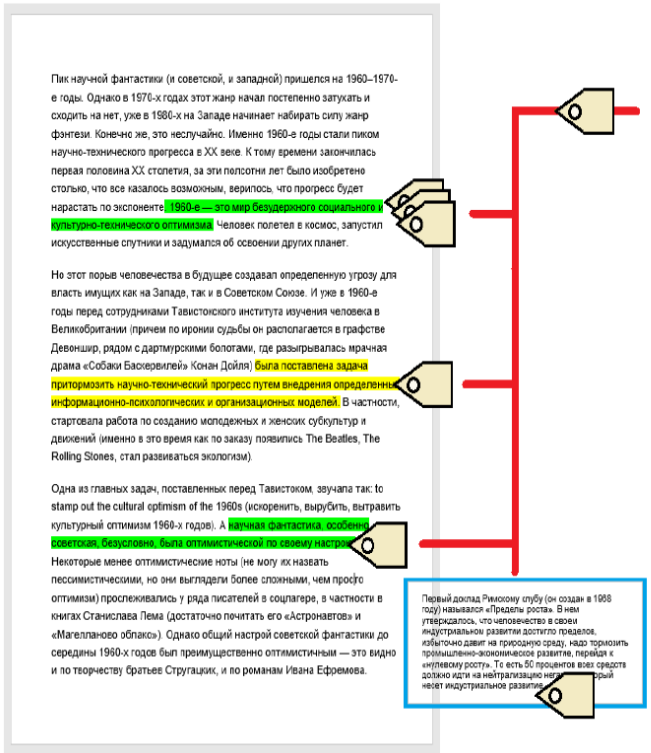
1. выделить фрагмент
2. классифицировать (тегировать) фрагмент по рубрикатору
3. связать несколько фрагментов
4. дать комментарий (затекст) к фрагменту или связи

Задачи универсализации обучаемой модели разметки:

1. унификация правил разметки и инструментария разметки
2. унификация нейросетевой архитектуры модели разметки
3. унификация методики оценивания моделей разметки

Унификация правил разметки и инструментария разметки

Обобщение классических задач компьютерной лингвистики (NER, SentAn, SemRL, SyntPars), задач выявления манипуляций, поляризации, смысловых ошибок в академических эссе и др.



Разметка состоит из элементов

Элемент разметки может содержать любое число фрагментов, затекстов и тегов

Теги (классы) выбираются из словаря тегов

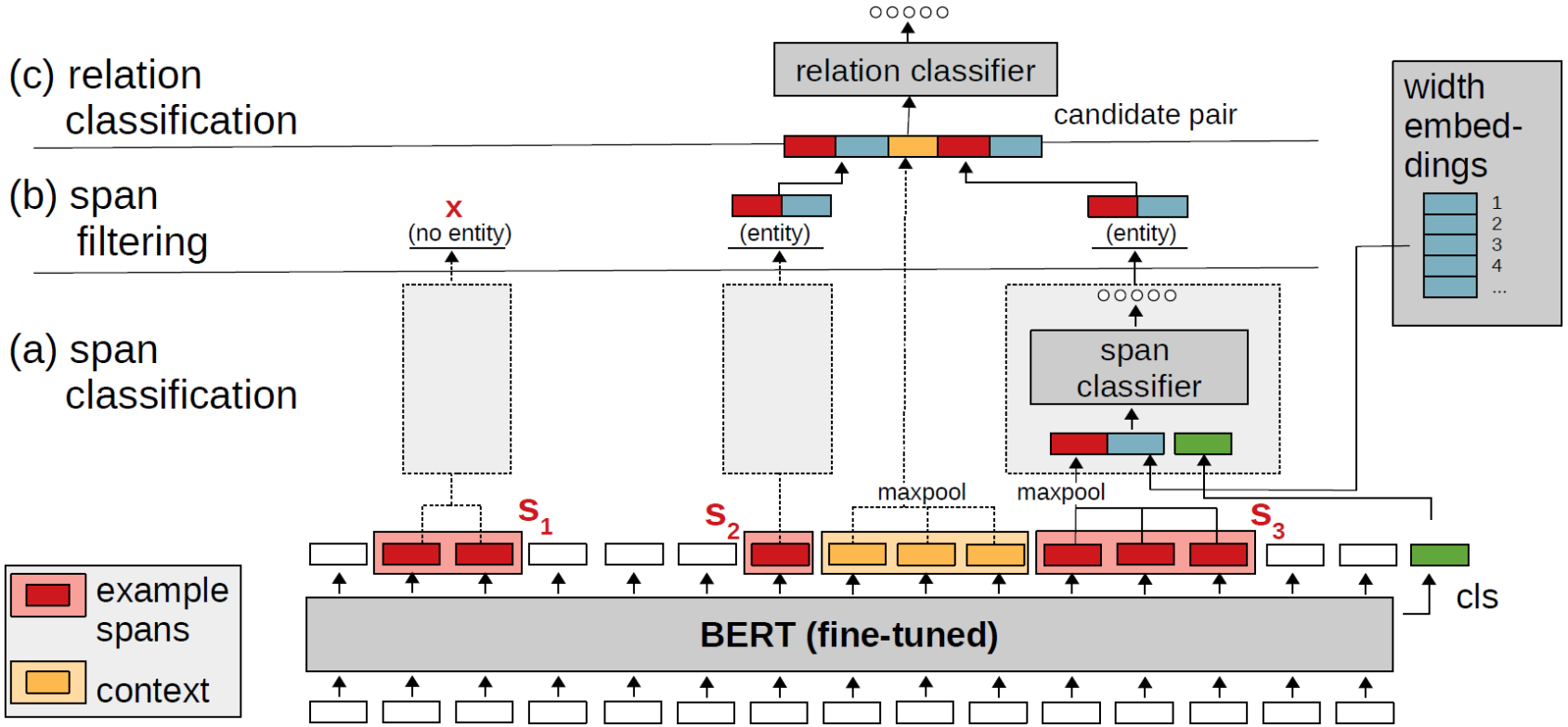
Фрагмент задаётся началом и концом, может иметь один или несколько тегов:

BEGIN была поставлена задача
притормозить научно-технический прогресс путем внедрения определен
информационно-психологических и организационных моделей. END TAG

Затекст может выбираться из словаря фраз или свободно генерироваться по контексту, может иметь один или несколько тегов

Технический регламент конкурса ПРО//ЧТЕНИЕ (http://ai.upgreat.one)

