

Ежегодная конференция Российской библиотечной ассоциации XXVI
Совместное заседание: Секция 08/11 по автоматизации, форматам и каталогизации 23-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

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

Эволюция подходов в обработке естественного языка

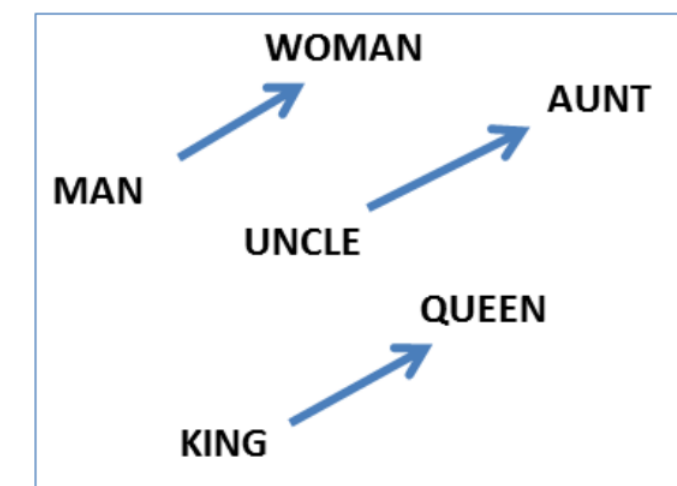
Декомпозиция задач по уровням «пирамиды NLP»

- 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],...

$$\text{softmax} \left(\frac{\begin{matrix} \mathbf{Q} \\ \text{grid} \end{matrix} \times \begin{matrix} \mathbf{K}^T \\ \text{grid} \end{matrix}}{\sqrt{d}} \right) \mathbf{V}$$

The diagram shows a matrix multiplication of a query matrix Q (purple grid) and a key matrix K^T (orange grid), divided by the square root of the dimension d. The result is passed through a softmax function to produce a value vector V (blue grid).

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**

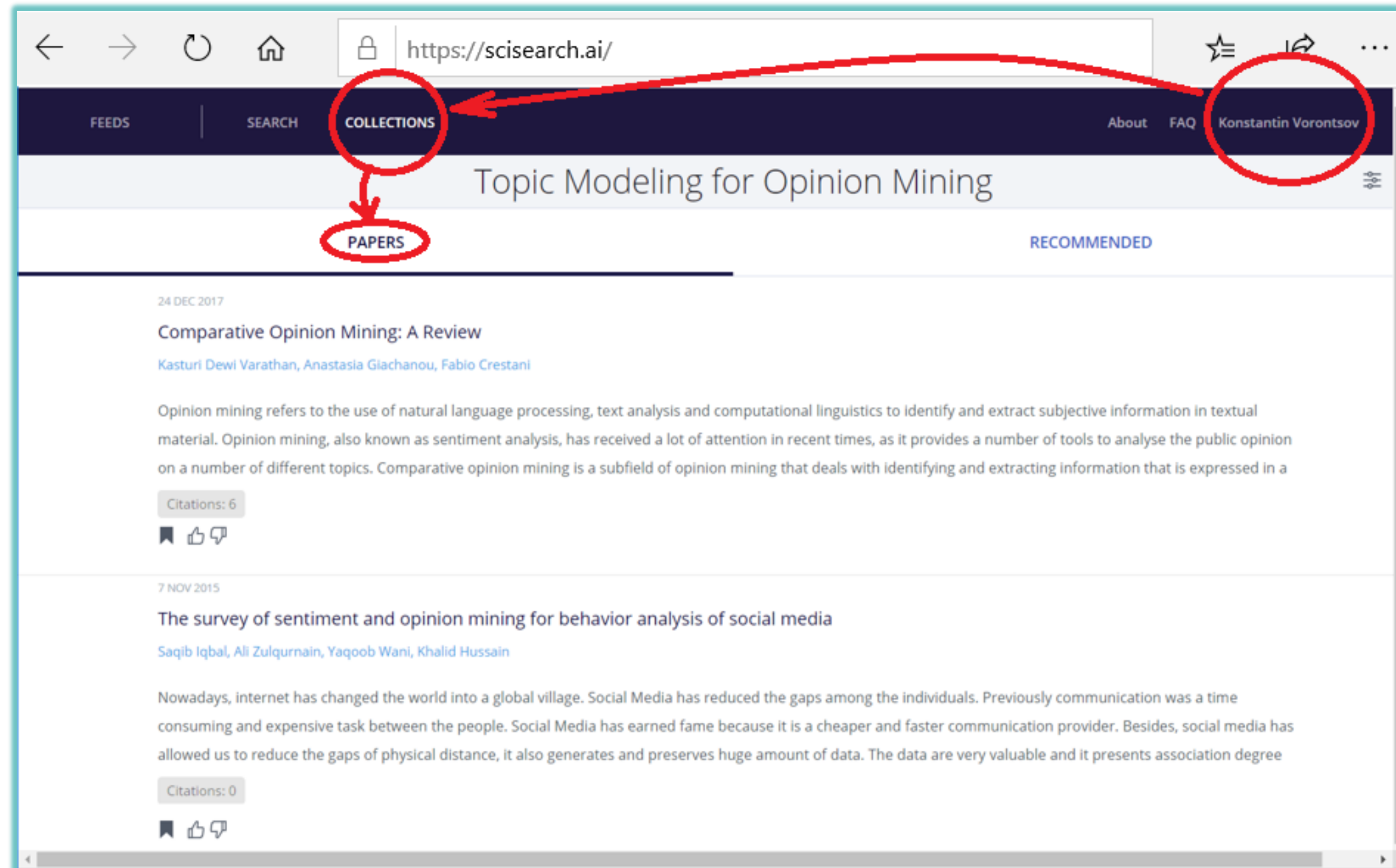
Аналитические функции:

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

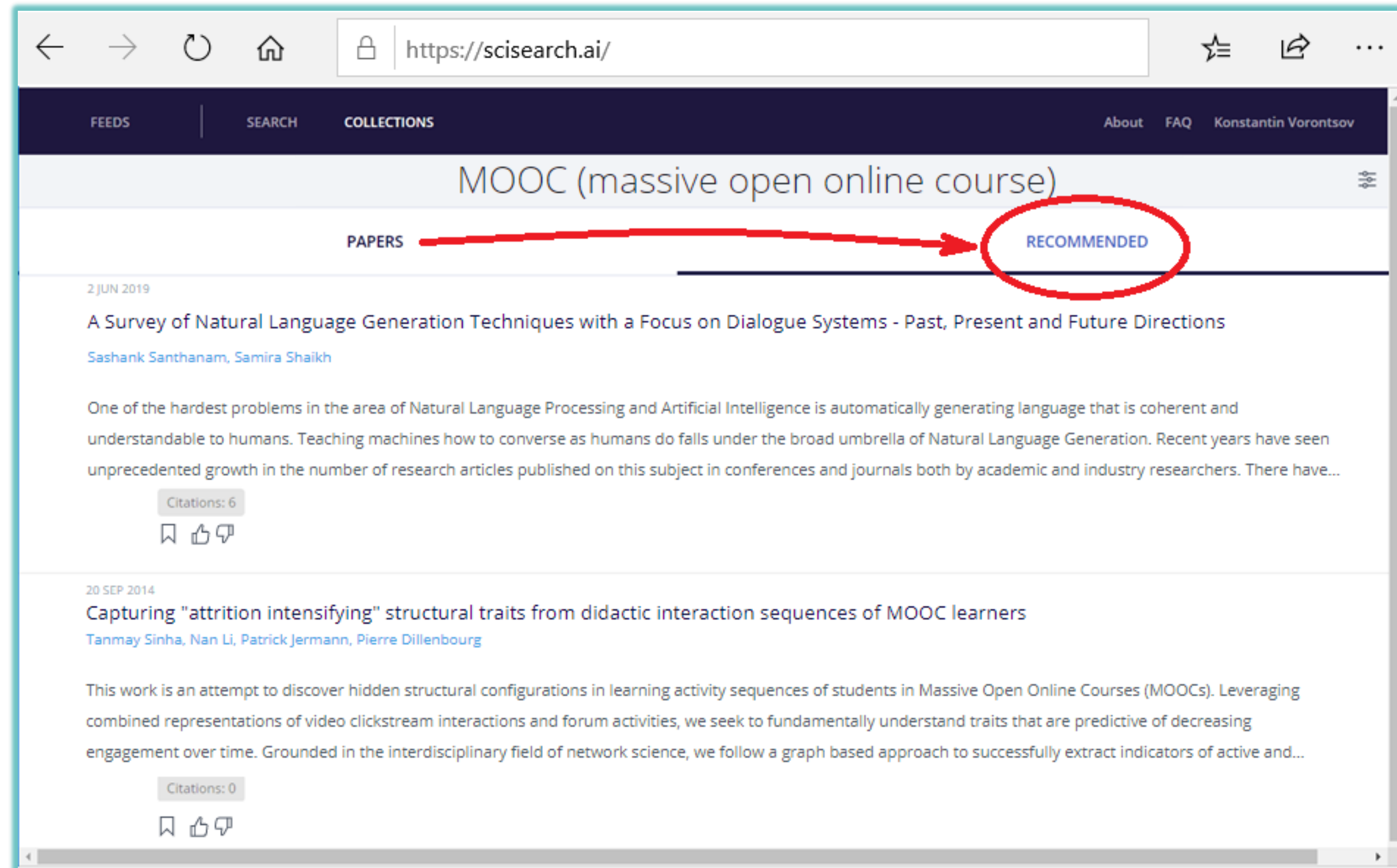
Коммуникативные функции:

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

Поиск и рекомендации в SciSearch.ai



Поиск и рекомендации в SciSearch.ai



The screenshot shows the SciSearch.ai website interface. The browser address bar displays <https://scisearch.ai/>. The navigation menu includes FEEDS, SEARCH, and COLLECTIONS. The search results are for the query "MOOC (massive open online course)". Two tabs are visible: "PAPERS" and "RECOMMENDED". The "RECOMMENDED" tab is circled in red, and a red arrow points from the "PAPERS" tab to it. Below the tabs, two search results are displayed:

2 JUN 2019
A Survey of Natural Language Generation Techniques with a Focus on Dialogue Systems - Past, Present and Future Directions
Sashank Santhanam, Samira Shaikh
One of the hardest problems in the area of Natural Language Processing and Artificial Intelligence is automatically generating language that is coherent and understandable to humans. Teaching machines how to converse as humans do falls under the broad umbrella of Natural Language Generation. Recent years have seen unprecedented growth in the number of research articles published on this subject in conferences and journals both by academic and industry researchers. There have...
Citations: 6
Bookmark, Like, Dislike icons

