

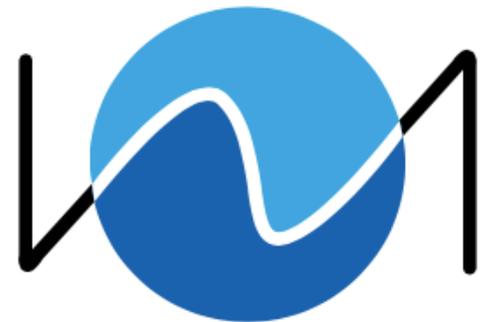
Интеллектуализация обработки информации, г.Гаэта

# Тематический информационный поиск

Анастасия Янина  
Константин Воронцов



09.10.2018



# Поиск близких текстов

## Для чего он нужен?

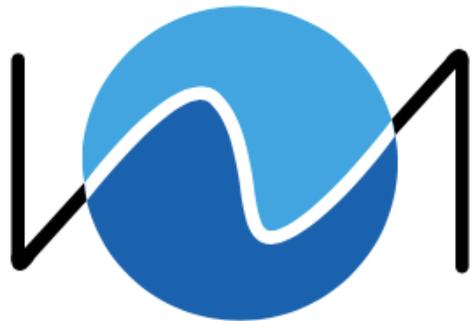
- Рекомендательные системы
- Ранжирование
- Поиск в интернете

## Актуальные подходы:

- Word embeddings: word2vec, glove, fasttext
- Нейросети: DSSM, LSTM, RNN, Attention
- Тематические модели: pLSA, LDA, ARTM

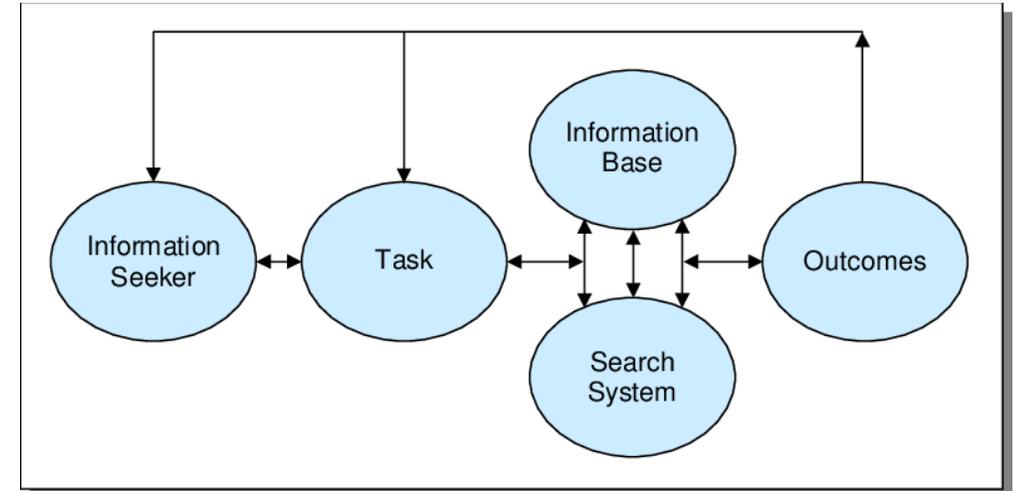
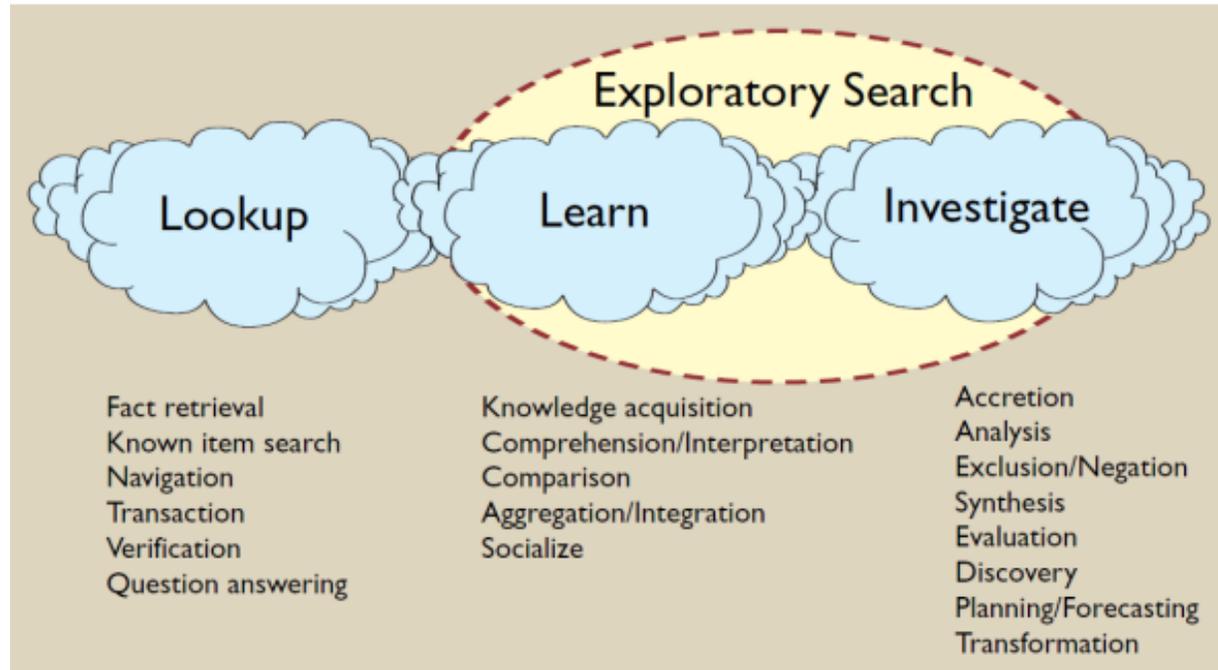


# Разведочный ПОИСК



Чем разведочный поиск  
отличается от поиска близких  
текстов и документного  
поиска?

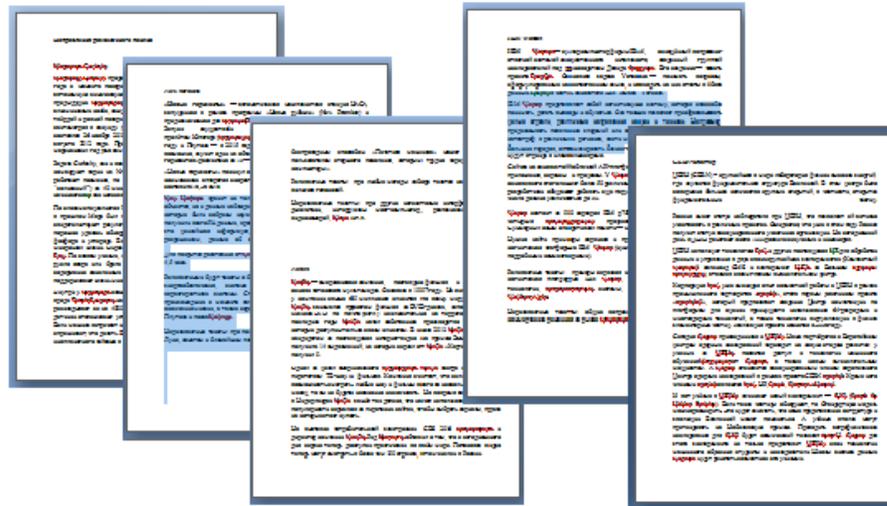
# Разведочный поиск



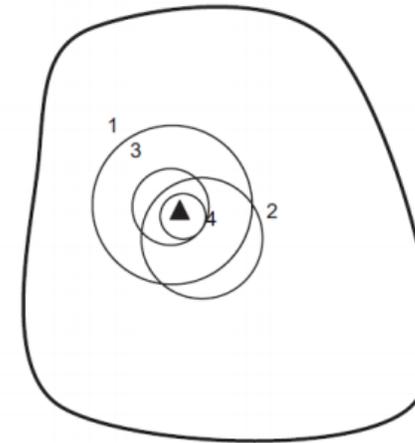
Gary Marchionini. Exploratory Search: from finding to understanding. Communications of the ACM. 2006, 49(4), p. 41–46.

