Ежегодная конференция Российской библиотечной ассоциации XXVI
Совместное заседание: Секция 08/11 по автоматизации, форматам и каталогизации23-К
Межрегиональный комитет по каталогизации

Современные методы и проблемы тематического моделирования и разведочного поиска

Воронцов Константин Вячеславович

д.ф.-м.н., профессор РАН, профессор МГУ и МФТИ,
зав. лаб. Машинного обучения и семантического анализа
Института Искусственного Интеллекта МГУ

Нижний Новгород 17 мая 2022

Содержание

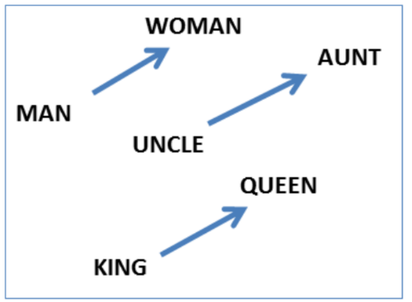
Эволюция подходов в обработке естественного языка

Декомпозиция задач по уровням «пирамиды NLP»

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

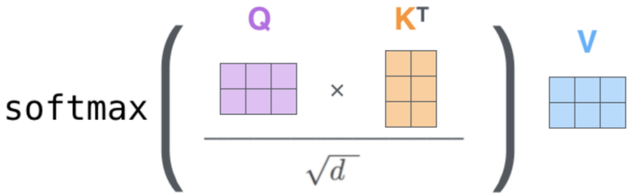
Модели векторных представлений слов (эмбедингов)

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



Нейросетевые векторные модели локальных контекстов

- рекуррентные нейронные сети: LSTM, GRU,…
- «end-to-end» модели внимания и трансформеры:
  машинный перевод, BERT [2018], GPT-3 [2020],…



Содержание

1. Эволюция подходов в обработке естественного языка
— от «стека технологий» к векторным моделям семантики

2. Проект «Мастерская знаний»
— составление и анализ тематических подборок научных статей

3. О некоторых задачах разведочного поиска
— поиск документов по документам
— многоязычный поиск и категоризация
— выявление научных трендов

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*)

Сегодня технологии IR/ML/NLP позволяют решать такие задачи

Функции «Мастерской знаний»

Подборка – долгосрочный поисковый интерес пользователя или группы

Поисково-рекомендательные функции:

- поиск тематически близких документов по подборке
- мониторинг новых документов по тематике подборки
- выявление новых научных трендов по тематике подборки
- контекстная рекомендация ссылок «см.также» в документах подборки

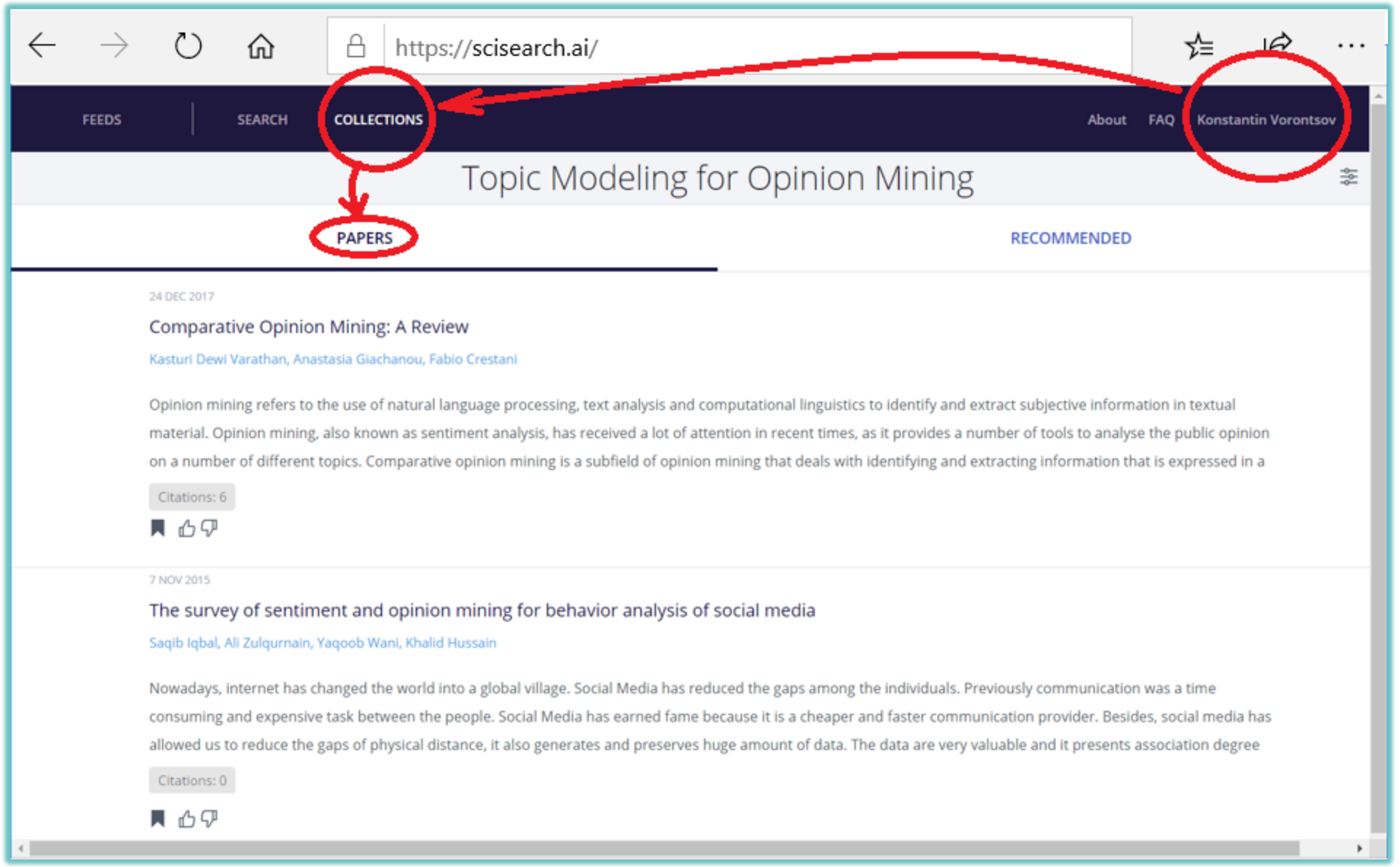
Аналитические функции:

- полуавтоматическая суммаризация подборки
- рекомендация порядка чтения документов внутри подборки
- систематизация подтем, идей, моделей, решений, мнений внутри подборки

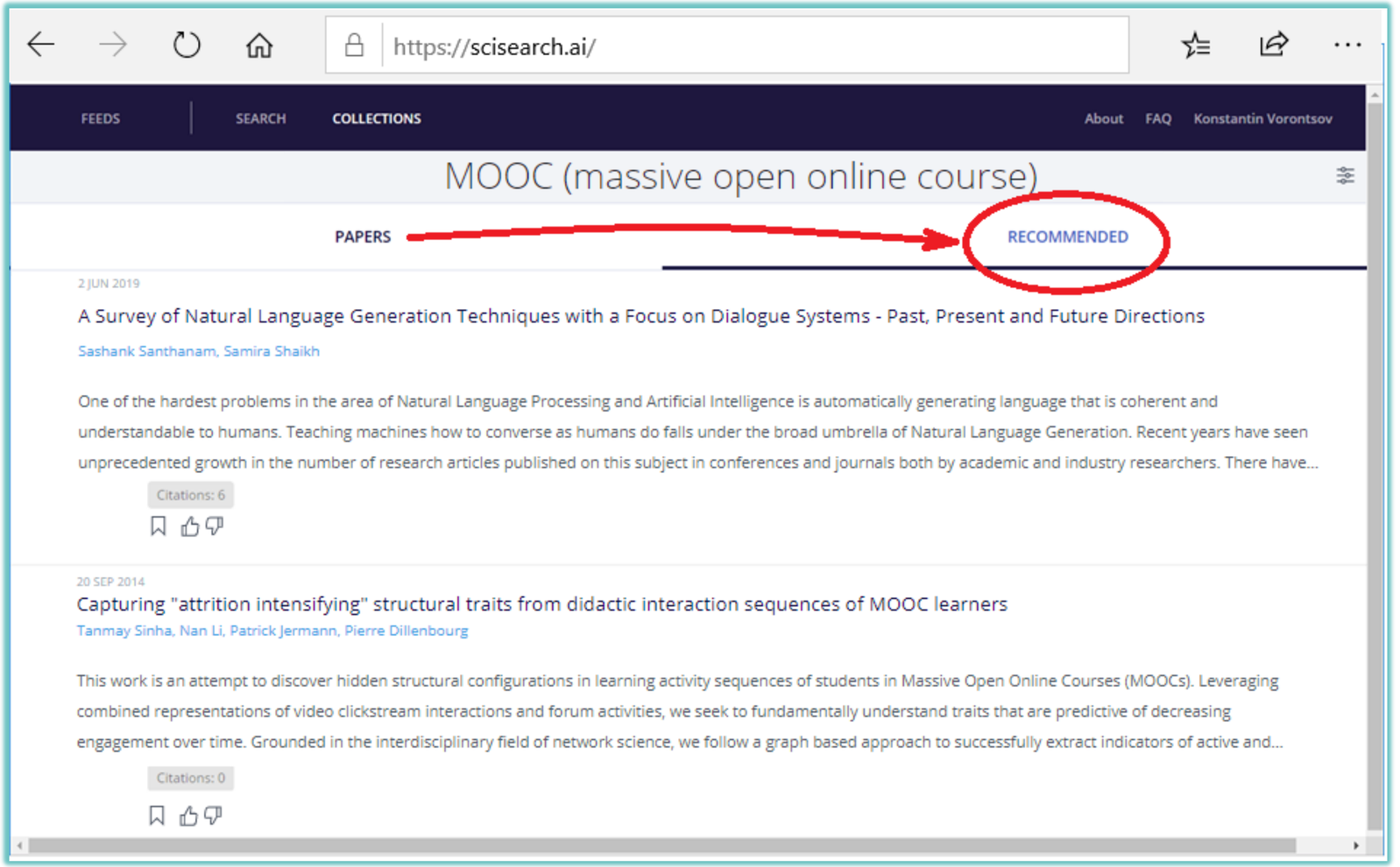
Коммуникативные функции:

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

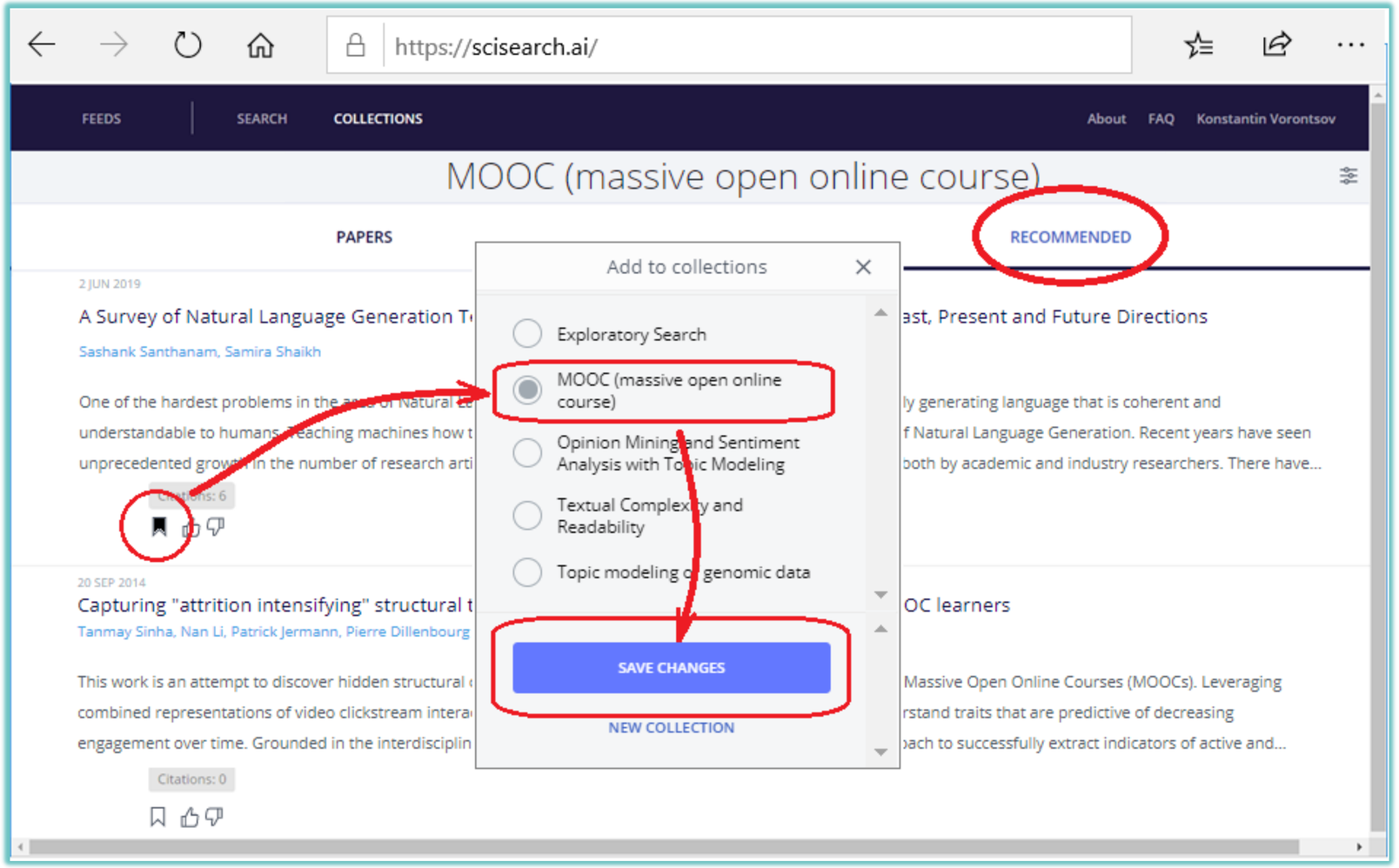
Поиск и рекомендации в SciSearch.ai



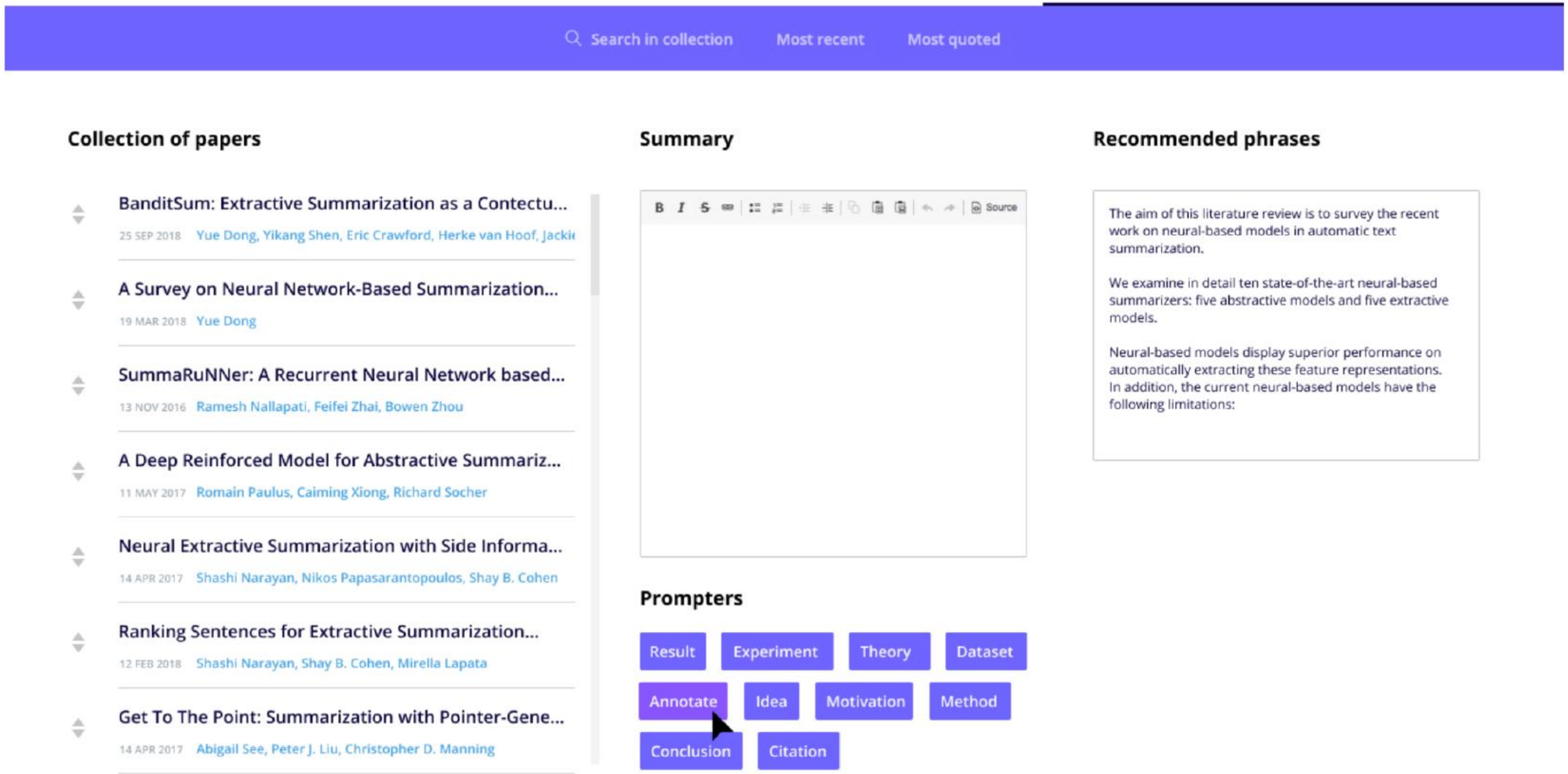
Поиск и рекомендации в SciSearch.ai



Поиск и рекомендации в SciSearch.ai



Полуавтоматическая суммаризация подборки



Андрей Власов. Методы полуавтоматической суммаризации подборок научных статей. МФТИ, 2020

Светлана Крыжановская. Технология полуавтоматической суммаризации подборок научных статей. МГУ, 2022

Полуавтоматическая суммаризация подборки

Концепция MAHS (Machine Aided Human Summarization)

1. Система рекомендует *сценарий реферата*, то есть
   в каком порядке процитировать статьи из подборки
2. Пользователь корректирует план в соответствии со своими целями
3. В цикле по ранжированным статьям подбоорки:
   - пользователь вызывает (кликает кнопку) одного из *суфлёров* по статье:
     «как другие авторы обычно ссылаются на эту статью», «основная идея статьи»,
     «метод», «достоинство», «недостаток», «результат», «вывод» и т.д.
   - система строит ранжированный список фраз
   - пользователь выбирает фразу из ранжированного списка
   - пользователь корректирует фразу и контекст в соответствии со своими целями

Андрей Власов. Методы полуавтоматической суммаризации подборок научных статей. МФТИ, 2020

Полуавтоматическая суммаризация подборки

Основные задачи машинного обучения:

- Формирование обучающей выборки: **paper → (refs, survey)**
- Ранжирование статей для сценария реферата
- Выбор релевантных фраз из текста статьи для каждого суфлёра
- Ранжирование выбранных фраз для каждого суфлёра
- Выбор релевантного контекста по данной ссылке, например:

> 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.