20 SEP 2014
Capturing "attrition intensifying" structural traits from didactic interaction sequences of MOOC learners
Tanmay Sinha, Nan Li, Patrick Jermann, Pierre Dillenbourg
This work is an attempt to discover hidden structural configurations in learning activity sequences of students in Massive Open Online Courses (MOOCs). Leveraging combined representations of video clickstream interactions and forum activities, we seek to fundamentally understand traits that are predictive of decreasing engagement over time. Grounded in the interdisciplinary field of network science, we follow a graph based approach to successfully extract indicators of active and...
Citations: 0
Bookmark, Like, Dislike icons

Поиск и рекомендации в SciSearch.ai

The screenshot displays the SciSearch.ai website interface. The browser address bar shows the URL <https://scisearch.ai/>. The navigation menu includes FEEDS, SEARCH, and COLLECTIONS. The main content area features a search result for "MOOC (massive open online course)" by Sashank Santhanam and Samira Shaikh, dated 2 JUN 2019. The article title is "A Survey of Natural Language Generation T...". The abstract begins with "One of the hardest problems in the area of Natural Language Generation is to generate language that is coherent and understandable to humans. Teaching machines how to generate language that is coherent and understandable to humans is a challenging task. In recent years, there has been an unprecedented growth in the number of research articles on this topic." The article has 6 citations. A red circle highlights the citation count, and a red arrow points from it to the "Add to collections" dialog box. The dialog box is titled "Add to collections" and contains a list of collection categories: Exploratory Search, MOOC (massive open online course), Opinion Mining and Sentiment Analysis with Topic Modeling, Textual Complexity and Readability, and Topic modeling of genomic data. The "MOOC (massive open online course)" option is selected. A red circle highlights this option, and a red arrow points from it to the "SAVE CHANGES" button. Below the list is a "NEW COLLECTION" link. The background shows a "RECOMMENDED" section with a red circle around the word "RECOMMENDED".

Полуавтоматическая суммаризация подборки

The interface features a blue header bar with a search icon and the text "Search in collection", and two buttons labeled "Most recent" and "Most quoted".

Collection of papers

- BanditSum: Extractive Summarization as a Contextu...
25 SEP 2018 Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, Jackie
- A Survey on Neural Network-Based Summarization...
19 MAR 2018 Yue Dong
- SummaRuNNer: A Recurrent Neural Network based...
13 NOV 2016 Ramesh Nallapati, Feifei Zhai, Bowen Zhou
- A Deep Reinforced Model for Abstractive Summariz...
11 MAY 2017 Romain Paulus, Caiming Xiong, Richard Socher
- Neural Extractive Summarization with Side Informa...
14 APR 2017 Shashi Narayan, Nikos Papasarantopoulos, Shay B. Cohen
- Ranking Sentences for Extractive Summarization...
12 FEB 2018 Shashi Narayan, Shay B. Cohen, Mirella Lapata
- Get To The Point: Summarization with Pointer-Gener...
14 APR 2017 Abigail See, Peter J. Liu, Christopher D. Manning

Summary

A rich text editor with a toolbar containing icons for Bold (B), Italic (I), Strikethrough (ABC), Bulleted List, Numbered List, Indent, Outdent, Undo, Redo, and Source. The editor area is currently empty.

Recommended phrases

The aim of this literature review is to survey the recent work on neural-based models in automatic text summarization.

We examine in detail ten state-of-the-art neural-based summarizers: five abstractive models and five extractive models.

Neural-based models display superior performance on automatically extracting these feature representations. In addition, the current neural-based models have the following limitations:

Prompters

Result Experiment Theory Dataset

Annotate Idea Motivation Method

Conclusion Citation

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 подборки:
 - 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

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.