R.W.White, R.A.Roth. Exploratory Search: beyond the Query-Response paradigm. San Rafael, CA: Morgan and Claypool, 2009.

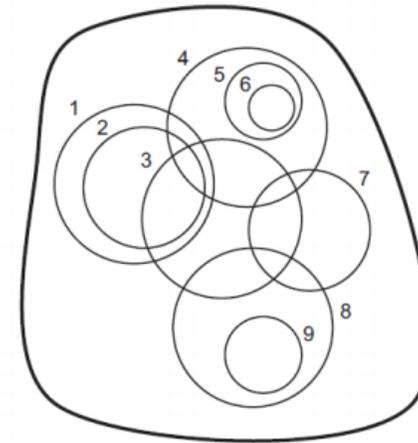
# Разведочный поиск



Iterative Search



Exploratory Search



▲ Search target    ◻ Information space  
○ Result sets (larger = more results, intersection = overlap, # = iteration)

Запросы для разведочного поиска

**Запрос:** 1-2 страницный документ с описанием поисковой задачи

**Результат поиска:** набор релевантных документов, после ознакомления с которыми у пользователя должна сложиться «карта» предметной области, сформироваться понимание основной терминологии и понятий.

# Тематический разведочный поиск

- Коллекция документов  $D$
- Запрос:  $q = (w_1, \dots, w_{n_q})$
- Тематический вектор запроса:  $\theta_{tq} = p(t|q)$
- Тематический вектор документа:  $\theta_{td} = p(t|d)$



...

# Мультимодальная тематическая модель

- Коллекция документов  $D$
- Набор тем  $T$
- Набор модальностей  $M$
- $W^1, \dots, W^m$  - словари для каждой модальности
- Примеры модальностей: слова, авторы, теги, категории, ссылки.

Матрица терминов в темах для каждой модальности:

$$\Phi_m = (\phi_{wt}^m)_{W^m \times T} \quad \phi_{wt}^m = p(w|t) \quad \forall m \in M$$

Матрица тем в документах:

$$\Theta = (\theta_{td})_{T \times D}, \quad \theta_{td} = p(t|d)$$

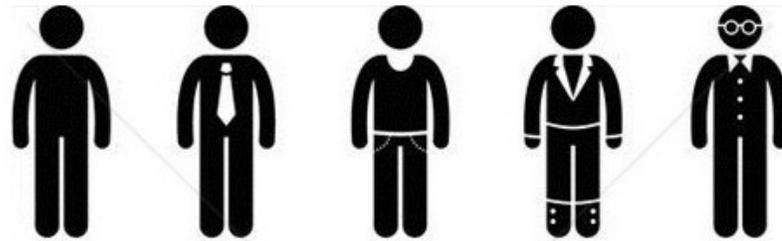
# Мультимодальная тематическая модель

Максимизируем сумму логарифма правдоподобия модели и взвешенной суммы регуляризаторов:

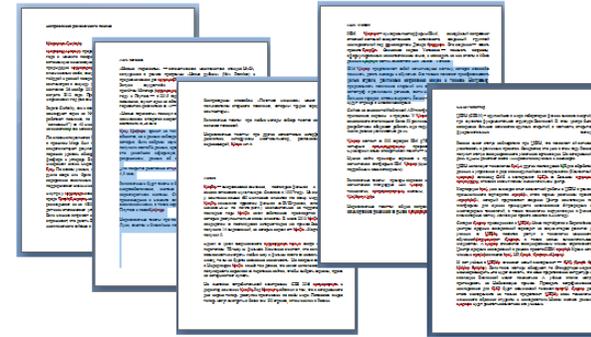
$$\sum_{m \in M} \lambda_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

$$\begin{cases} \text{E-step:} & p_{tdw} = \mathop{\text{norm}}_{t \in T}(\phi_{wt} \theta_{td}) \\ \text{M-step:} & \begin{cases} \phi_{wt} = \mathop{\text{norm}}_{w \in W^m} \left( \sum_{d \in D} \lambda_{m(w)} n_{dw} p_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right) \\ \theta_{td} = \mathop{\text{norm}}_{t \in T} \left( \sum_{w \in d} \lambda_{m(w)} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) \end{cases} \end{cases}$$

# Оценка качества разведочного поиска



Асессоры



Запросы

Два задания для асессоров:

- 1) Найти как можно больше релевантных документов, пользуясь любыми поисковыми средствами (Google/Yandex, поиск по категориям, теги)
- 2) Разметить SERP'ы нашего поисковика по тем же запросам

# Датасеты



## ХабраХабр

- 175143 статей
- 5 модальностей: 10552 слов, 742000 биграмм, 524 авторов, 10000 комментаторов, 2546 тегов, 123 категории



## TechCrunch

- 759324 статей
- 5 модальностей: 11523 слов, 1.2 млн.биграмм, 605 авторов и 184 категории

# Примеры запросов для разведочного поиска

Algorithms for coloring graphs

Netflix

Techniques for fast typing

Elon Musk space projects

Hadoop MapReduce

Self-driving Google car

Public-key cryptography

Platforms for online education

Data Science Meetups in Moscow

Educational projects mail.ru

Interplanetary station New horizons

Word2vec

Watson is a question answering computer system capable of answering questions posed in natural language, developed in IBM's DeepQA project by a research team led by principal investigator David Ferrucci. Watson was named after IBM's first CEO, industrialist Thomas J. Watson. The computer system was specifically developed to answer questions on the quiz show Jeopardy! and, in 2011, the Watson computer system competed on Jeopardy! against former winners Brad Rutter and Ken Jennings winning the first place prize of \$1 million.

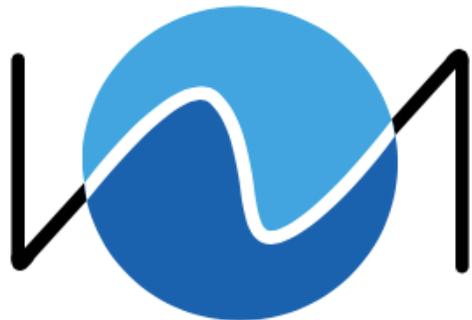
The sources of information for Watson include encyclopedias, dictionaries, thesauri, newswire articles, and literary works. Watson also used databases, taxonomies, and ontologies. Specifically, DBPedia, WordNet, and Yago were used. The IBM team provided Watson with millions of documents, including dictionaries, encyclopedias, and other reference material that it could use to build its knowledge.

... ..