Унификация нейросетевых архитектур моделей разметки

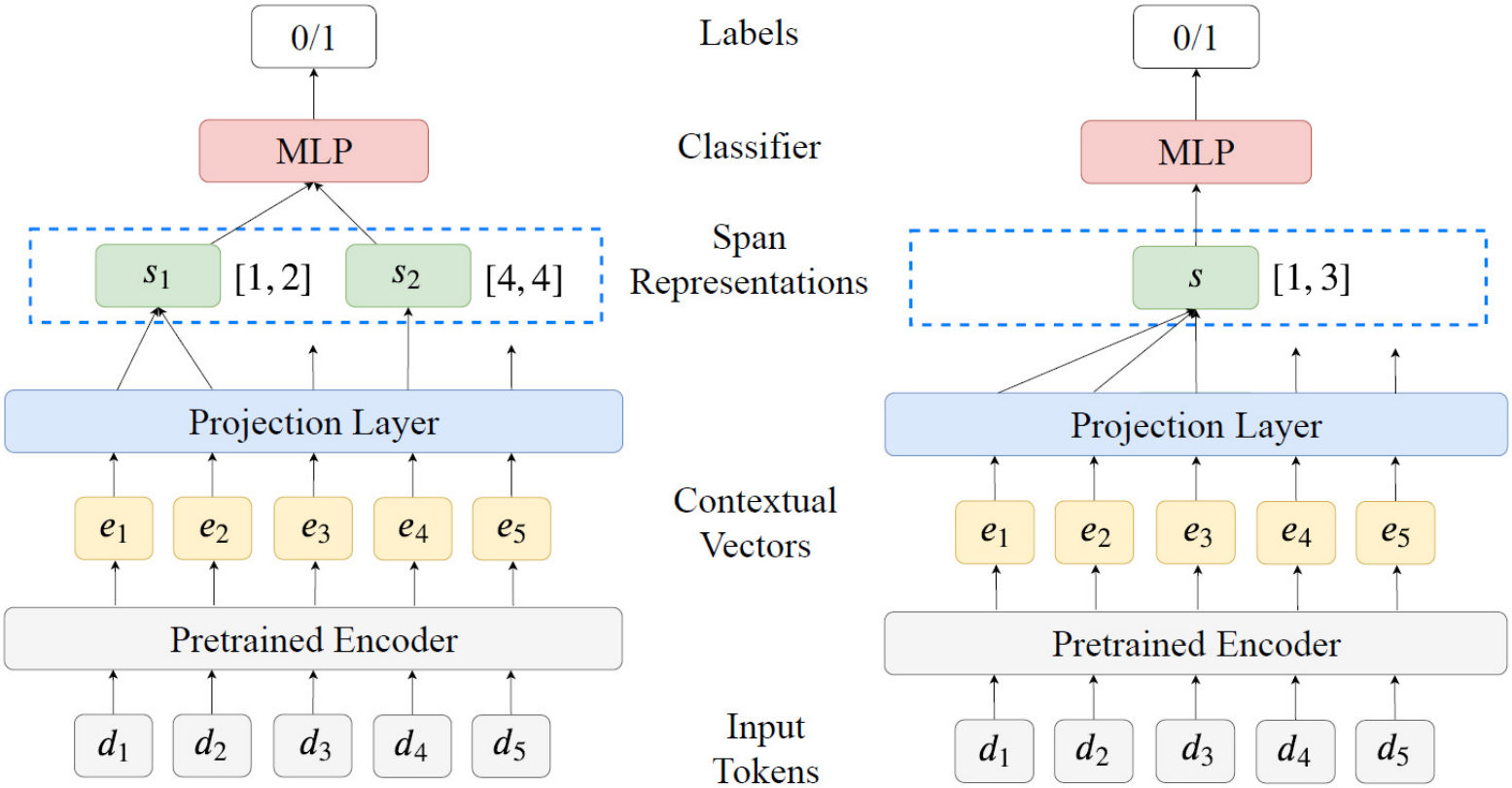


M.Eberts, A.Ulges. Span-based joint entity and relation extraction with transformer pre-training. 2020.

L.Anisiutin, T.Batura, N.Shvarts. Information extraction from news texts using a joint deep learning model. 2021.

Wayne Xin Zhao et al. A Survey of Large Language Models. ArXiv, 29 Jun 2023.

Сравнение методов формирования эмбедингов фрагментов



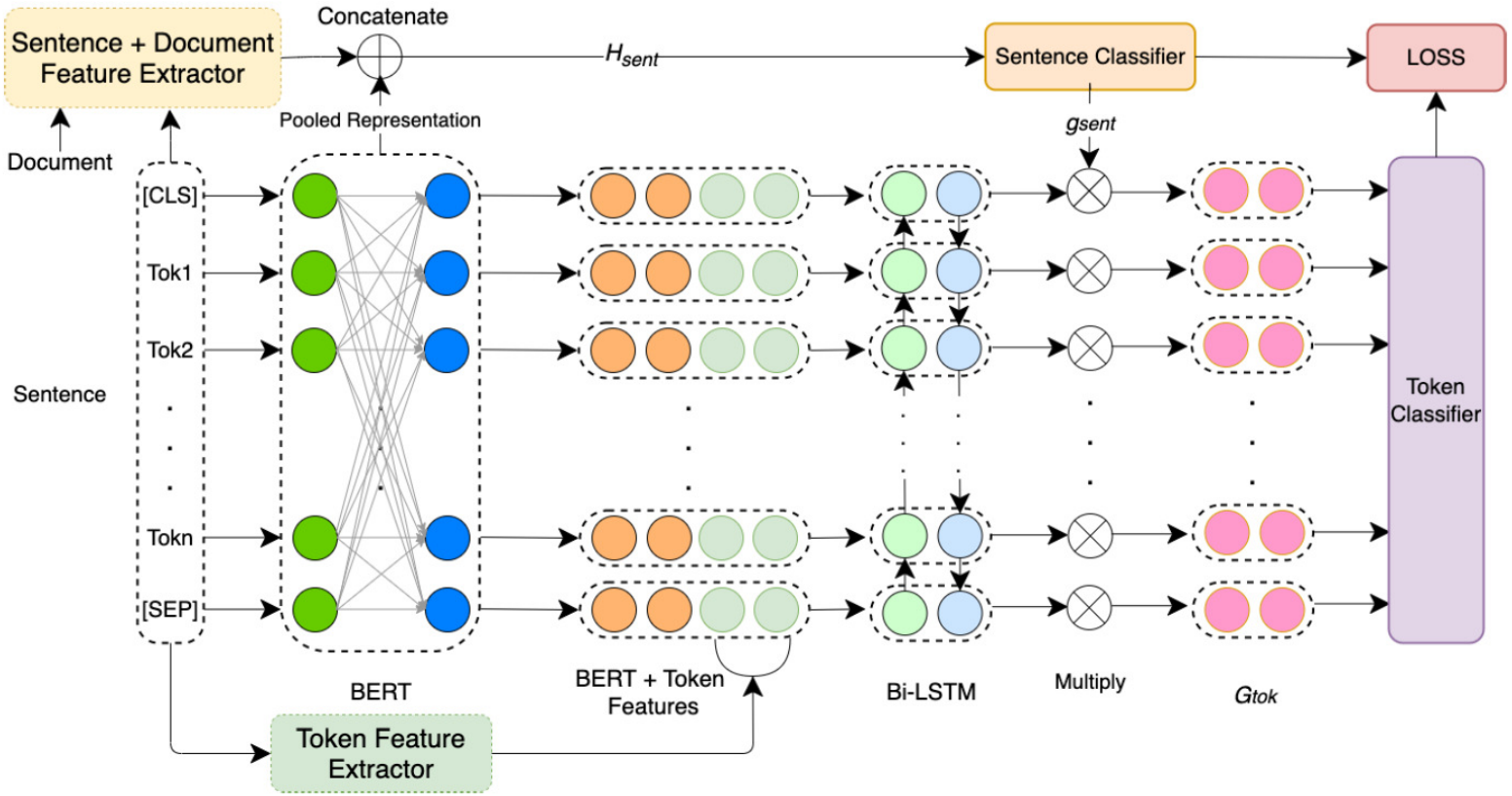
Полнота (recall) до 90% на задачах NER, SRL, Mention Detection

---

Xiaoya Li et al. A Unified MRC Framework for Named Entity Recognition. 2022.
S.Toshniwal et al. A Cross-Task Analysis of Text Span Representations. 2020.

Извлечение признаков предложений и документов

Задача детекции фрагментов с приёмами пропаганды



Sopan Khosla et al. LTIatCMU at SemEval-2020 Task 11: Incorporating Multi-Level Features for Multi-Granular Propaganda Span Identification. 2020.

Унификация методики оценивания моделей разметки

- В основе методики — сравнение пар разметок текста: «алгоритм – эксперт», «эксперт-1 – эксперт-2», путём оптимального сопоставления их элементов
- Вводятся меры согласованности пары разметок $\mathrm{Con}_k(A, B)$
- Вводится их средневзвешенная согласованность $\mathrm{Con}(A, B)$
- СТАР (Средняя Точность Алгоритмической Разметки) — средняя по размеченной выборке согласованность $\mathrm{Con}(A, E)$ разметки модели A и разметки эксперта E
- СТЭР (Средняя Точность Экспертной Разметки) — средняя по размеченной выборке согласованность $\mathrm{Con}(E_1, E_2)$ разметок двух экспертов, E_1 и E_2
- ОТАР = СТАР / СТЭР (Относительная Точность Алгоритмической Разметки) — если выше 100%, то это означает, что алгоритм работает не хуже экспертов

Технический регламент конкурса ПРО//ЧТЕНИЕ (http://ai.upgreat.one)

Направления исследований

- Технология формализации гуманитарных знаний (разметка + языковая модель + оценивание)
- Универсальная нейросетевая модель для автоматической разметки текстов (на основе архитектуры трансформера)
- Выявление ценностей социокультурного кода
- Выявление манипулятивных приёмов
- Выявление поляризации общественного мнения
- Выявление психо-эмоциональных воздействий
- Обнаружение новых тем и событий (topic detection and tracking)

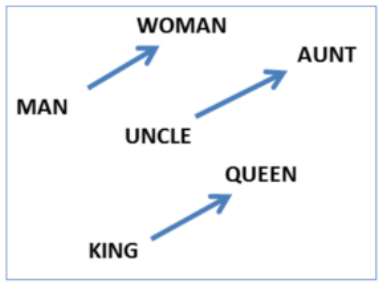
Эволюция подходов машинного обучения в анализе текстов

Декомпозиция задач по уровням пирамиды NLP

- морфологический анализ, лемматизация, опечатки
- синтаксический анализ, выделение терминов, NER
- семантический анализ, выделение фактов, тем

Модели векторных представлений (эмбедингов) слов на основе матричных разложений

- модели дистрибутивной семантики:
  word2vec [Mikolov, 2013], FastText [Bojanowski, 2016]
- тематические модели LDA [Blei, 2003], ARTM [2014]



Нейросетевые модели локальных контекстов

- рекуррентные нейронные сети
- модели внимания и трансформеры:
  BERT [2018], GPT-3 [2020] и др.

$$\text{softmax}\left(\frac{Q \times K^T}{\sqrt{d}}\right) V$$

Постановка задачи тематического моделирования

Дано:

- W — конечное множество (словарь) термов (слов, токенов)
- D — конечное множество (коллекция) документов
- n_{dw} — частота терма $w \in W$ в документе $d \in D$