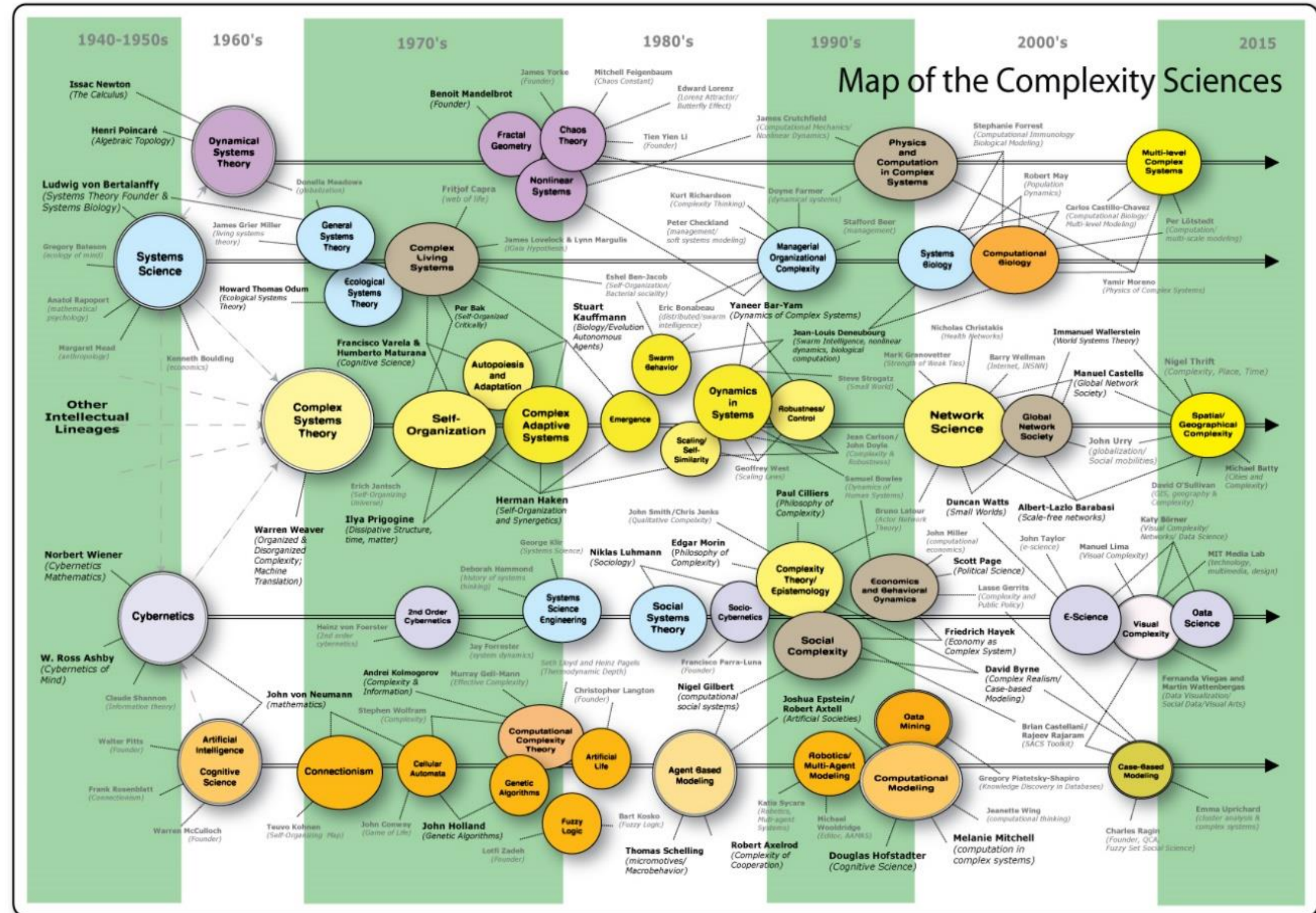
M.Yasunaga, J.Kasai, R.Zhang, A.Fabbri, I.Li, D.Friedman, D.Radev. ScisummNet: A Large Annotated Corpus and Content-Impact Models for Scientific Paper Summarization with Citation Networks. 2019.

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 VigARTM

Схема эксперимента:

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



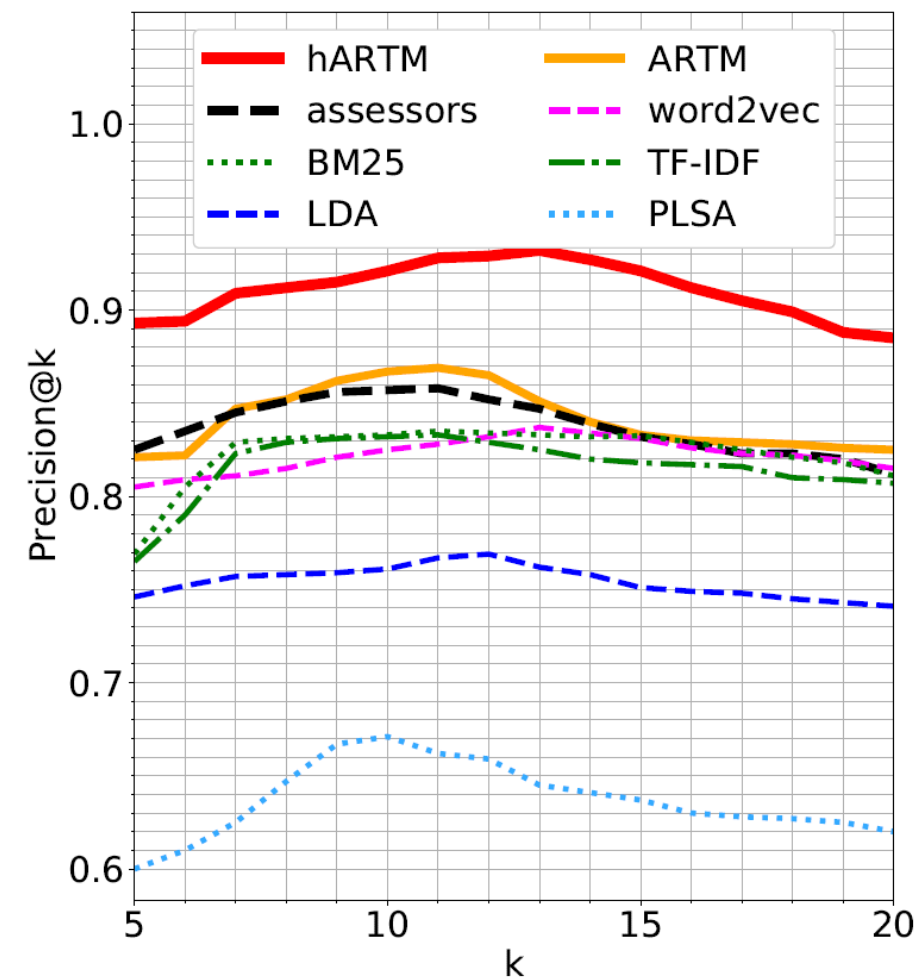
(170K Russian docs)