**Relevant texts:** examples of services and applications which are based on IBM Watson cognitive platform, QA systems, comparison between IBM Watson and Wolfram Alpha

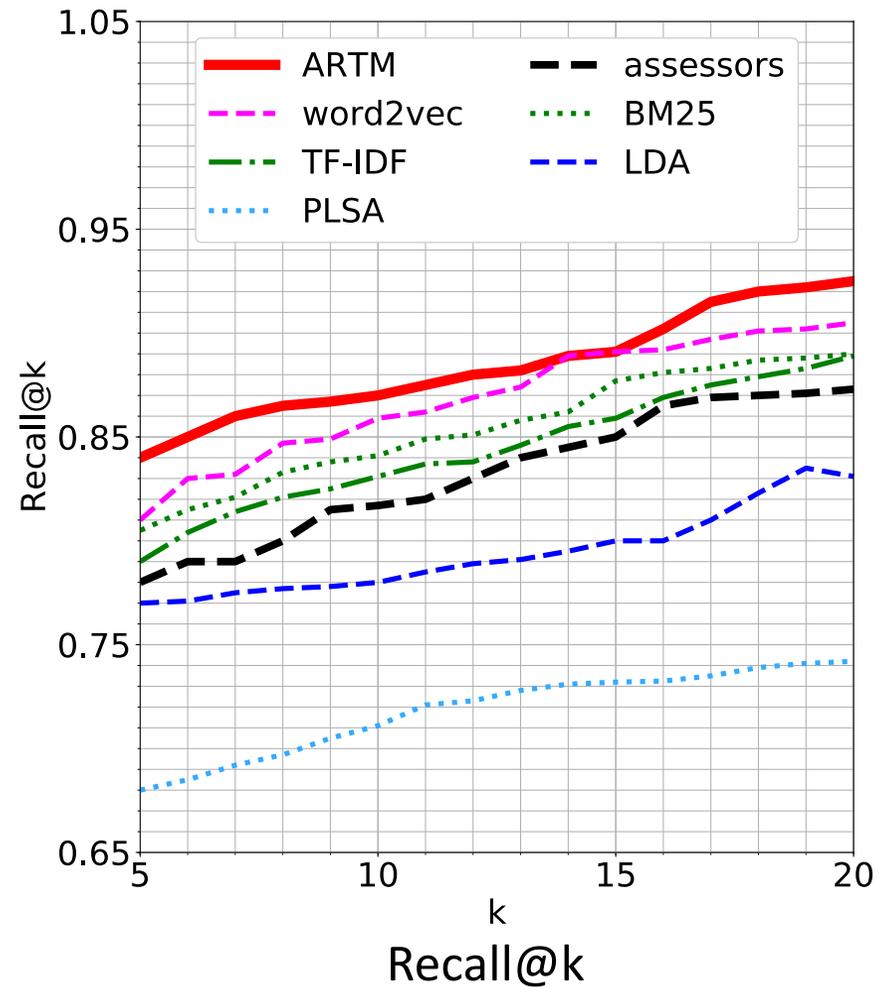
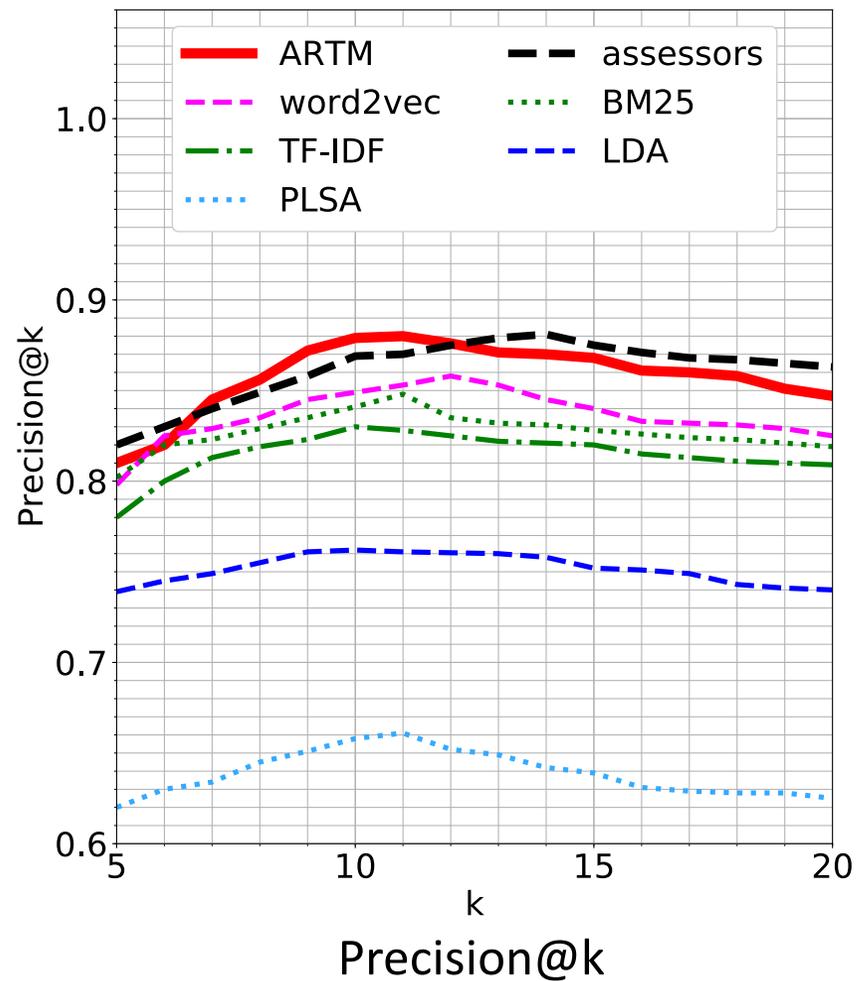
**Irrelevant texts:** common facts about artificial intelligence, other commercial solutions for solving business and analytic tasks.