Найти: вероятностную тематическую языковую модель

$$p(w|d) = \sum_{t \in T} p(w \mid \cancel{d}, t)\, p(t|d) = \sum_{t \in T} \phi_{wt} \theta_{td}$$

где $\phi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$ — параметры модели

Критерий: максимум логарифма правдоподобия

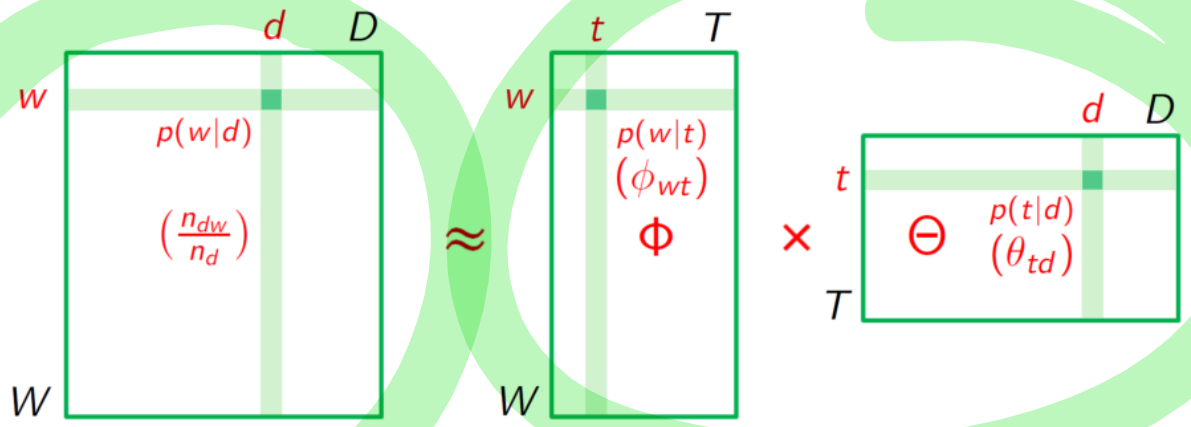
$$L(\Phi, \Theta) = \ln \prod_{d,w} p(w|d)^{n_{dw}} = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \to \max_{\Phi, \Theta}$$

при ограничениях $\phi_{wt} \geqslant 0$, $\sum_w \phi_{wt} = 1$, $\theta_{td} \geqslant 0$, $\sum_t \theta_{td} = 1$

Hofmann T. Probabilistic Latent Semantic Indexing. ACM SIGIR, 1999.

Некорректно поставленная задача матричного разложения

Низкоранговое стохастическое матричное разложение:



Если Φ, Θ — решение, то стохастические Φ', Θ' — тоже решения

- $\Phi'\Theta' = (\Phi S)(S^{-1}\Theta)$, $\operatorname{rank} S = |T|$
- $L(\Phi', \Theta') = L(\Phi, \Theta)$ — линейно не зависимые решения
- $L(\Phi', \Theta') \geqslant L(\Phi, \Theta) - \varepsilon$ — приближённые решения

Регуляризация — стандартный приём доопределения решения с помощью добавления дополнительных критериев.

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

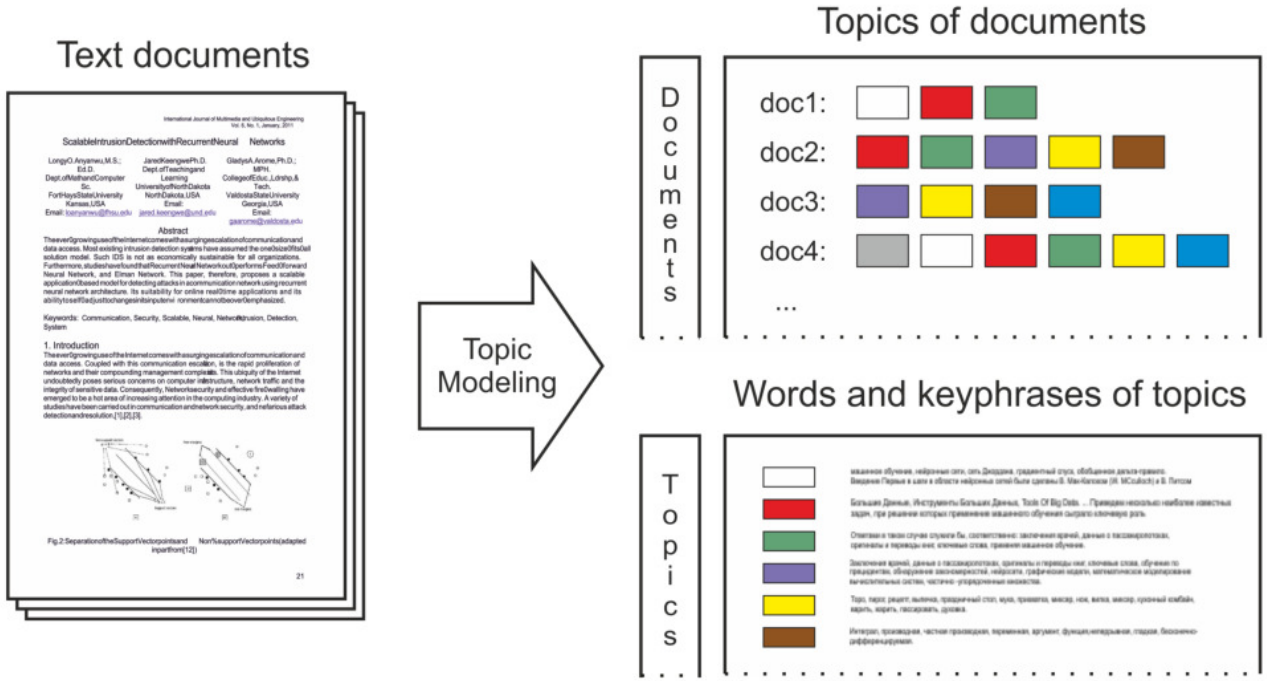
$$\text{E-шаг:} \quad p_{tdw} \equiv p(t|d,w) = \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