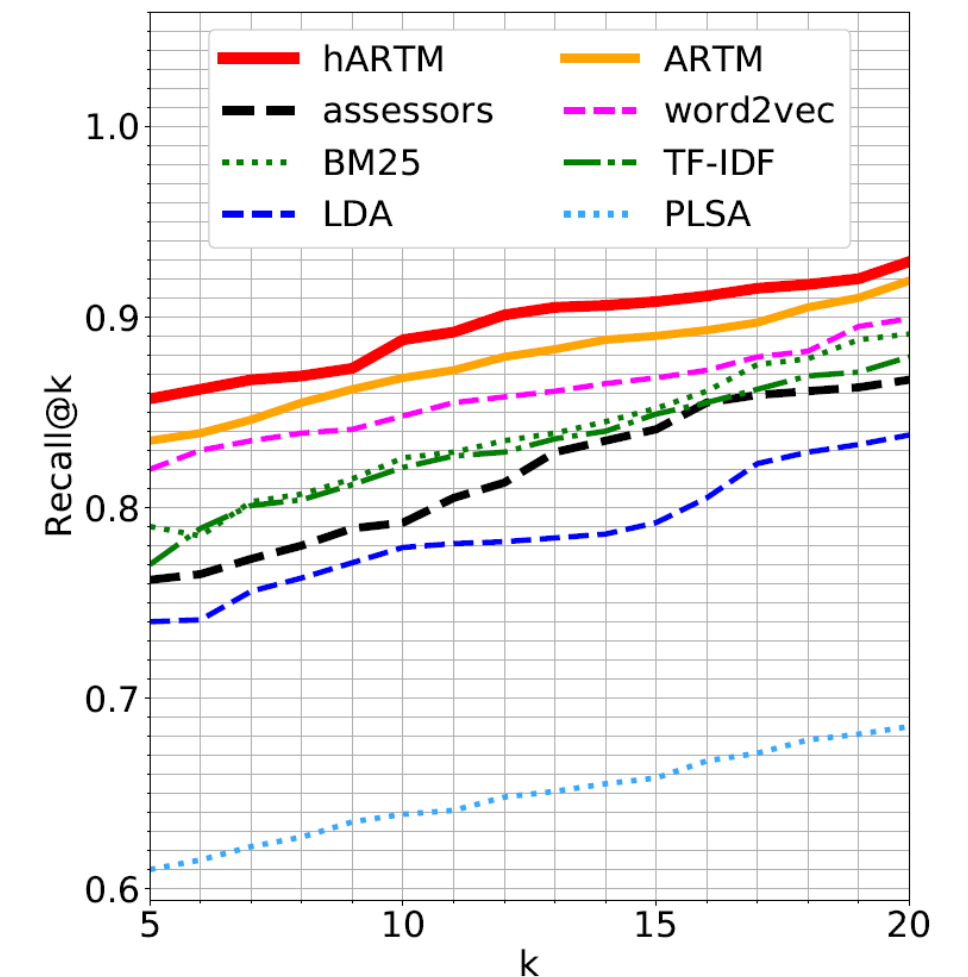
(750K English docs)

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:

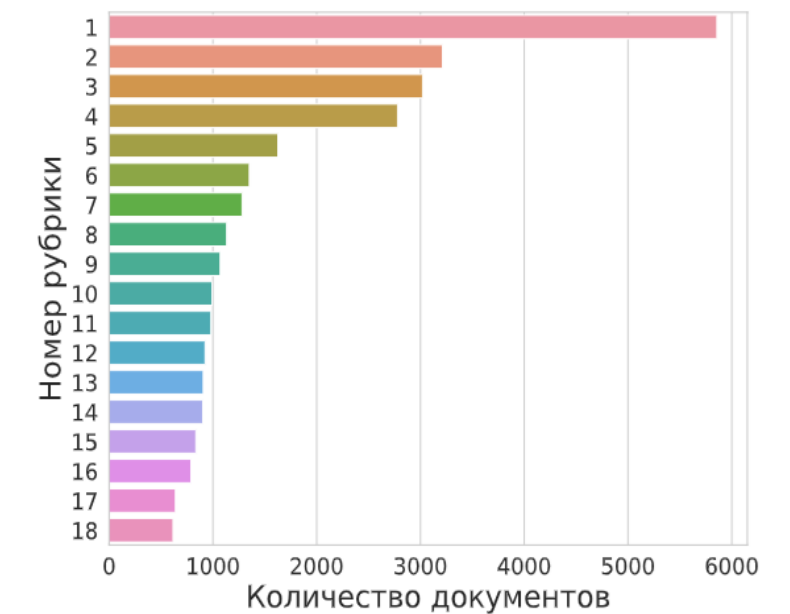
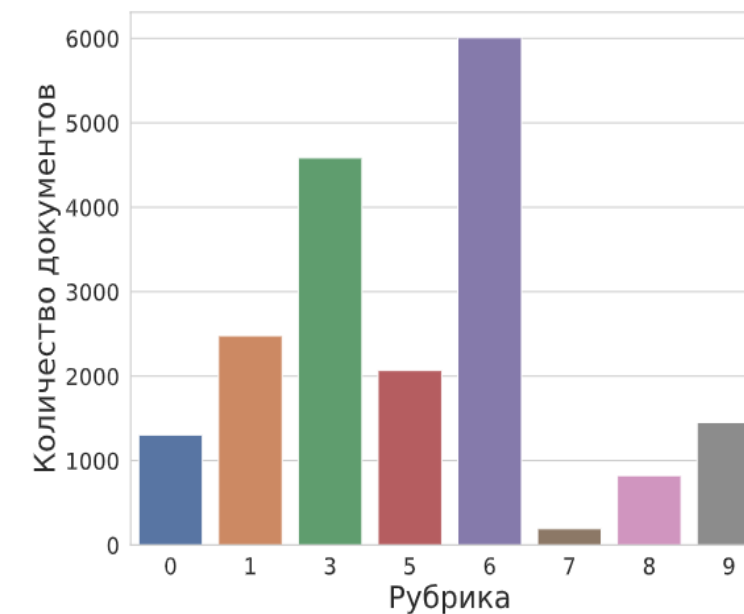
- научные статьи eLibrary и статьи Wikipedia (100 языков)
- рубрики УДК и ГРНТИ