# Результаты

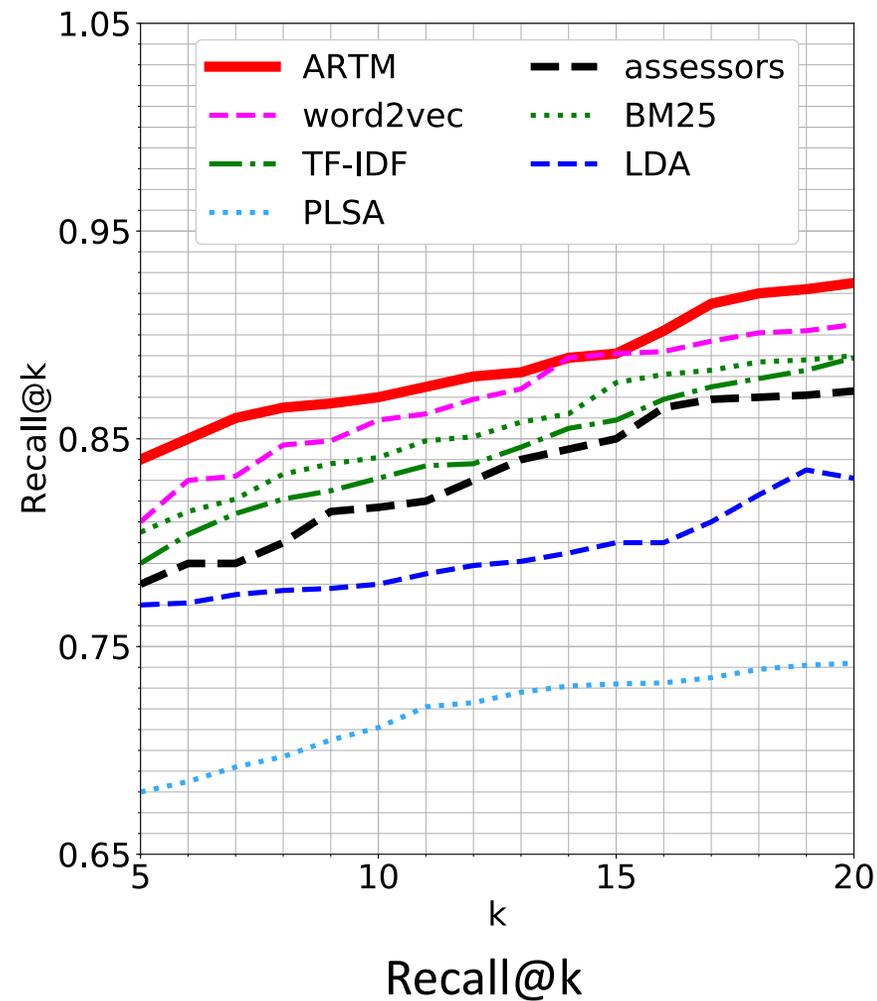
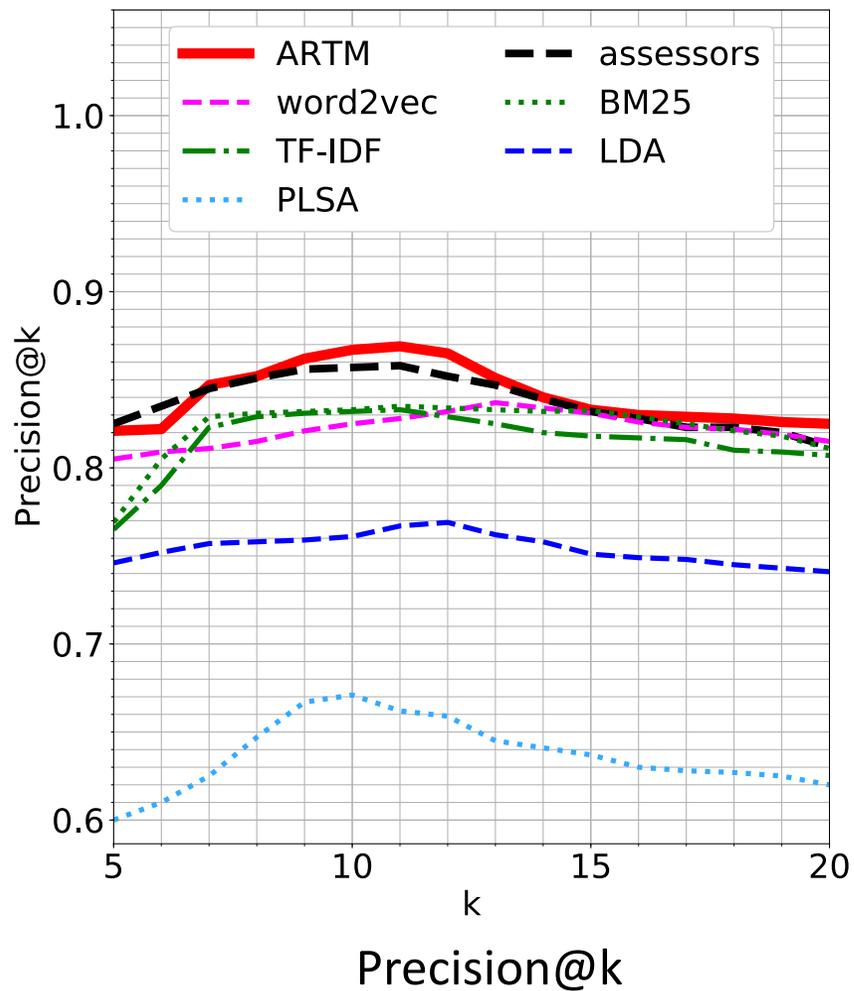


Разведочный поиск по  
текстовым коллекциям  
коллективных блогов  
Habrhabr и TechCrunch

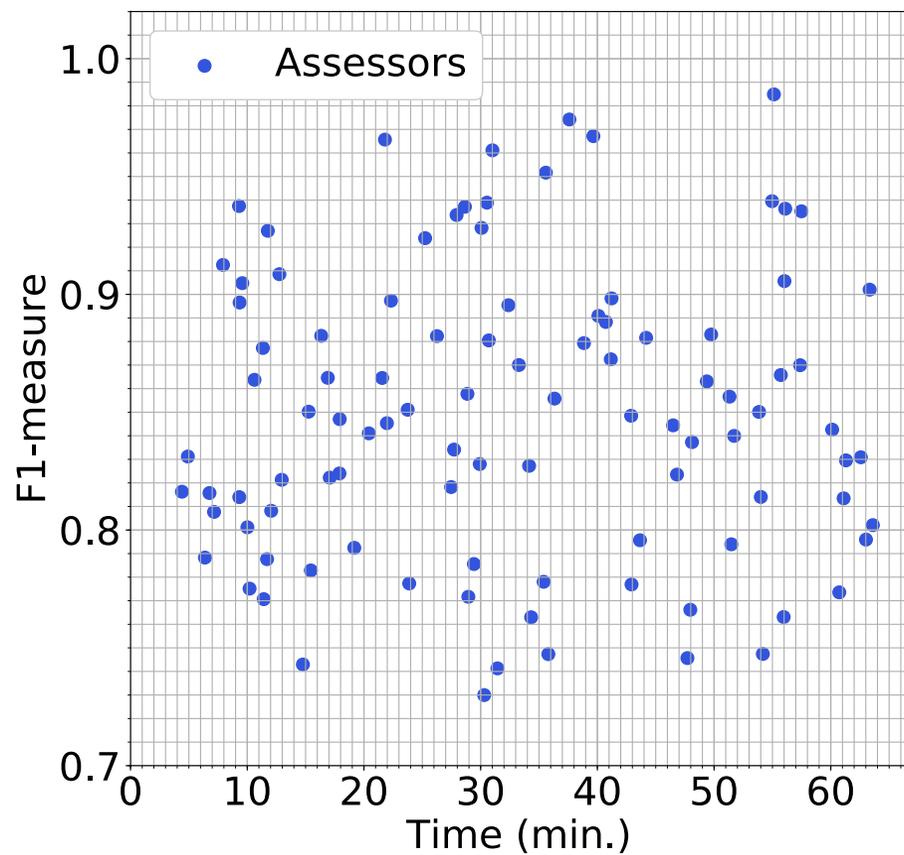
# Хабр: тематический поиск vs. ассессоры



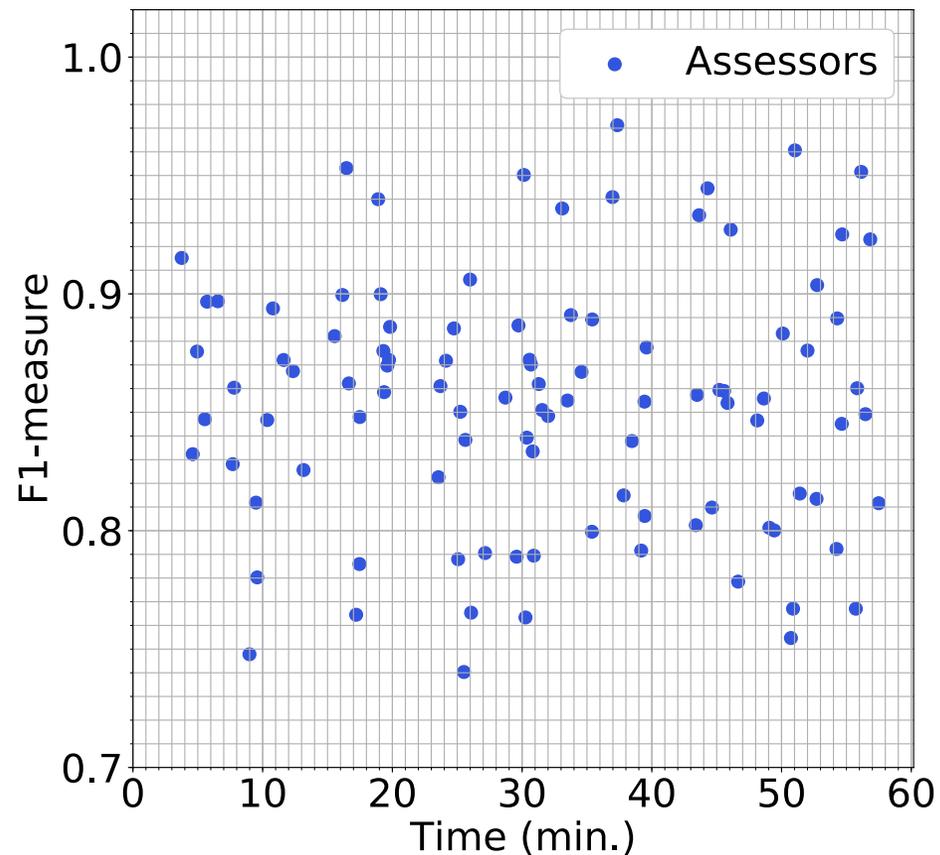
# TechCrunch: тематический поиск vs. ассессоры