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

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

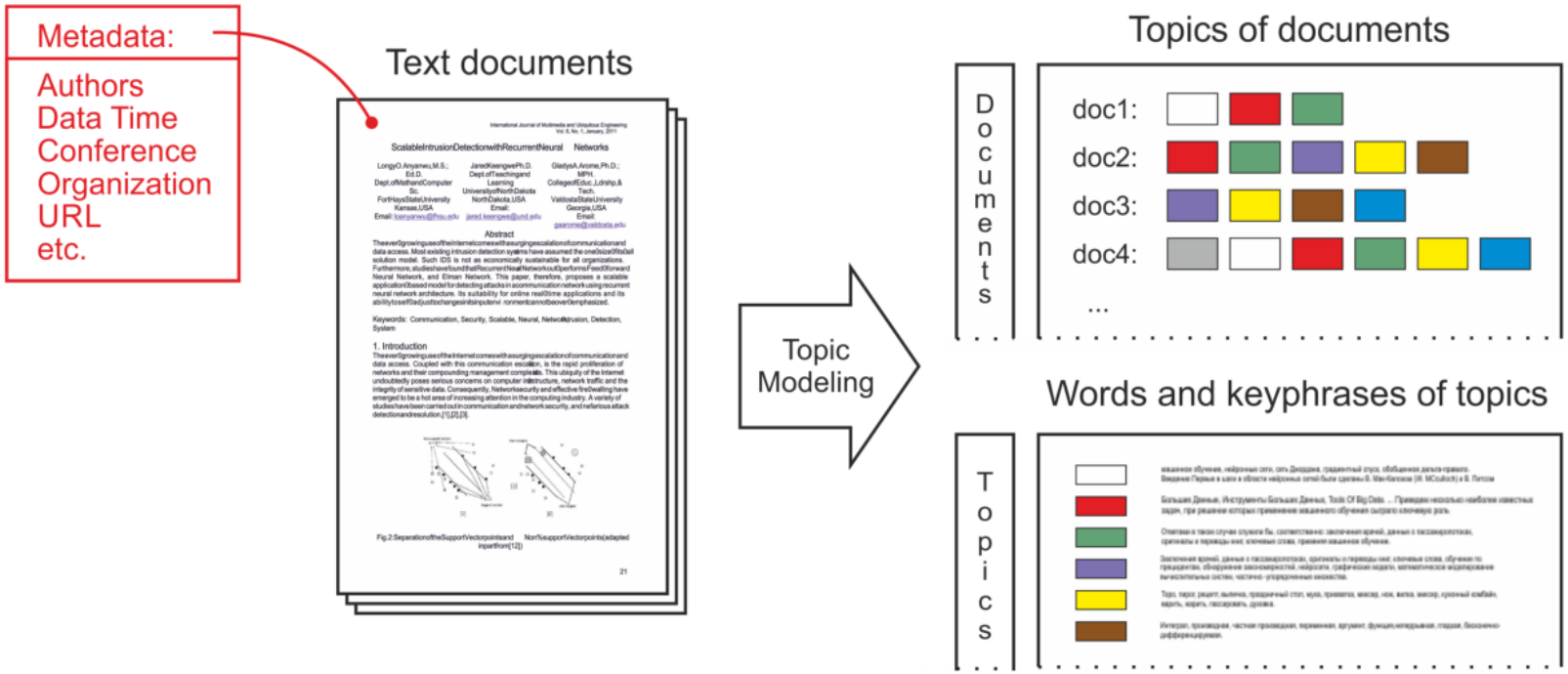
Мультимодальная тематическая модель

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



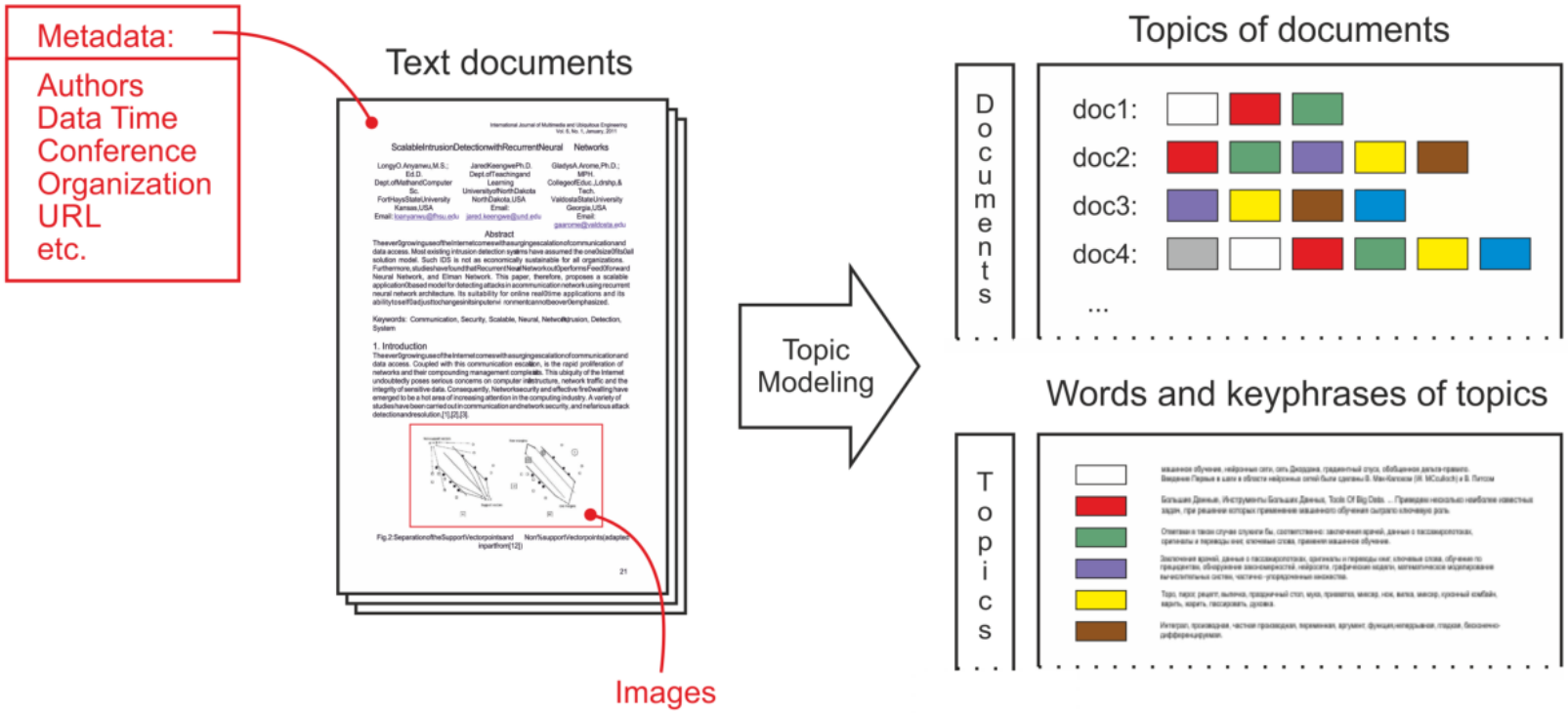
Мультимодальная тематическая модель

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



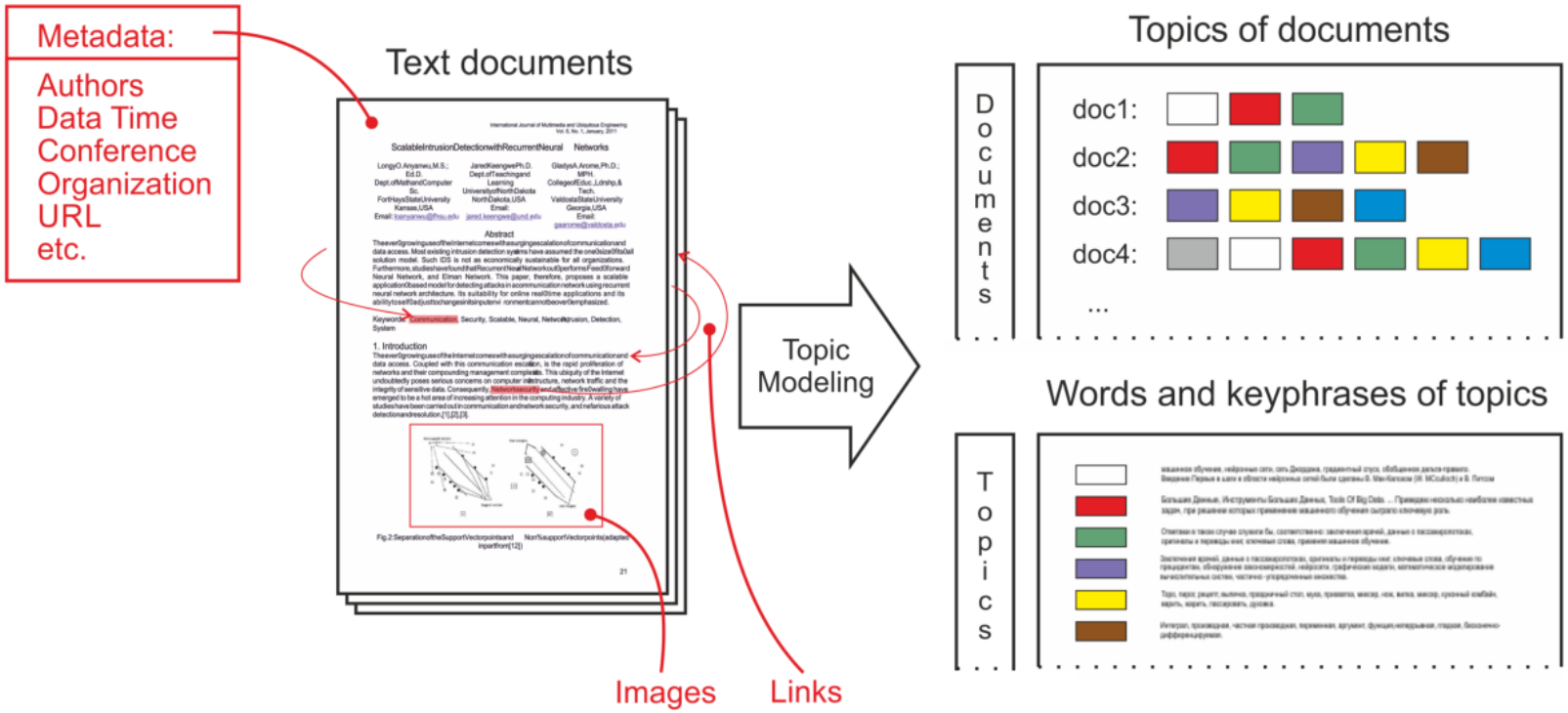
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,
$p(\text{объект}|t)$,