Две задачи, одна модель:

- тематический поиск документов по документам
- категоризация документов

Особенности решения:

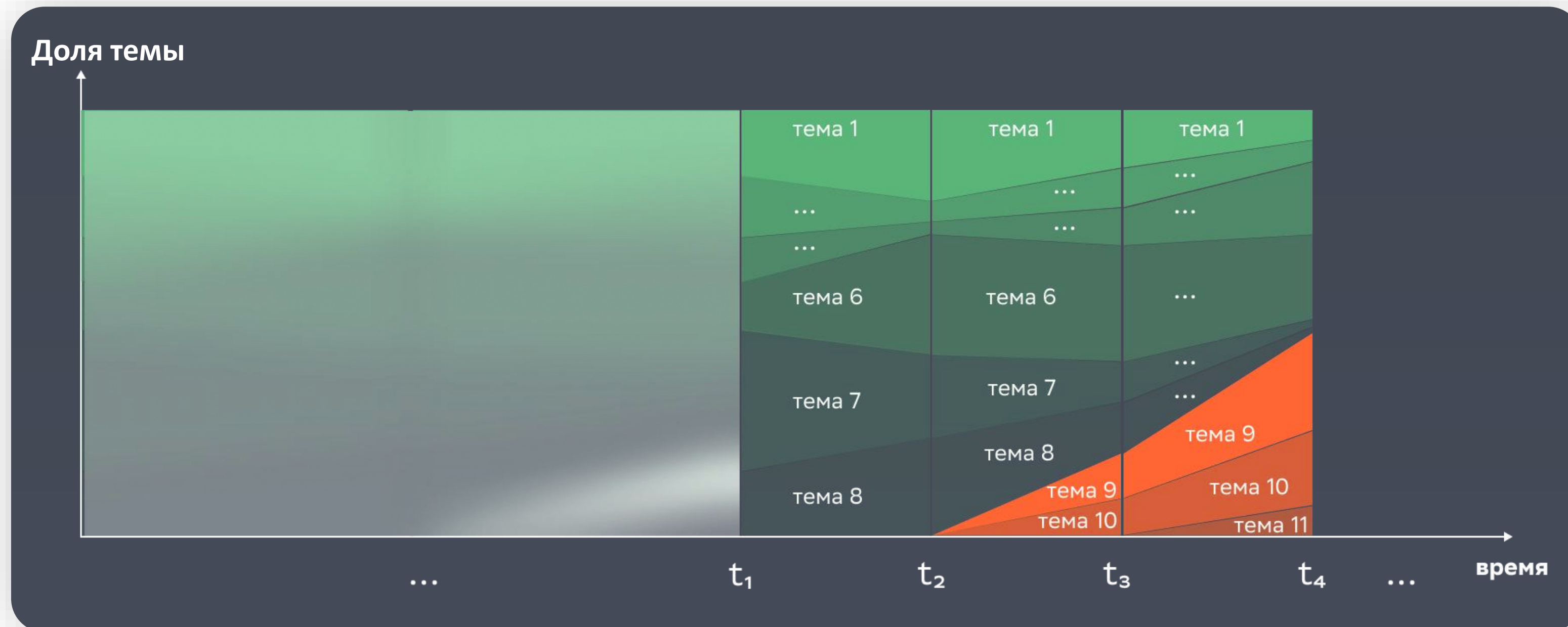
- модальности: языки, рубрики
- редукция словарей (VPE-токенизация) до 11 тыс. токенов на каждый язык
- сокращение модели с 128 Гб до 4.8 Гб