# Тематический поиск: выигрыш по времени

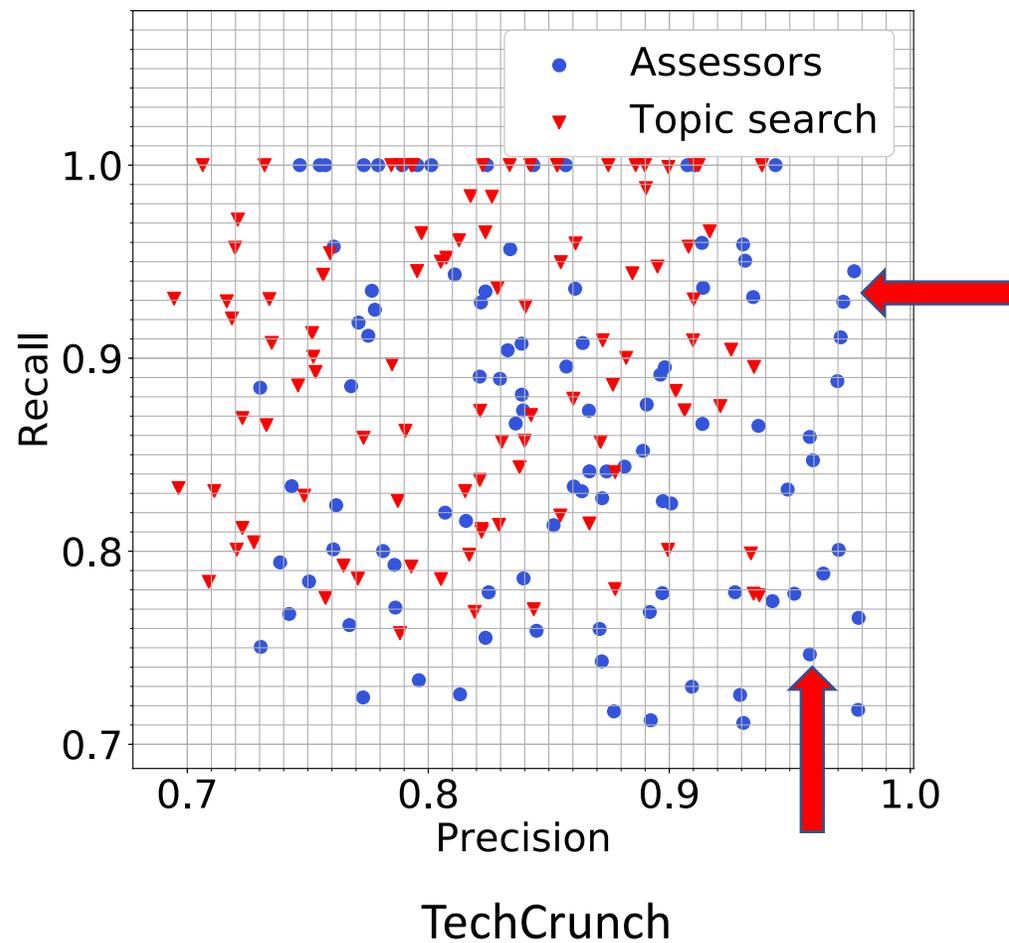
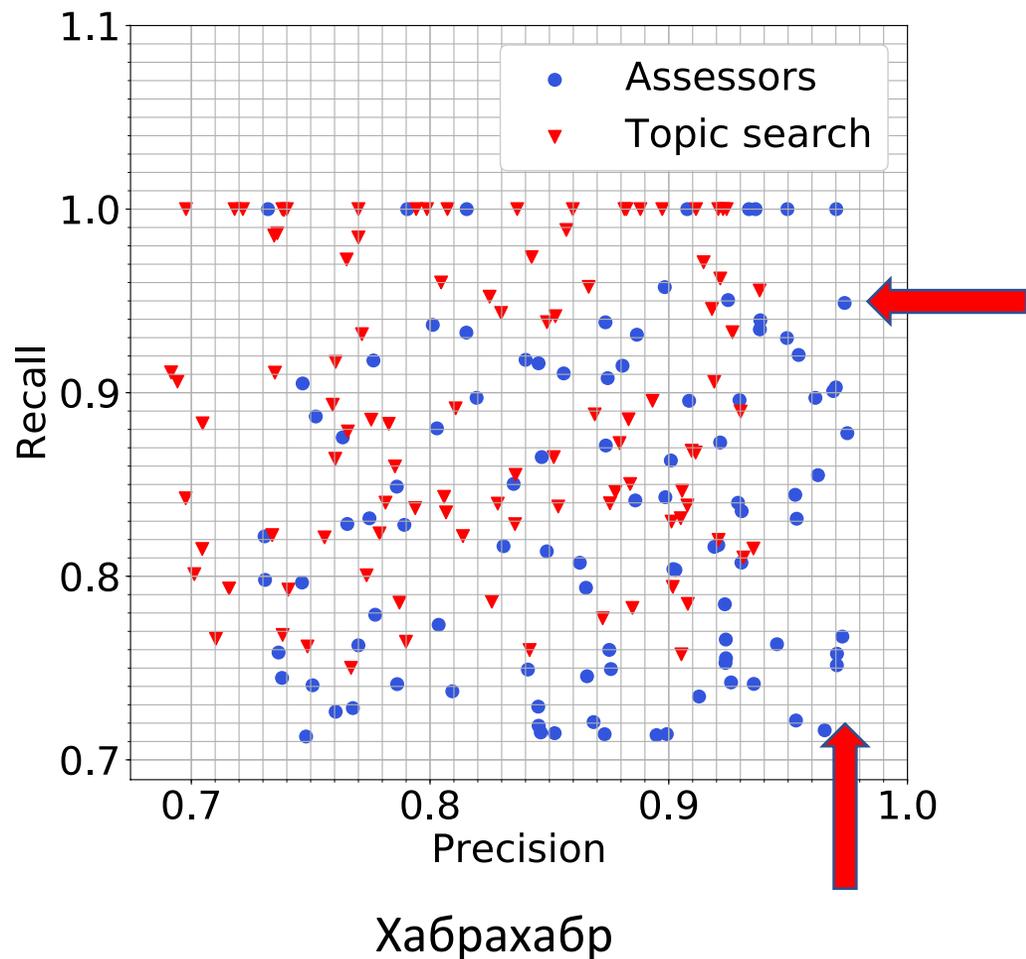