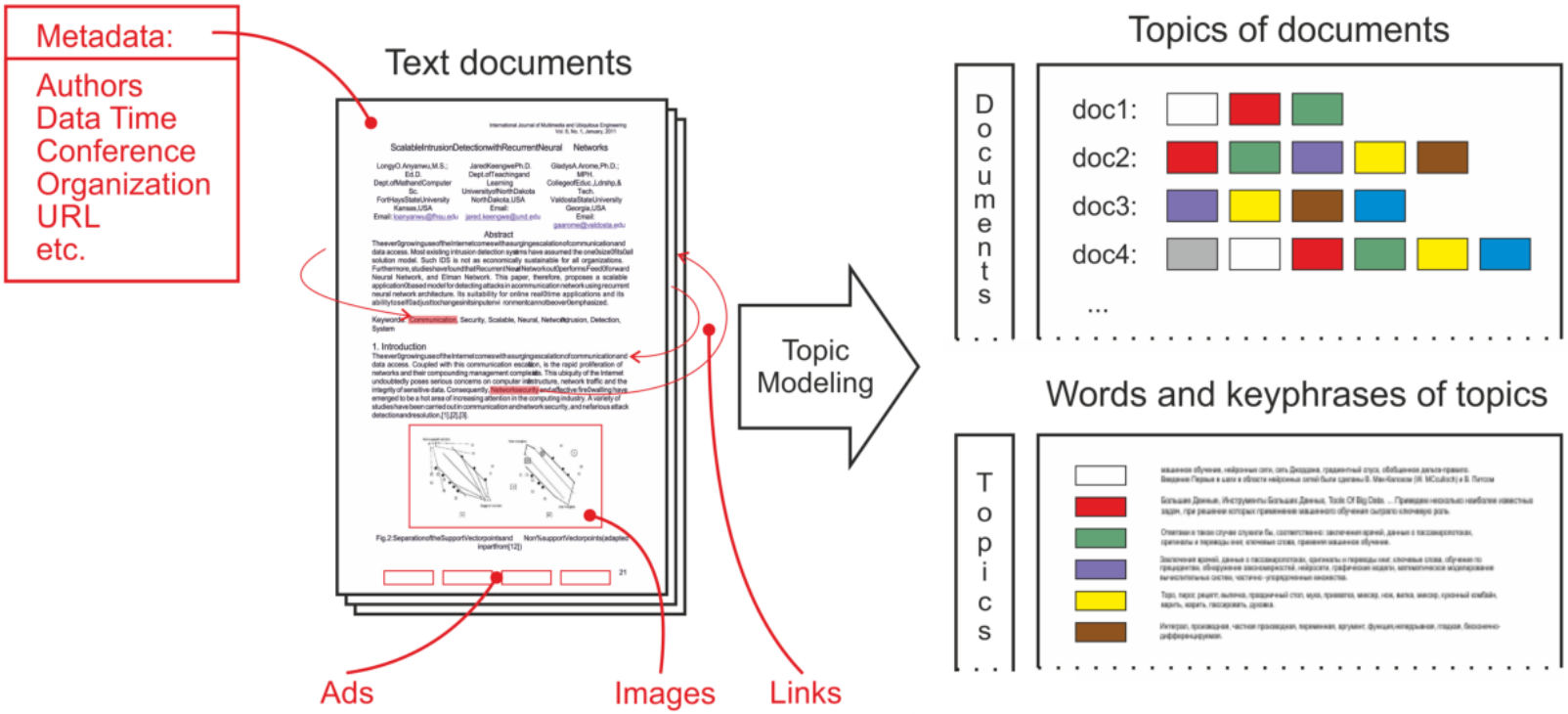
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,
$p(\text{объект}|t)$, $p(\text{ссылка}|t)$,



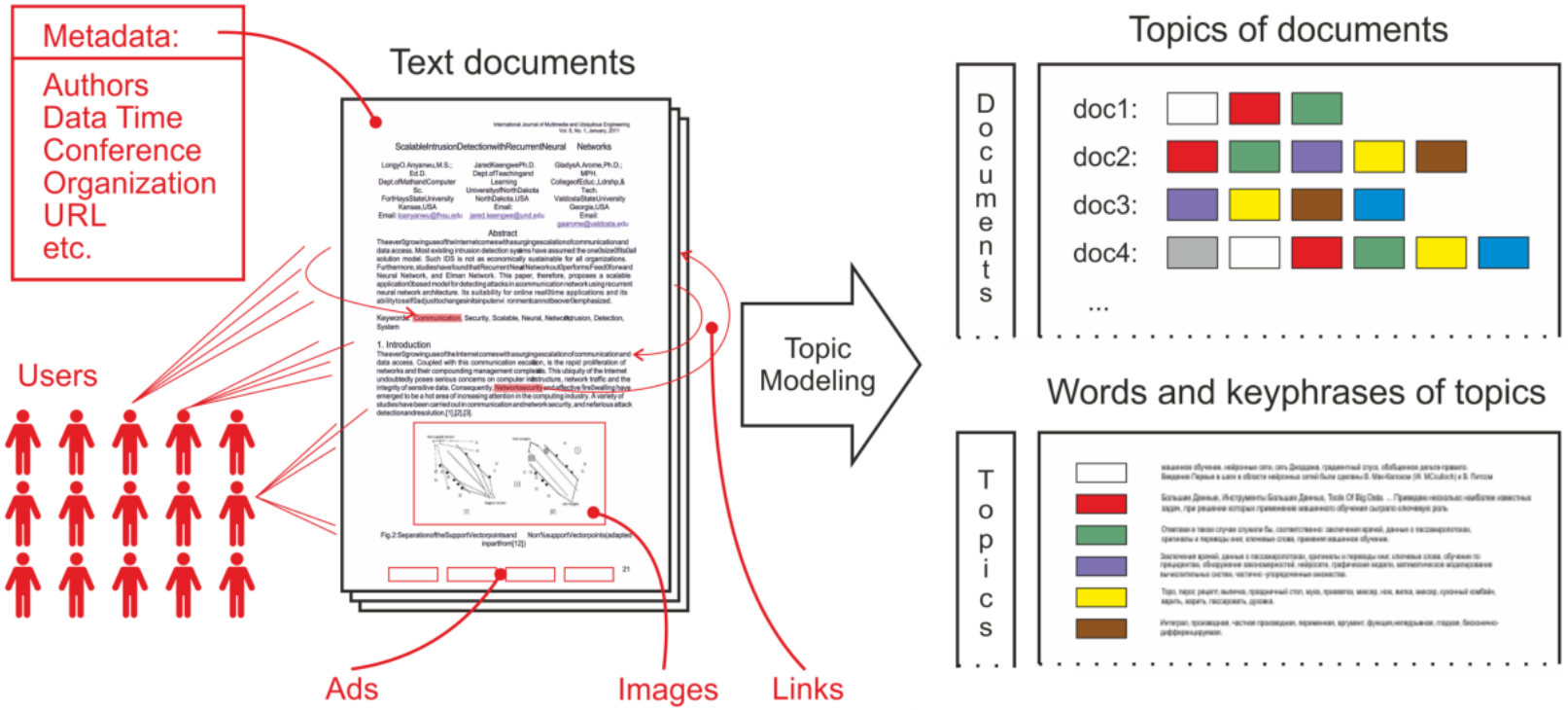
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,
$p(\text{объект}|t)$, $p(\text{ссылка}|t)$, $p(\text{баннер}|t)$,



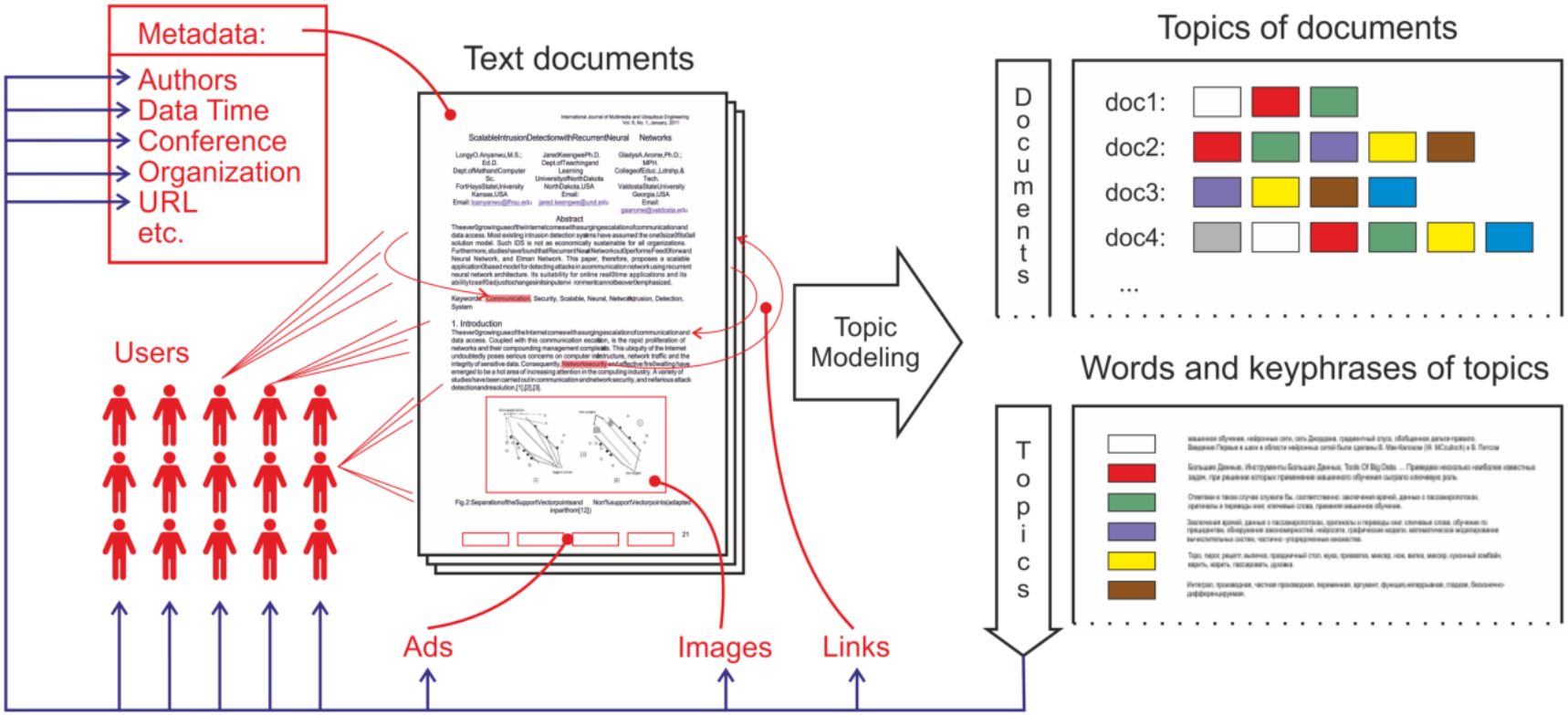
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,
$p(\text{объект}|t)$, $p(\text{ссылка}|t)$, $p(\text{баннер}|t)$, $p(\text{пользователь}|t)$



Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,
$p(\text{объект}|t)$, $p(\text{ссылка}|t)$, $p(\text{баннер}|t)$, $p(\text{пользователь}|t)$



Мультимодальная ARTM

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

Максимизация суммы log-правдоподобий с регуляризацией:

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

EM-алгоритм: метод простой итерации для системы уравнений

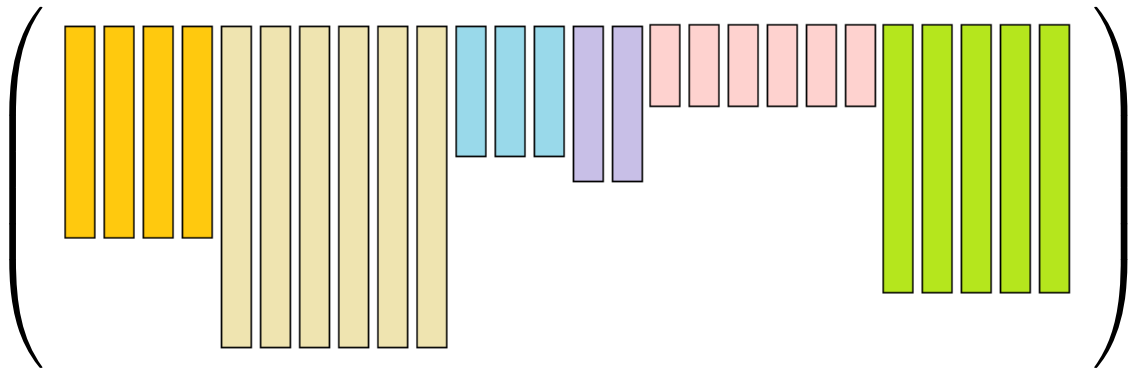
$$\text{E-шаг:} \quad p_{tdw} = \underset{t \in T}{\text{norm}} \left(\phi_{wt} \theta_{td} \right)$$

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

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

Задача максимизации функции на единичных симплексах

Пусть $\Omega = (\omega_j)_{j \in J}$ — набор нормированных неотрицательных векторов $\omega_j = (\omega_{ij})_{i \in I_j}$ различных размерностей $|I_j|$:

$\Omega =$ 

Задача максимизации функции $f(\Omega)$ на единичных симплексах:

$$\begin{cases} f(\Omega) \to \max\limits_{\Omega}; \\ \sum\limits_{i \in I_j} \omega_{ij} = 1, \quad j \in J; \\ \omega_{ij} \geqslant 0, \quad i \in I_j, \quad j \in J. \end{cases}$$

Необходимые условия экстремума и метод простых итераций

Операция нормировки вектора: $p_i = \underset{i \in I}{\text{norm}}(x_i) = \dfrac{\max(x_i, 0)}{\sum_k \max(x_k, 0)}$

Лемма. Пусть $f(\Omega)$ непрерывно дифференцируема по Ω.
Если ω_j — вектор локального экстремума задачи $f(\Omega) \to \max$ и $\exists i$: $\omega_{ij} \dfrac{\partial f}{\partial \omega_{ij}} > 0$, то ω_j удовлетворяет системе уравнений

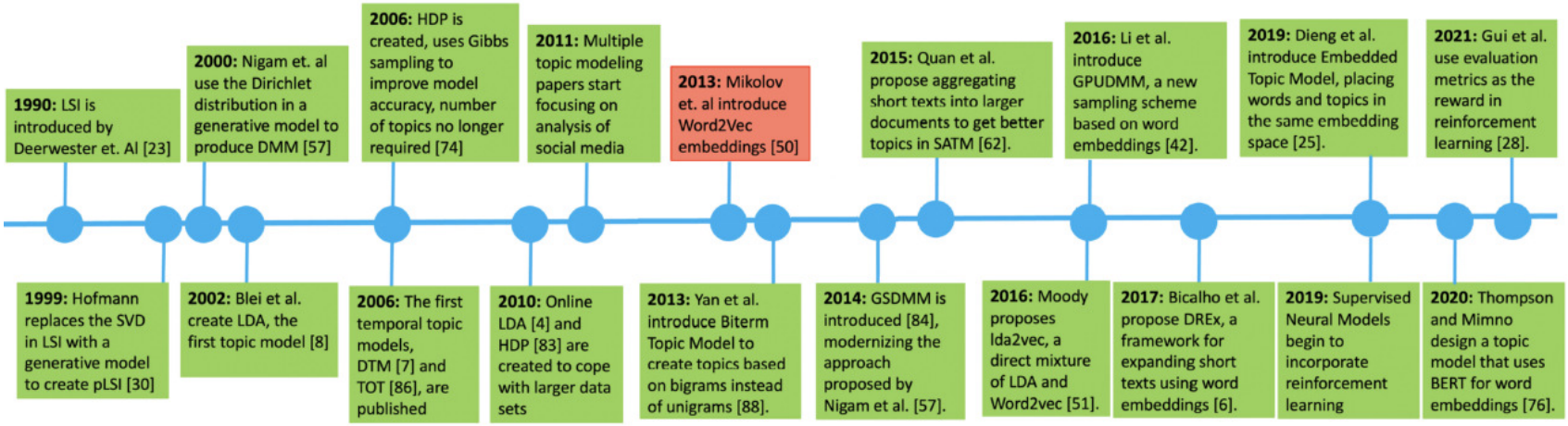
$$\omega_{ij} = \underset{i \in I_j}{\text{norm}} \left(\omega_{ij} \frac{\partial f}{\partial \omega_{ij}} \right).$$

- Численное решение системы — методом простых итераций
- Векторы $\omega_j = 0$ отбрасываются как вырожденные решения
- Итерации похожи на градиентную оптимизацию:

$$\omega_{ij} := \omega_{ij} + \eta \frac{\partial f}{\partial \omega_{ij}},$$

но учитывают ограничения и не требуют подбора шага η

Эволюция тематического моделирования



Neural Topic Models — поток публикаций начиная с 2016

Как «объединить лучшее от двух миров»?

- **Neural:** качество, универсальность, генеративность
- **Topic:** скорость, интерпретируемость, простота

Что объединяет: векторизация, оптимизация, регуляризация, гомогенизация, локализация (контекст и внимание)

Rob Churchill, Lisa Singh. The Evolution of Topic Modeling. November, 2022.

Направления исследований

- Строить любые тематические модели в pyTorch, где градиентный шаг на симплекс уже реализован
- Neural Topic Modeling (NTM) — как совмещать интерпретируемость тематических моделей с преимуществами глубоких нейросетевых моделей языка
- В том числе, создание тематических моделей внимания
- Открытая проблема — несбалансированность тем
- Автоматическое именование и аннотирование тем (каждая тема должна уметь «рассказать о себе»)
- Как приблизить долю интерпретируемых тем к 100%
- Динамическое обнаружение новых тем и трендов (Topic Detection & Tracking, First Story Detection)
- Автоматический подбор гиперпараметров, AutoML

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:

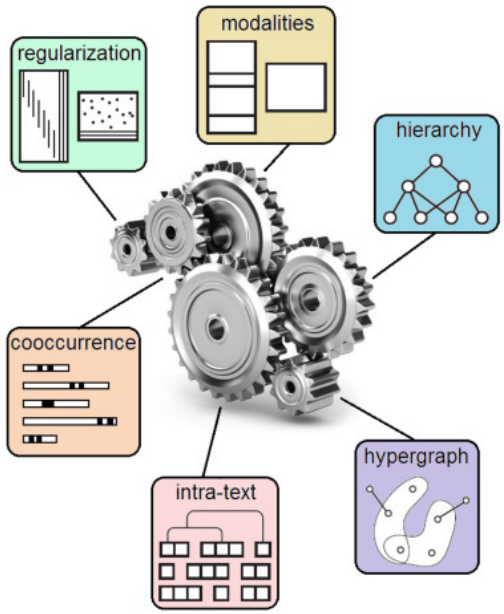
- Свободная коммерческая лицензия (BSD 3-Clause)
- Кросс-платформенность: Windows, Linux, MacOS (32/64 bit)
- Интерфейсы API: command-line, C++, and Python

K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova. BigARTM: Open source library for regularized multimodal topic modeling of large collections. AIST 2015.