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

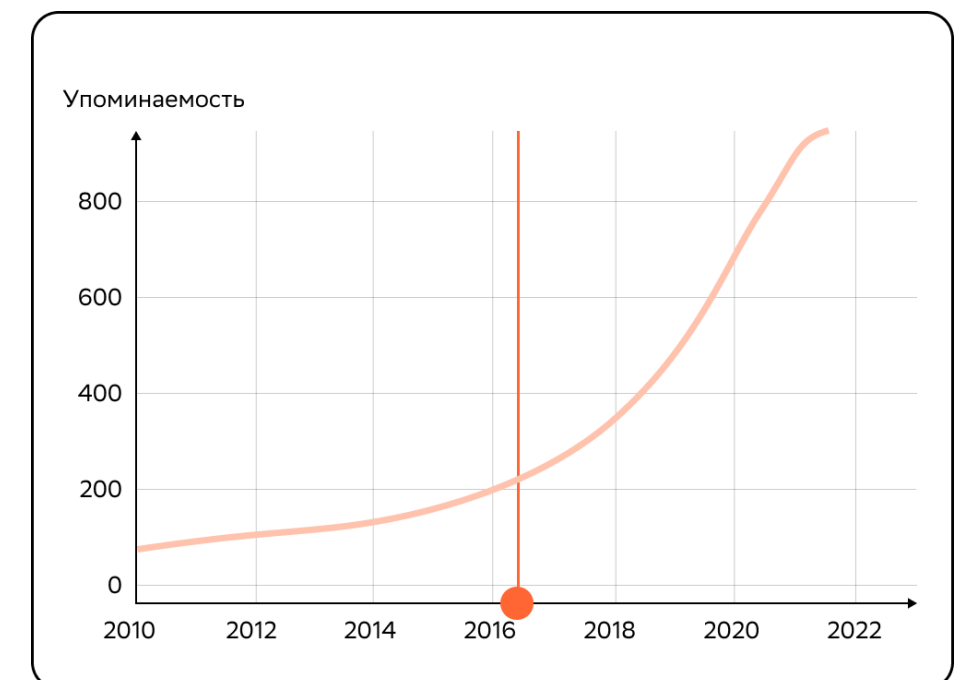
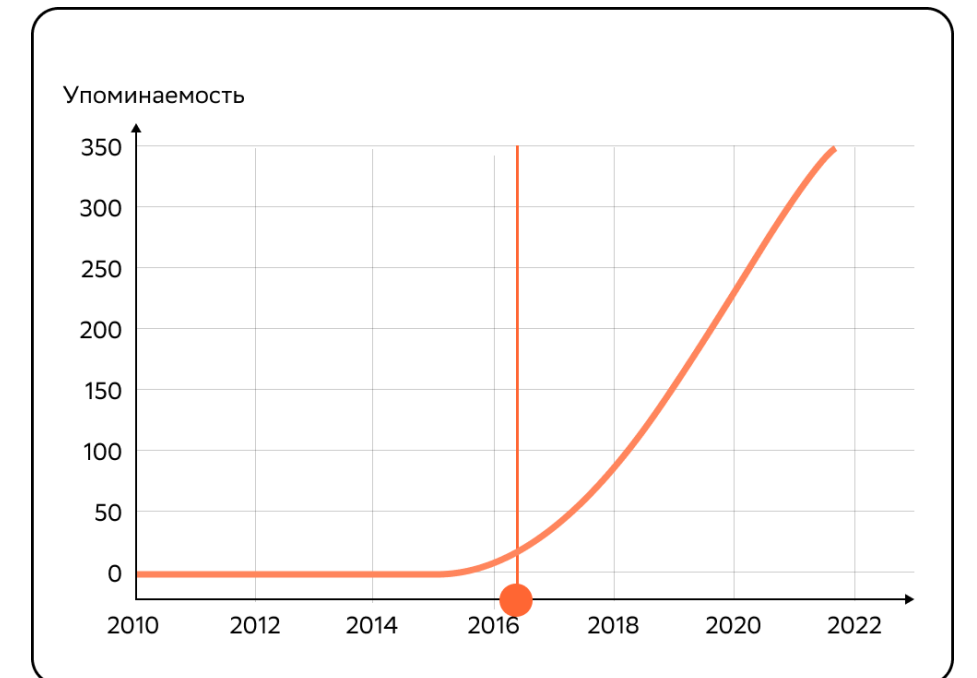
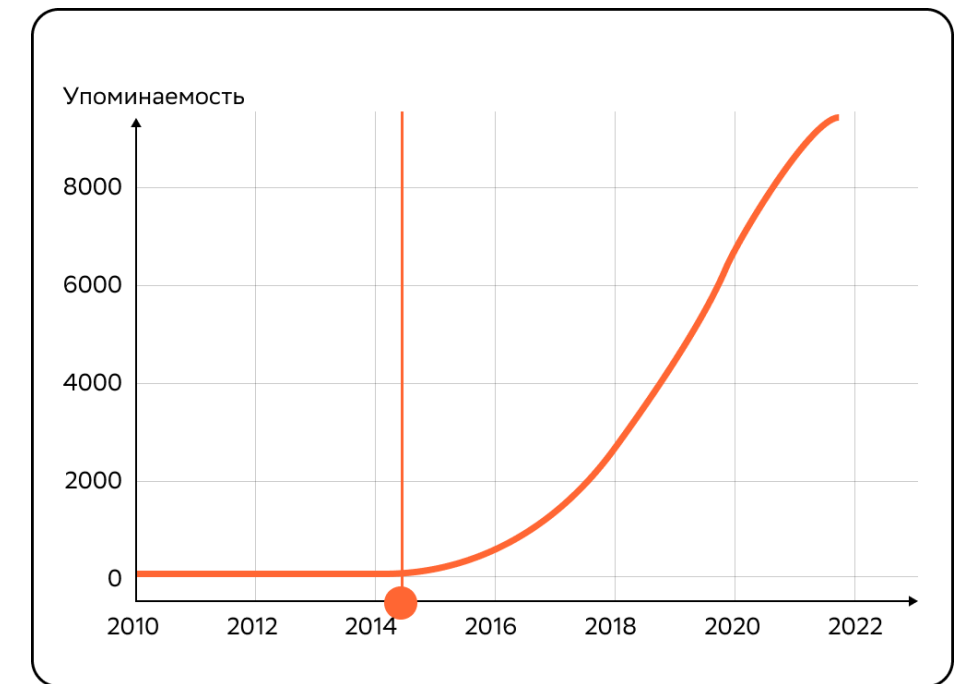
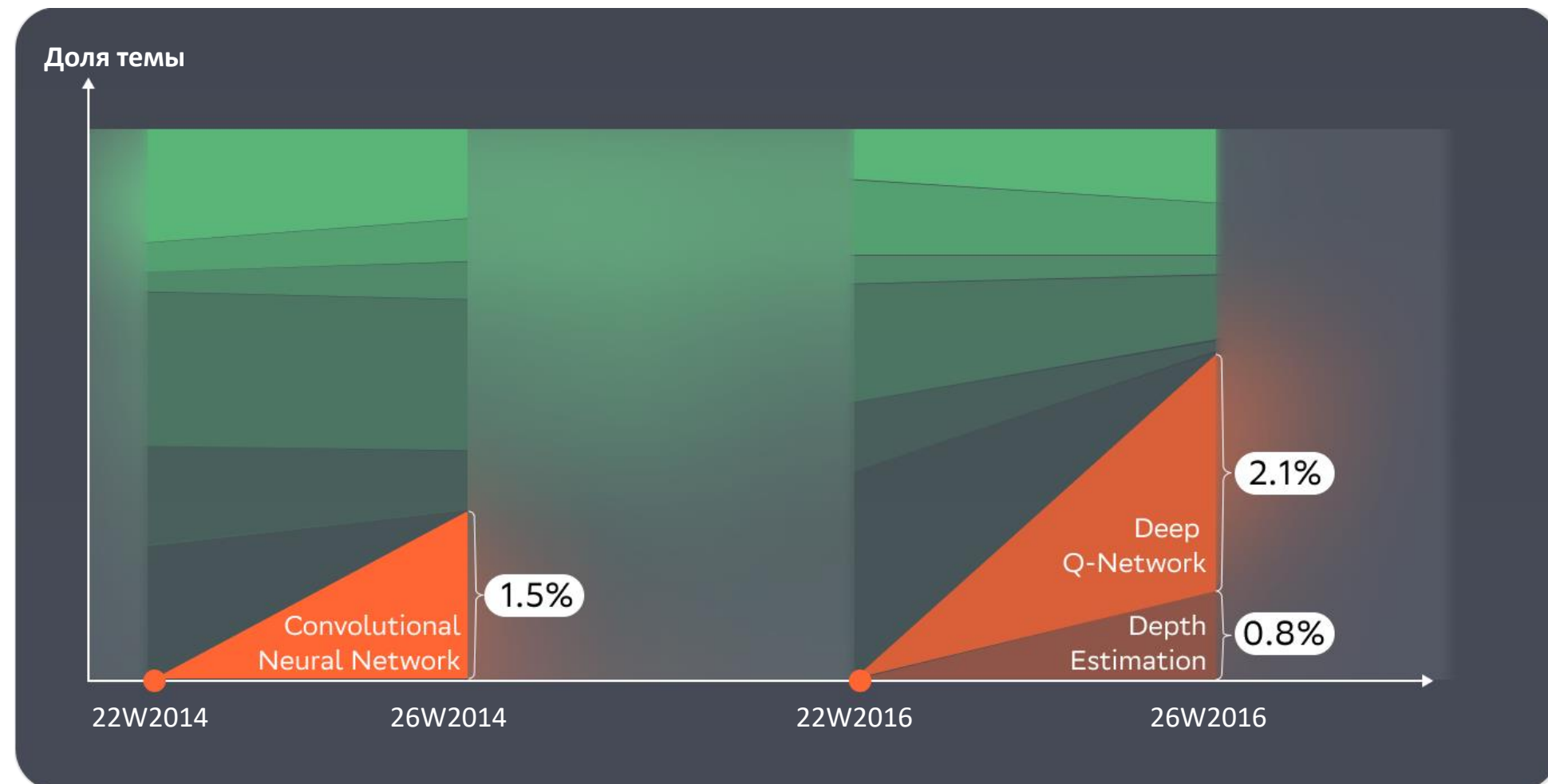


Поиск научных трендов

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

stylegan

latent code

mapping network

ablation study

text generation

generation quality

generator architecture

mask

encoder

gan model

Meta Learning

meta model

meta train

meta optimization

meta update

meta testing

training task

continual learning

previous task

catastrophic forgetting

ablation study

NERF

neural radiance field

accurate depth estimation

additional qualitative result

novel loss function

optical flow prediction

image reconstruction loss

monocular depth prediction

geometric consistency loss

depth estimation method

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 улучшить 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](mailto:k.v.vorontsov@phystech.edu)

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