Хабрахабр

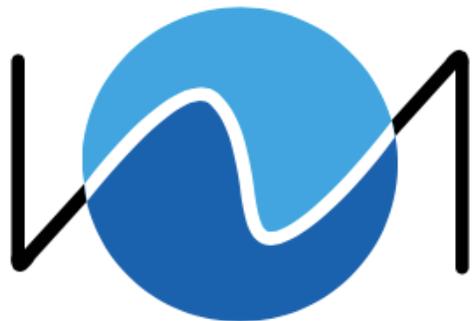


TechCrunch

# Тематический поиск vs. ассессоры: примеры

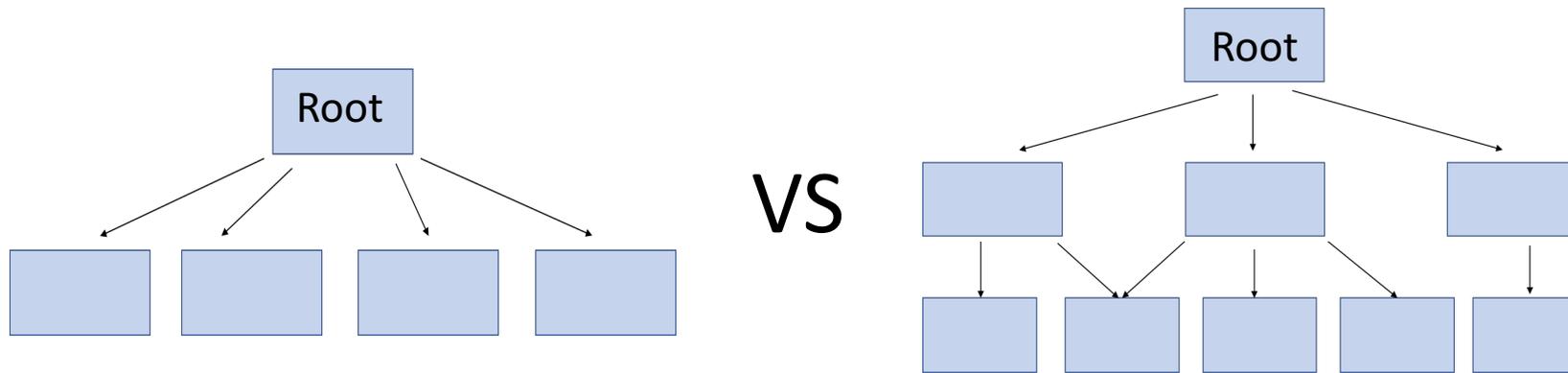


# Иерархические тематические модели



Как hARTM помогает искать  
близкие тексты?

# Тематизация запроса



Иерархические модели:

- Более гибкие
- Дочерняя тема может иметь более одного родителя (дочерняя тема «Социальные сети» наследуется от «Общение» и «IT»).
- Каскадная система поиска позволяет отсеивать откровенно нерелевантные документы на первом проходе, когда размер инвертированного индекса очень маленький
- В итоговом тематическом векторе гораздо меньше мусорных тем

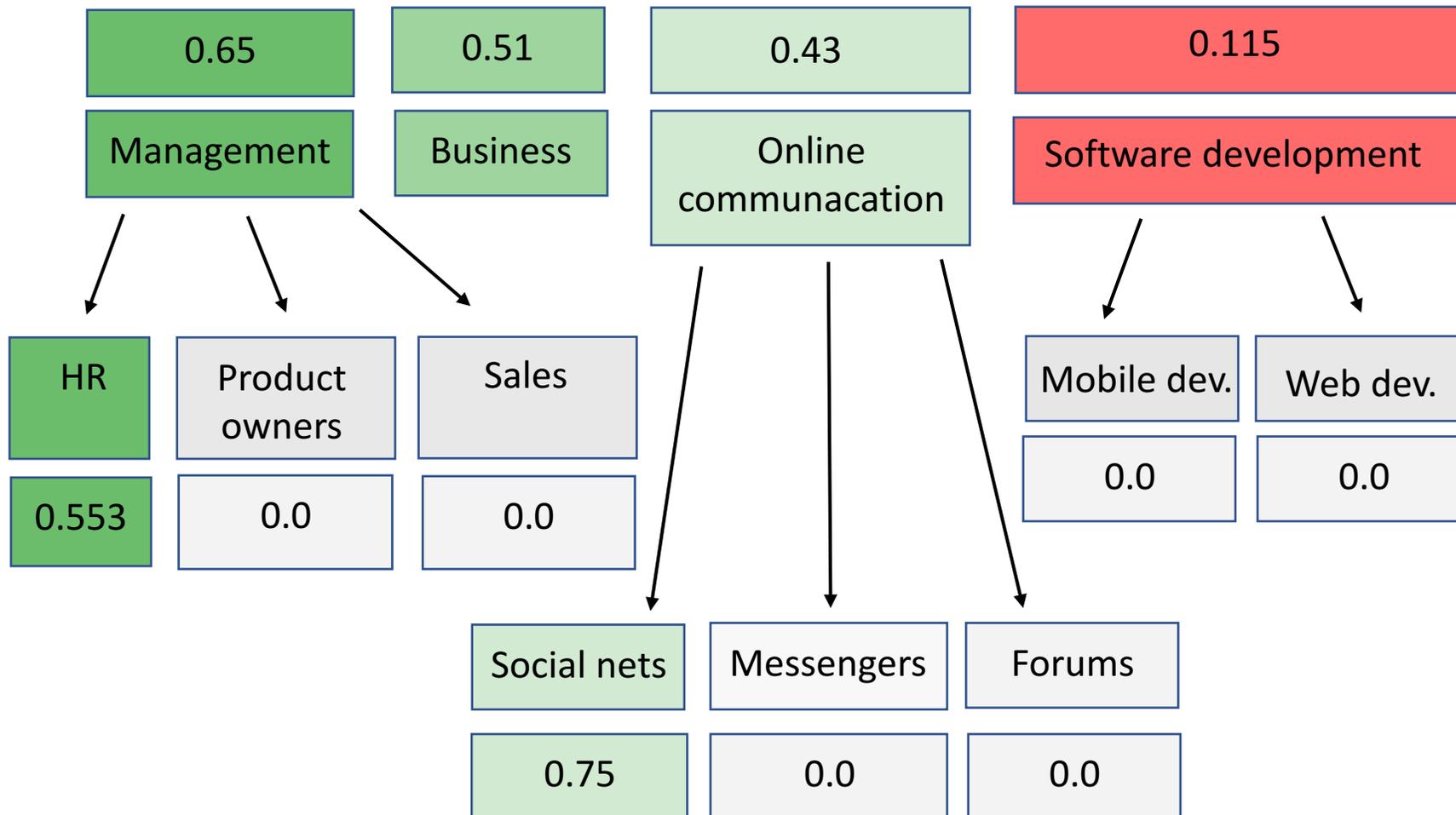
# Пример тематизации запроса

LinkedIn is a professional networking site that allows its members to create business connections, search for jobs, and find potential clients.