Ключевые возможности библиотек BigARTM и TopicNet

BigARTM

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

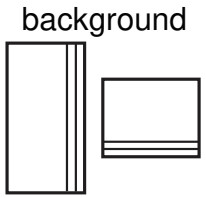


TopicNet

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

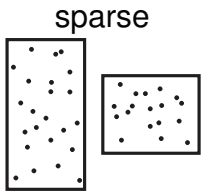
V.Bulatov, E.Egorov, E.Veselova, D.Polyudova, V.Alekseev, A.Goncharov, K.Vorontsov.
TopicNet: making additive regularisation for topic modelling accessible. LREC-2020

Регуляризаторы для улучшения интерпретируемости тем



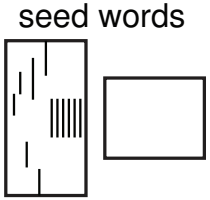
Сглаживание фоновых тем $B \subset T$:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in B} \sum_w \beta_w \ln \phi_{wt} + \alpha_0 \sum_d \sum_{t \in B} \alpha_t \ln \theta_{td}$$

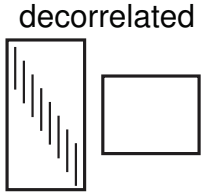


Разреживание предметных тем $S = T \setminus B$:

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in S} \sum_w \beta_w \ln \phi_{wt} - \alpha_0 \sum_d \sum_{t \in S} \alpha_t \ln \theta_{td}$$

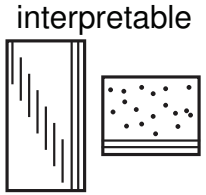


Сглаживание для выделения релевантных тем
с помощью словаря «затравочных» ключевых слов



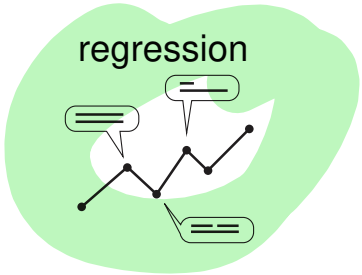
Декоррелирование для повышения различности тем:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t,s} \sum_w \phi_{wt} \phi_{ws}$$



Сглаживание + разреживание + декоррелирование
для улучшения интерпретируемости тем

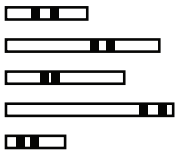
Регуляризаторы для учёта дополнительной информации



Линейная модель регрессии $\hat{y}_d = \langle v, \theta_d \rangle$ документов:

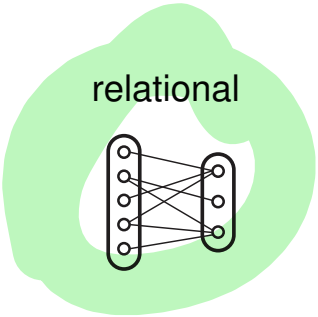
$$R(\Theta, v) = -\tau \sum_{d \in D} \left(y_d - \sum_{t \in T} v_t \theta_{td} \right)^2$$

biterm



Связи сочетаемости слов (n_{uv} — частота битерма):

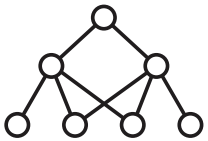
$$R(\Phi) = \tau \sum_{u \in W} \sum_{v \in W} n_{uv} \ln \sum_{t \in T} n_t \phi_{ut} \phi_{vt}$$



Связи или ссылки между документами:

$$R(\Theta) = \tau \sum_{d,c \in D} n_{dc} \sum_{t \in T} \theta_{td} \theta_{tc}$$

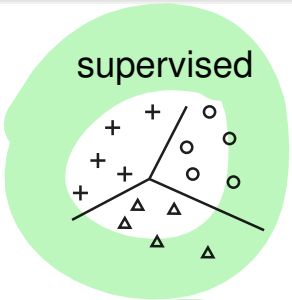
hierarchy



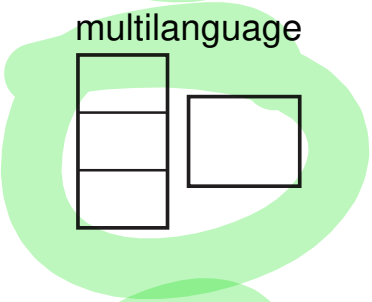
Связи родительских тем t с дочерними подтемами s:

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

Регуляризаторы для мультимодальных тематических моделей

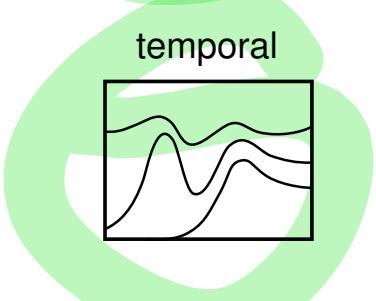


Модальности меток классов или категорий для задач классификации и категоризации текстов.



Модальность языков и регуляризация со словарём $\pi_{uwt} = p(u|w,t)$ переводов с языка k на ℓ:

$$R(\Phi,\Pi) = \tau \sum_{u\in W^k} \sum_{t\in T} n_{ut} \ln \sum_{w\in W^\ell} \pi_{uwt}\phi_{wt}$$



Темпоральные модели с модальностью времени i:

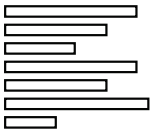
$$R(\Phi) = -\tau \sum_{i\in I} \sum_{t\in T} \left|\phi_{it} - \phi_{i-1,t}\right|.$$

Модальность геолокаций g с близостью $S_{gg'}$:

$$R(\Phi) = -\frac{\tau}{2} \sum_{g,g'\in G} S_{gg'} \sum_{t\in T} n_t^2 \left(\frac{\phi_{gt}}{n_g} - \frac{\phi_{g't}}{n_{g'}}\right)^2$$

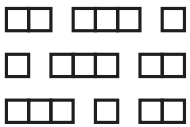
Регуляризаторы для моделирования последовательного текста

sentence



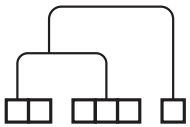
Тематические модели, учитывающие границы предложений, абзацев и секций документов

n-gram



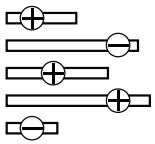
Модели с модальностями *n*-грамм, коллокаций, именованных сущностей (используем TopMine)

syntax



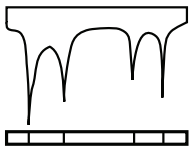
Модели, учитывающие результаты автоматического синтаксического разбора (используем UDPipe)

sentiment



Модели выделения мнений на основе тональностей, фактов, семантических ролей именованных сущностей

segmentation



Тематические модели сегментации с автоматическим определением границ сегментов

ARTM — модульный подход к синтезу требуемых моделей

Для построения композитных моделей в BigARTM не нужны ни математические выкладки, ни программирование «с нуля».

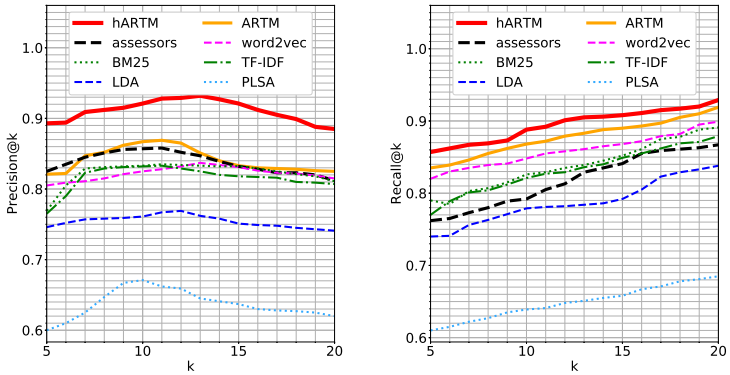
Этапы моделирования	Bayesian TM	ARTM	
	Анализ требований	Анализ требований	
Формализация:	Вероятностная модель порождения данных	Стандартные критерии	Свои критерии
Алгоритмизация:	Байесовский вывод для данной порождающей модели (VI, GS, EP)	Единый регуляризованный EM-алгоритм для любых моделей и их композиций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

-- *нестандартизуемые этапы, уникальная разработка для каждой задачи*

-- *стандартизуемые этапы*

Разведочный поиск в технологических блогах

Цель: поиск документов
по длинным текстовым запросам
— Habr.ru (175K документов),
— TechCrunch.com (760K док.).