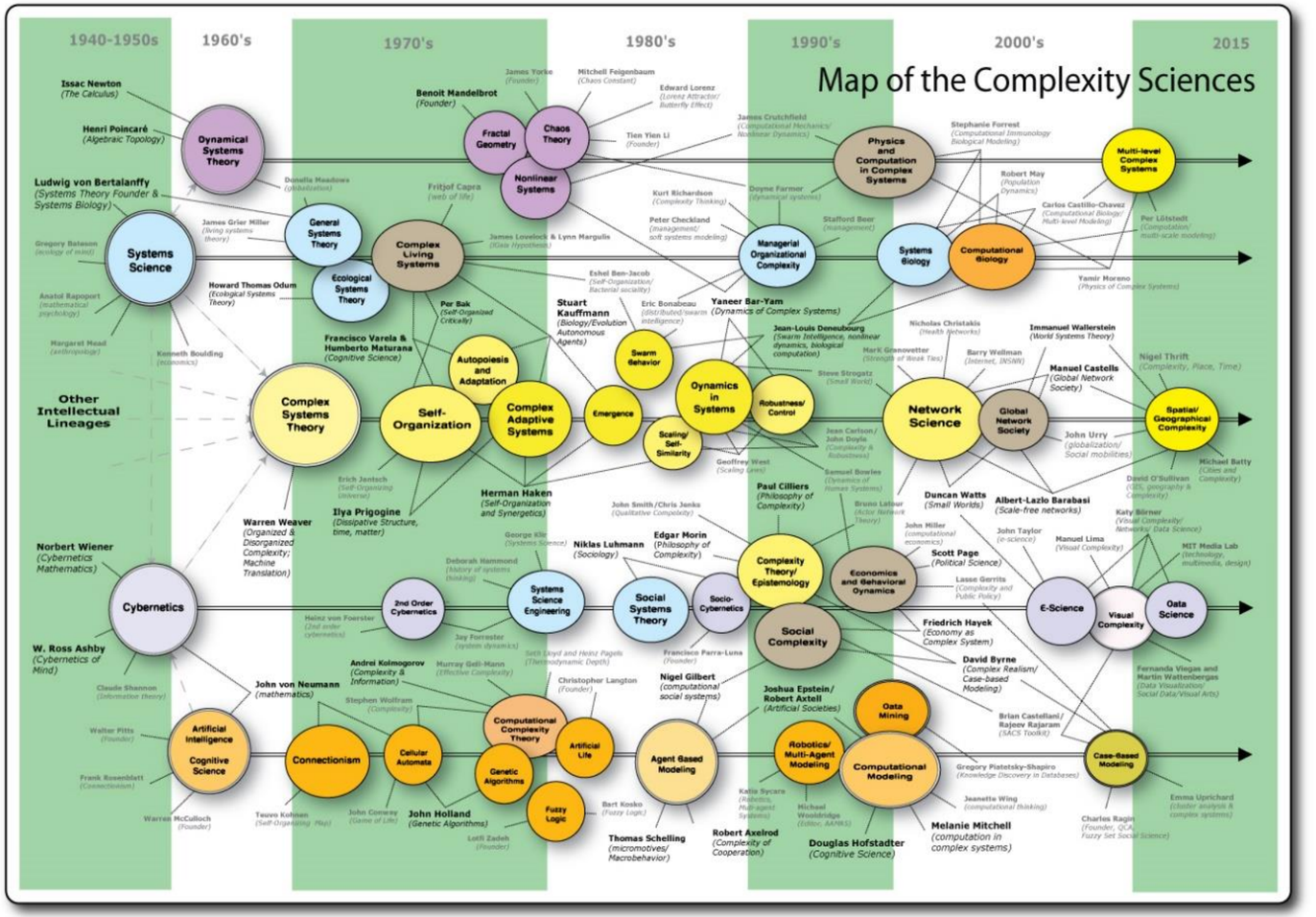
M.Yasunaga, J.Kasai, R.Zhang, A.Fabbri, I.Li, D.Friedman, D.Radev. ScisummNet: A Large Annotated Corpus and Content-Impact Models for Scientific Paper Summarization with Citation Networks. 2019.

Андрей Власов. Методы полуавтоматической суммаризации подборок научных статей. МФТИ, 2020

Визуализация и дистантное чтение (distant reading)

Осями на карте могут быть:

- **время**
- **спектр тем**
- сложность
- обзорность
- актуальность
- «хайповость»
- цитируемость



Содержание

1. Эволюция подходов в обработке естественного языка
— от «стека технологий» к векторным моделям семантики

2. Проект «Мастерская знаний»
— составление и анализ тематических подборок научных статей

3. О некоторых задачах разведочного поиска
— поиск документов по документам
— многоязычный поиск и категоризация
— выявление научных трендов

Технология тематического поиска BigARTM

Схема эксперимента:

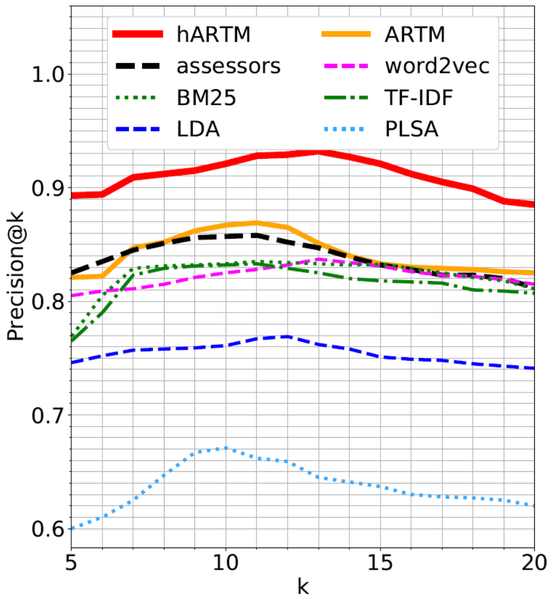
- длинные запросы (1 стр. А4)
- 100 запросов на коллекцию
- 3 асессора на каждый запрос
- от 10 до 60 минут на запрос
- разметка на Яндекс.Толока
- две коллекции техно-новостей:

(170K Russian docs)

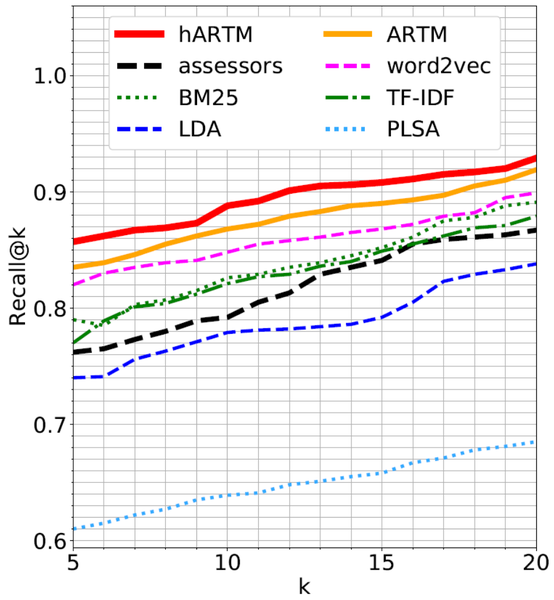
(750K English docs)

Оценки качества поиска:

точность (precision@k)



полнота (recall@k)



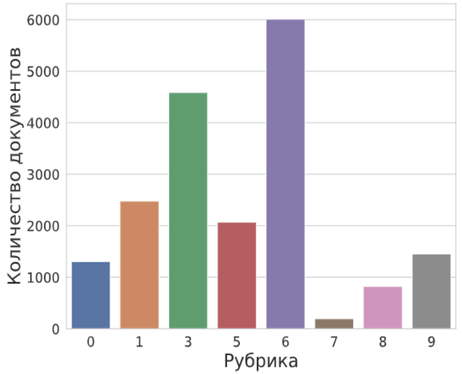
Ianina A., Golitsyn L., Vorontsov K. Multi-objective topic modeling for exploratory search in tech news. AINL 2017.

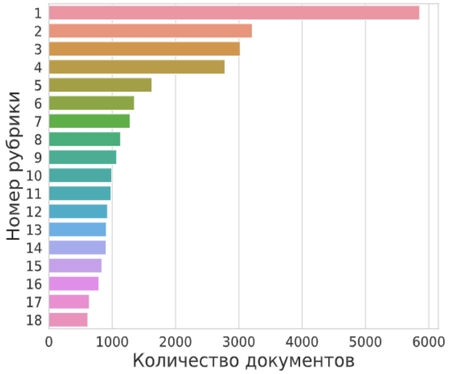
Ianina A., Vorontsov K. Regularized multimodal hierarchical topic model for document-by-document exploratory search. 2019.

Мультиязычный тематический поиск и категоризация

Данные:

- научные статьи eLibrary
  и статьи Wikipedia (100 языков)
- рубрики УДК и ГРНТИ





Две задачи, одна модель:

- тематический поиск документов по документам
- категоризация документов

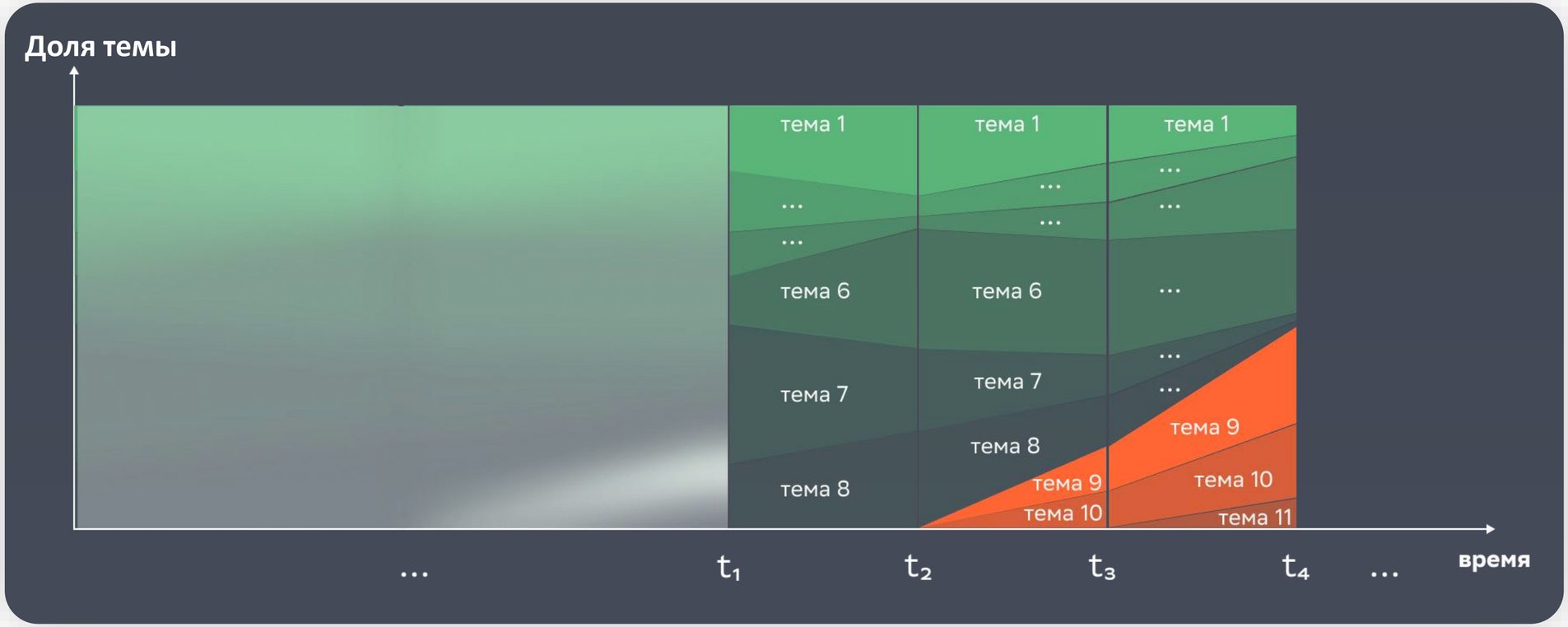
Особенности решения:

- модальности: языки, рубрики
- редукция словарей (BPE-токенизация)
  до 11 тыс. токенов на каждый язык
- сокращение модели с 128 Гб до 4.8 Гб

Название модели	Средняя частота УДК	Средний процент УДК	Средняя частота ГРНТИ	Средний процент ГРНТИ
XLM-RoBERTa	0.835	0.179	0.832	0.288
Тематическая модель	**0.995**	**0.225**	**0.852**	**0.366**

Поиск научных трендов

- *Темпоральная тематическая модель* последовательно дообучается на статьях, вышедших за 30 дней
- Удаётся детектировать >60% из 87 трендовых тем (из области Data Science), выделенных экспертами в течение года после появления темы