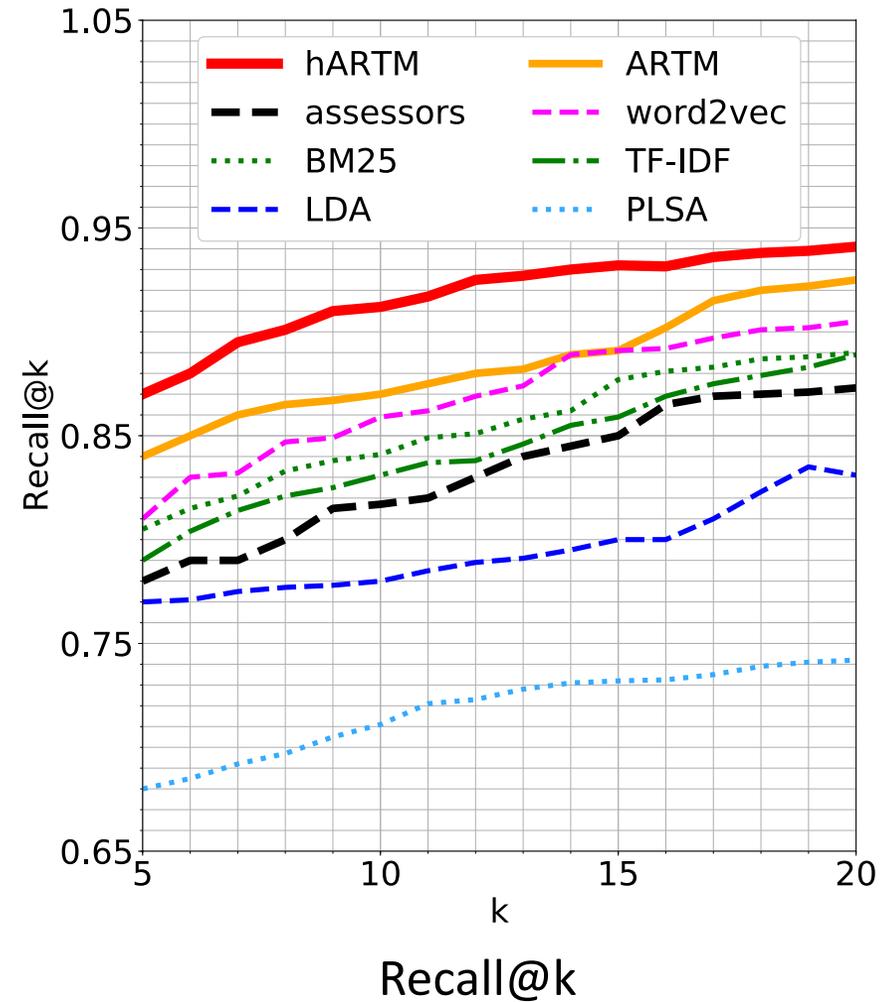
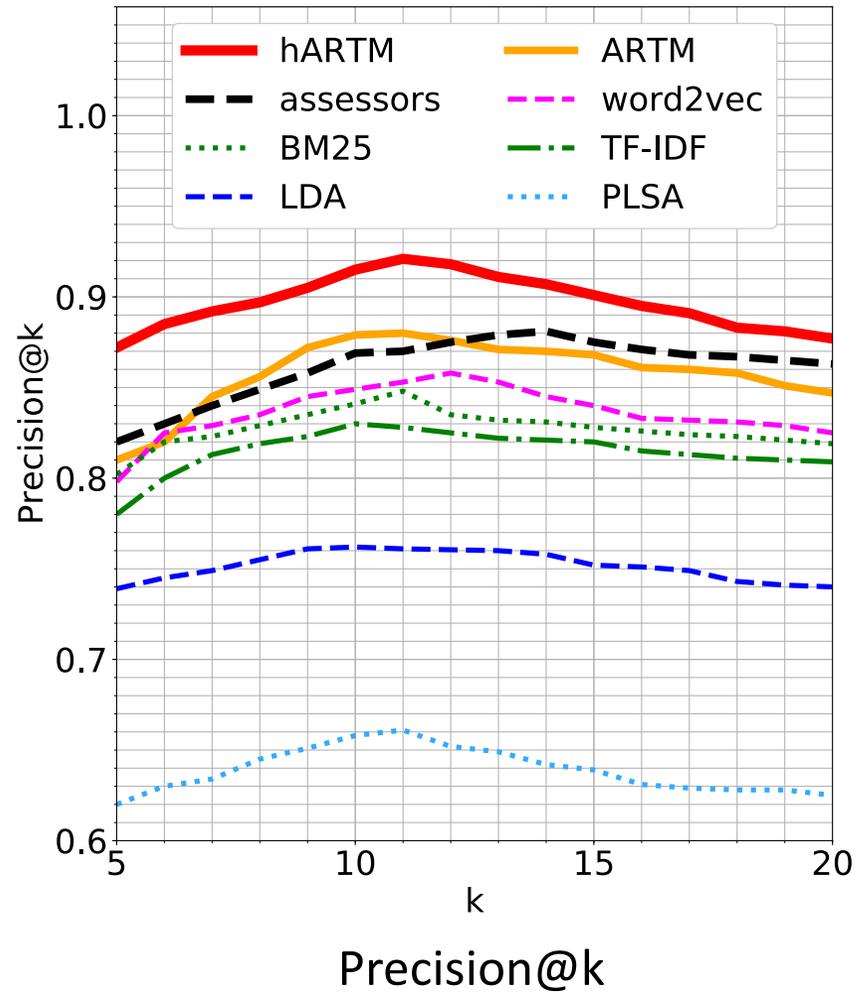
The site also enables its users to build and engage with their professional networks; access shared knowledge and insights; and find business opportunities. It offers LinkedIn mobile applications across various platforms and languages such as iOS, Android, Blackberry, Nokia Asha, and Windows Mobile, a public website that allows developers to integrate its content and services into their applications; and a set of embeddable widgets to allow web developers to include content from the company's network into their websites and applications.