Регуляризаторы:

$$\mathscr{L}\left(\Phi\,\Theta\right) + R\left(\text{hierarchy}\right) + R\left(\text{interpretable}\right) + R\left(\text{multimodal}\right) + R\left(\text{n-gram}\right) \to \max$$

Результаты:

- Точность и полнота 93%, превосходит асессоров и другие методы (tf-idf, BM25, word2vec, PLSA, LDA, ARTM).
- Увеличилась оптимальная размерность векторов:
  200 → 1400 (Habr.ru), 475 → 2800 (TechCrunch.com).

A.Ianina, K.Vorontsov. Regularized multimodal hierarchical topic model for document-by-document exploratory search. FRUCT–ISMW, 2019.

Поиск и классификация этно-релевантных тем в соцсетях

Цель: выявление как можно большего числа тем о национальностях и межнациональных отношениях (затравка — словарь 300 этнонимов).

(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, болеция, 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,

Регуляризаторы:

$$\mathscr{L}\left(\Phi\,\Theta\right) + R\left(\text{seed words}\right) + R\left(\text{interpretable}\right) + R\left(\text{multimodal}\right)$$
$$+ R\left(\text{temporal}\right) + R\left(\text{geospatial}\right) + R\left(\text{sentiment}\right) \to \max$$

Результаты: число релевантных тем: 45 (LDA) $\to$ 83 (ARTM).

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.

Mining ethnic content online with additively regularized topic models. 2016.

Аналогичные исследования по выделению узкой тематики

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.

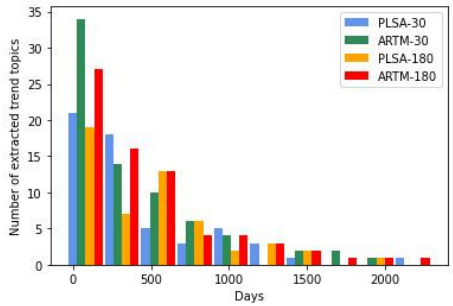
3. *M.A.Basher, A.Rahman, B.C.M.Fung.* Analyzing topics and authors in chat logs for crime investigation. 2014.

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.

Выявление трендов в коллекции научных публикаций

Цель: раннее обнаружение трендовых тем с начальным экспоненциальным ростом в области AI/ML 2009–2021 гг.



Регуляризаторы:

$$\mathscr{L}\left(\underset{\Phi\ \Theta}{\text{PLSA}}\right) + R\left(\underset{}{\text{interpretable}}\right) + R\left(\underset{}{\text{dynamic}}\right) + R\left(\underset{}{\text{multimodal}}\right) + R\left(\underset{}{\text{n-gram}}\right) \to \max$$

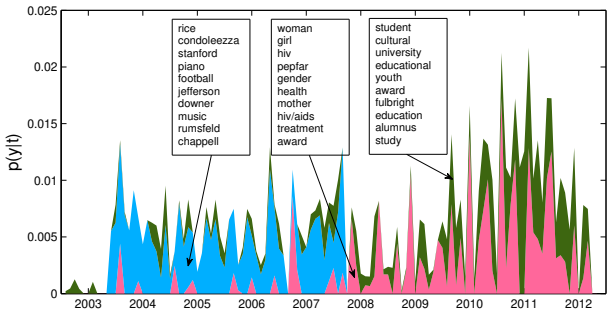
Результаты:

- выделение 90 из 91 тренда в области машинного обучения
- 63% тем выделяется за год, 79% за два года

Н.Герасименко, А.Чернявский, М.Никифорова, М.Никитин, К.Воронцов. Инкрементальное обучение тематических моделей для поиска трендовых тем в научных публикациях. Доклады РАН, 2022.

Выявление динамики тем в новостных потоках

Цель: выделение тем в коллекции пресс-релизов МИДов 4х стран, с привязкой ко времени.



Регуляризаторы:

$$\mathscr{L}\left(\text{PLSA}\right) + R\left(\text{interpretable}\right) + R\left(\text{temporal}\right) + R\left(\text{multimodal}\right)$$
$$+ R\left(\text{n-gram}\right) + R\left(\text{multilanguage}\right) \to \max$$

Результаты:

- разделение тем на событийные и перманентные
- когерентность тем: 5.5 $\to$ 6.5

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

Выделение поляризованных мнений в политических новостях

Цель: найти признаки, по которым событийная тема разделяется на кластеры-мнения

Modalities	*Pr*	*Rec*	*F1*
TF-IDF	0.51	0.95	0.67
SPO	0.59	0.7	0.64
FR	0.86	0.49	0.65
Sent	0.69	0.57	0.66
SPO+FR	0.86	0.68	0.76
SPO+Sent	0.83	0.78	0.81
FR+Sent	0.9	0.52	0.67
All	**0.77**	**0.97**	**0.86**

Регуляризаторы:

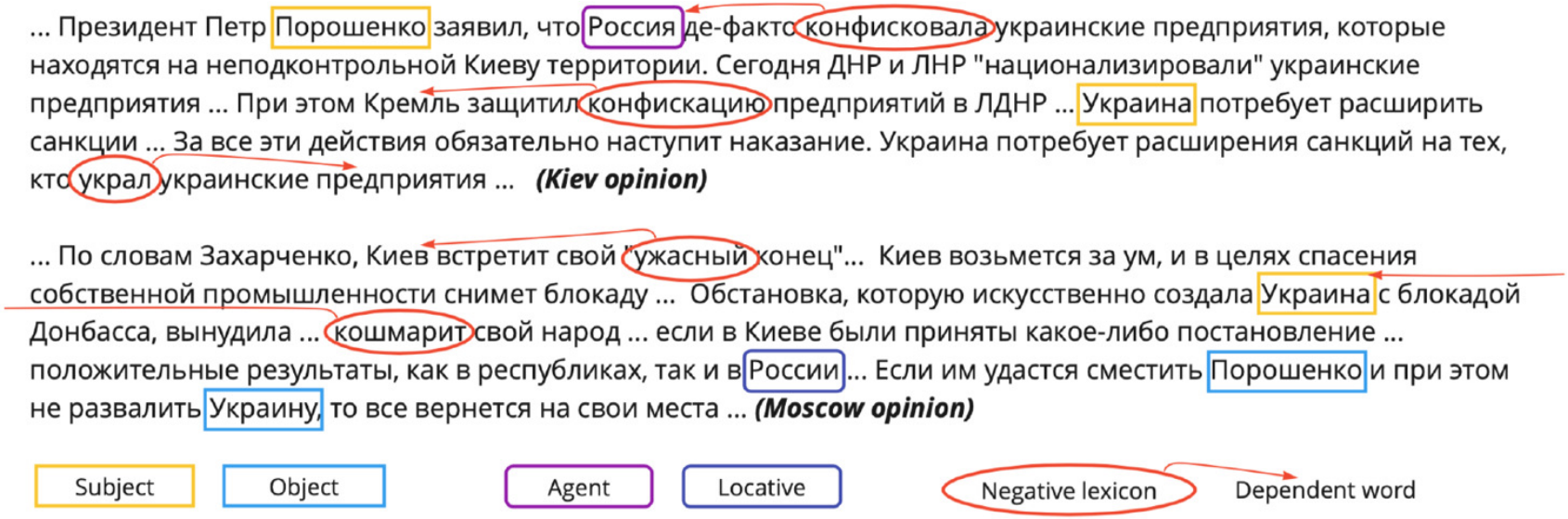
$$\mathscr{L}\left(\underset{\text{PLSA}}{\Phi\,\Theta}\right)+R\left(\underset{\text{interpretable}}{\quad}\right)+R\left(\underset{\text{multimodal}}{\quad}\right)+R\left(\underset{\text{n-gram}}{\quad}\right)+R\left(\underset{\text{syntax}}{\quad}\right)\to\max$$

Результаты:

- выделение мнений внутри тем: F1-мера = 0.86%
- совместное использование трёх модальностей:
  - SPO — факты как триплеты «субъект–предикат–объект»
  - FR — семантические роли слов по Филлмору
  - Sent — тональности именованных сущностей

D.Feldman, T.Sadekova, K.Vorontsov. Combining facts, semantic roles and sentiment lexicon in a generative model for opinion mining. Dialogue 2020.

Выделение поляризованных мнений в политических новостях



Слова «Порошенко», «Россия», «Украина» встречаются в тексте-1 и тексте-2 одинаково часто, однако:

- «Порошенко» — субъект в тексте-1 и объект в тексте-2;
- «Россия» — агенс в тексте-1 и локация в тексте-2;
- негативная тональность: «Россия», «Кремль» в тексте-1, «Киев», «Украина» в тексте-2.

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 (2022)
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

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**
   - отбор моделей, оценивание и отбор релевантных тем
   - рекомендации документов для тематических подборок

Направления исследований

- Участие в разработке «Тематизатора»
- Встраивание модуля BigARTM в Orange
- Встраивание модуля BigARTM в PolyAnalyst
- Разработка сценариев разведочного анализа текстовых данных в социо-гуманитарных исследованиях
- Разработка или адаптация средств визуализации тематических моделей в модуле «Тематизатора»
- Разработка адаптивных стратегий AutoML для оптимизации гиперпараметров тематических моделей

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