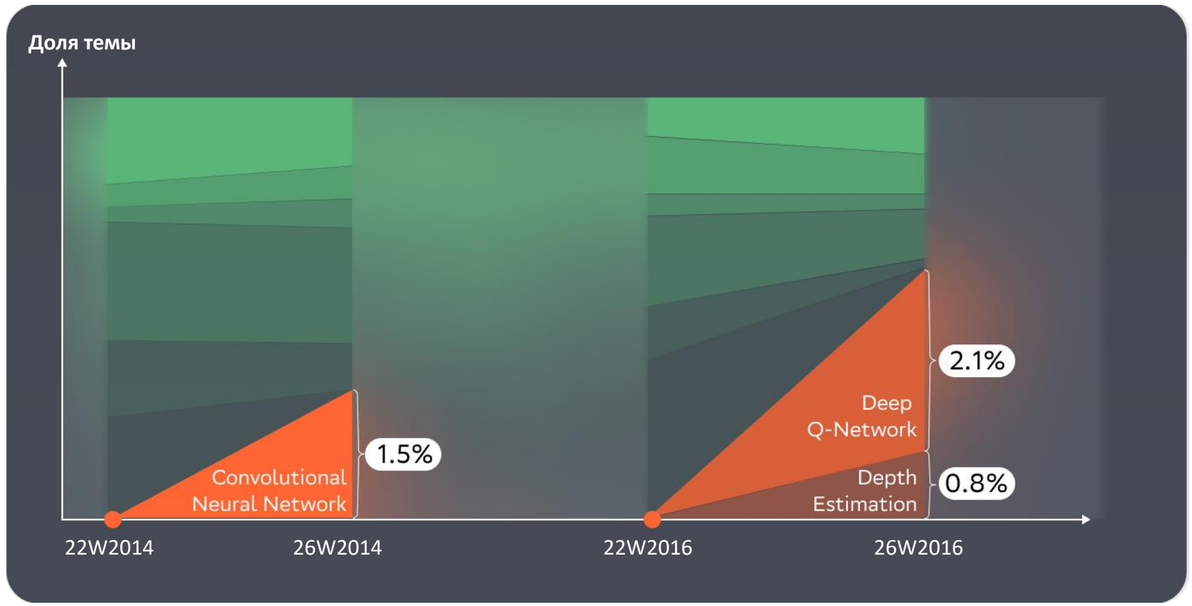


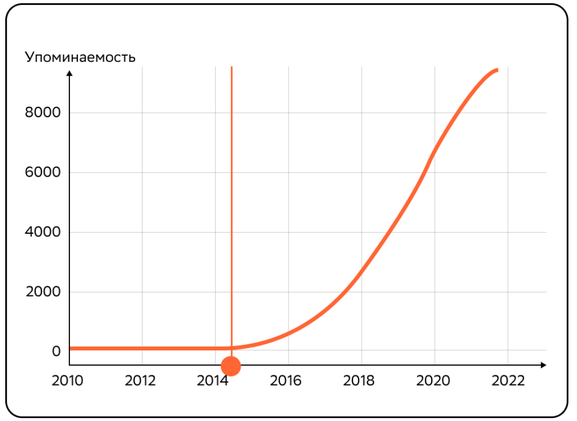
Поиск научных трендов

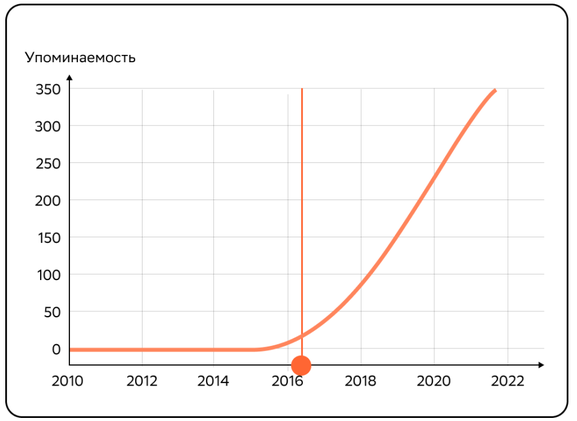
Трендовая тема:

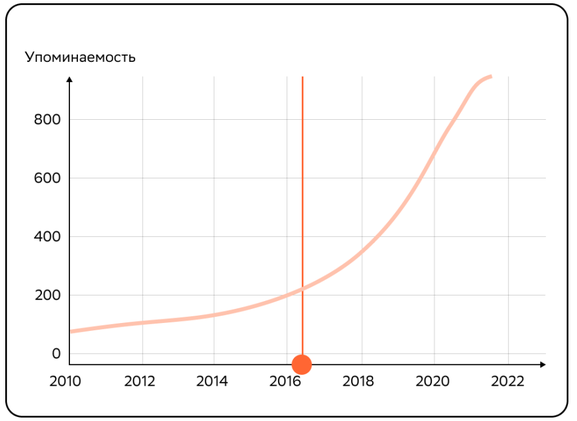
- наличие семантического ядра
- наличие аномального роста

Примеры: динамика упоминаний трендовых тем









Поиск научных трендов: примеры тем

Topic modeling	Speech recognition	Collaborative filtering	Machine translation
latent variable	prosodic feature	web page	word alignment
mixture model	speech signal	search result	target language
topic model	eye gaze	recommender system	bleu score
mixture component	audio signal	collaborative filtering	parallel corpus
Gibbs sampling	spontaneous speech	word sense	source sentence
multinomial distribution	topic segmentation	ranking model	translation model
Gibbs sampler	acoustic feature	web search	machine translation
generative process	ASR output	user preference	sentence pair
Dirichlet distribution	switchboard corpus	user profile	source language
Dirichlet process	audio data	ranking score	best list

Поиск научных трендов: примеры тем

StyleGan	Meta Learning	NERF
stylegan	meta model	neural radiance field
latent code	meta train	accurate depth estimation
mapping network	meta optimization	additional qualitative result
ablation study	meta update	novel loss function
text generation	meta testing	optical flow prediction
generation quality	training task	image reconstruction loss
generator architecture	continual learning	monocular depth prediction
mask	previous task	geometric consistency loss
encoder	catastrophic forgetting	depth estimation method
gan model	ablation study	optical flow network

Резюме

- *Цифровые технологии* (AI, ML, NLP, NLU) готовы для автоматизации широкого спектра задач «Мастерской знаний»

- Основное достоинство тематических моделей — покоординатная интерпретируемость векторных представлений текста

- Механизмы автоматического выделения терминов, регуляризации, модальностей позволяют существенно улучшать качество поиска, категоризации, выделения трендов

Спасибо за внимание!

Воронцов Константин Вячеславович
д.ф.-м.н., проф. РАН,
зав. лаб. Машинного обучения и семантического анализа
Института Искусственного Интеллекта МГУ

k.v.vorontsov @ phystech.edu

http://www.MachineLearning.ru/wiki?title=User:Vokov