HR	Business	Social nets	Mobile dev.	Web dev.
0.582	0.356	0.321	0.245	0.227

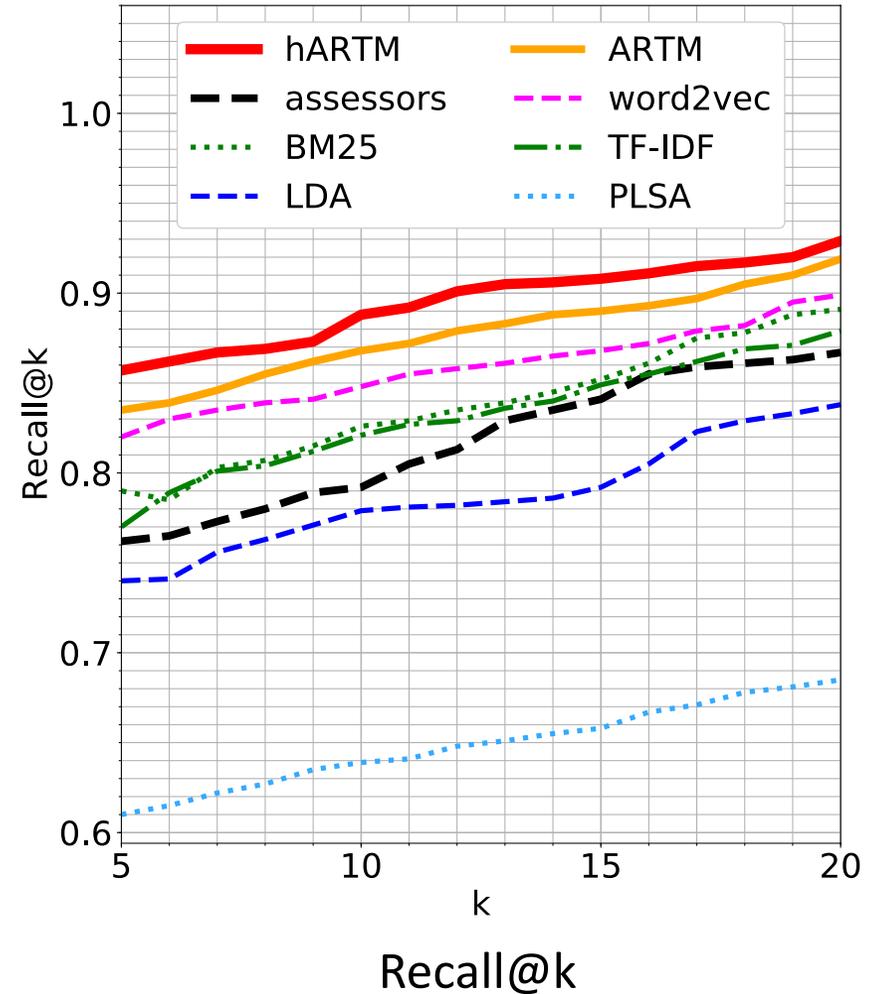
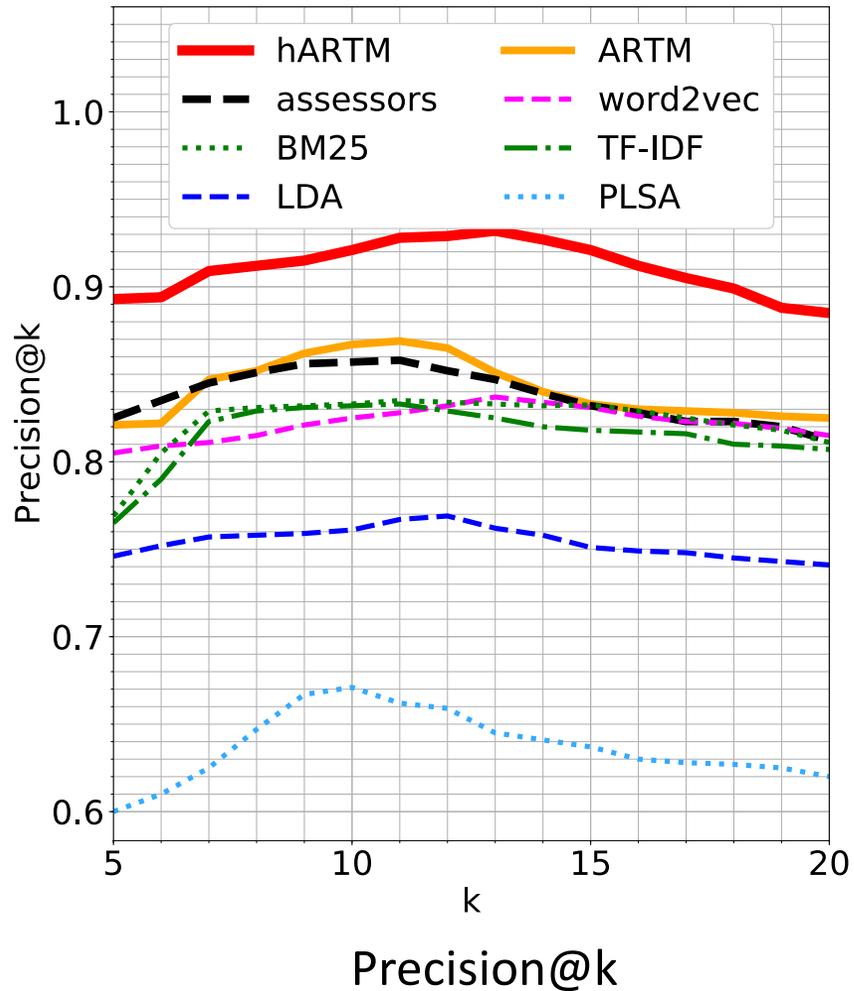
# Пример тематизации запроса



# Хабр: тематический поиск vs. ассессоры



# TechCrunch: тематический поиск vs. ассессоры



# Результаты

- Предложена технология и разработан тематический поисковик для разведочного информационного поиска
- Предложена технология оценки качества разведочного поиска (в терминах precision/recall) на основе ассессорских оценок
- Автоматический тематический поиск превосходит по качеству ассессорский поиск, а также дает значительный выигрыш по времени работы
- Тематический поиск дает выигрыш перед TF-IDF бейзлайном на ~15% (precision) и ~10% (recall)
- Разработанная каскадная система поиска эмулирует природу разведочного поиска, позволяя избавиться от итеративного уточнения и переформулировки запросов, что приводит к росту precision поиска более, чем на 7%
- Иерархическая тематизация позволяет решить проблему поиска по неоднородным текстовым коллекциям

# ССЫЛКИ

## Статьи:

- 1) Vorontsov, K., Frei, O., Apishev, M., Romov, P., Suvorova, M., & Yanina, A. (2015, October). Non-Bayesian additive regularization for multimodal topic modeling of large collections. In *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications* (pp. 29-37), International Conference on Information and Knowledge Management (CIKM), ACM. (Scopus, Web of Science, BAK)
- 2) Ianina A., Golitsyn L., Vorontsov K. [Multi-objective topic modeling for exploratory search in tech news](#) // Filchenkov A., Pivovarova L., Žižka J. (eds) Artificial Intelligence and Natural Language. AINL 2017, St. Petersburg, Russia, September 20-23, 2017. — Communications in Computer and Information Science, vol 789. Springer, Cham, 2017. — pp 181–193.
- 3) Янина, А. О., & Воронцов, К. В. (2016). Мультимодальные тематические модели для разведочного поиска в коллективном блоге. *Машинное обучение и анализ данных*, 2(2), 173-186 (<http://jmla.org/papers/doc/2016/no2/Yanina2016Multimodal.pdf>).
- 4) Ianina, A., & Vorontsov, K. Multimodal topic modeling for exploratory search in collective blog. In *Intelligent Data Processing: Theory and Applications: Book of abstracts of the 11th International Conference* (pp.186-187). <http://mmro.ru/files/2016-idp-11.pdf>
- 5) Vorontsov, K., Frei, O., Apishev, M., Romov, P., & Dudarenko, M. (2015, April). Bigartm: Open source library for regularized multimodal topic modeling of large collections. In *International Conference on Analysis of Images, Social Networks and Texts* (pp. 370-381). Springer, Cham.
- 6) Публикация в сборнике тезисов участников форума "Наука будущего - наука молодых" статьи "Multimodal topic model for exploratory search" ([http://sfy-conf.ru/fileadmin/downloads/Сборник\\_тезисов.pdf](http://sfy-conf.ru/fileadmin/downloads/Сборник_тезисов.pdf)) - индексируется РИНЦ

# ССЫЛКИ

## Конференции:

- 1) Выступление на конференции AINL ([ainlconf.ru](http://ainlconf.ru)) с докладом "Multi-objective Topic Modelling for Exploratory Search in Tech News".
- 2) Рассказ на мини-воркшопе: Mini-Workshop: Stochastic Processes and Probabilistic Models in Machine Learning (Topic modeling section).
- 3) Выступление с докладом «Multimodal topic models for exploratory search» на 11th International Conference on Intelligent Data Processing: Theory and Applications, октябрь 2016, Барселона
- 4) Выступление с докладом “Multimodal topic models for exploratory search” на The 5th Symposium on Conformal Probabilistic Prediction with Applications, Royal Holloway, University of London (UK) and CIEMAT (Spain), апрель 2016, Мадрид

Рассказ с DataFest’а: <https://www.youtube.com/watch?v=3Lxb-DqPtv4>