Методы поиска и анализа текстов в системах информационной поддержки принятия решений

Д.А.Девяткин
ФИЦ ИУ РАН
devyatkin@isa.ru

Информационная поддержка принятия решений

- Помощь лицам, принимающим решения, в использовании данных для решения неструктурированных или слабоструктурированных проблем.

(Gorry G, Thieraij R.J, Spraque R, Ларичев О.И., Петровский А.Б.)

Информационная поддержка принятия решений: работа с текстами

- В различных областях деятельности порождается большое количество текстов, которые могли бы быть полезны для информационной поддержки принятия решений.
- Необработанные массивы текстов затруднительно использовать для информационной поддержки принятия решений.

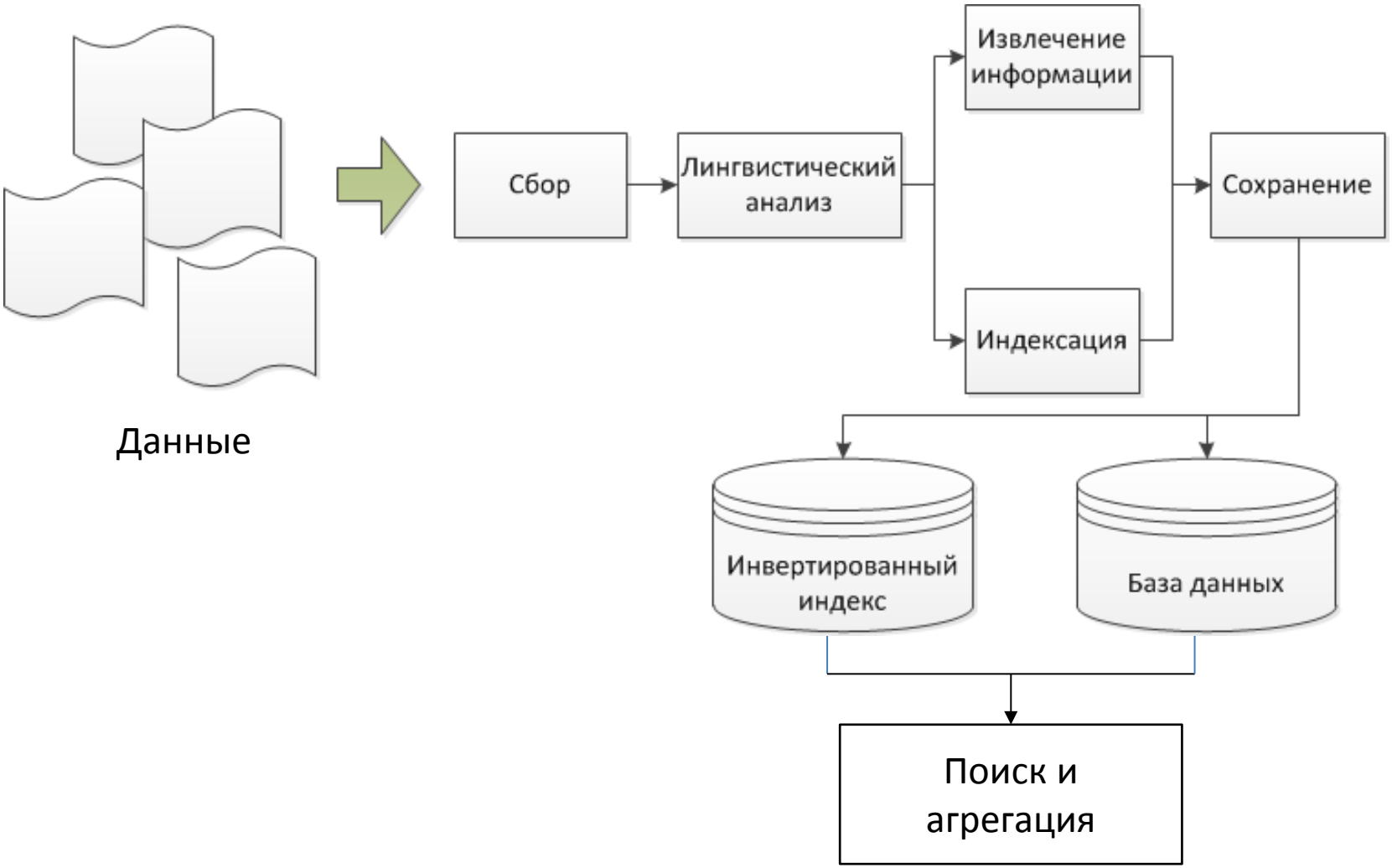
Проблемы

- Необходимость сфокусированного сбора информации, т.к. глубокая обработка всех порождаемых текстовых сообщений труднореализуема.
- Разнородность источников данных, разнородность программных интерфейсов (API).
- Небольшое количество готовых лингвистических инструментов, корпусов для обучения, особенно для русского языка.

Системы и инструменты

- **AIDR (Artificial Intelligence for Disaster Response)** – система сбора и анализа сообщений о ЧС в Twitter.
- **Tweedr** - библиотека для построения систем мониторинга соц. сетей (на примере Twitter).
- (Mitchell M, 1994), (Abdullah S., 2013) - прогнозирование динамики фондового рынка по сообщениям СМИ.
- **UMLS** содержит метатезаурус, словари, семантическую сеть и программные компоненты, которые позволяют сопоставлять концепты из разных медицинских и биомедицинских баз знаний друг с другом и находить их в текстах на естественном языке.
- **cTAKES** - платформа по анализу медицинских и биомедицинских текстов.

Основные этапы обработки данных

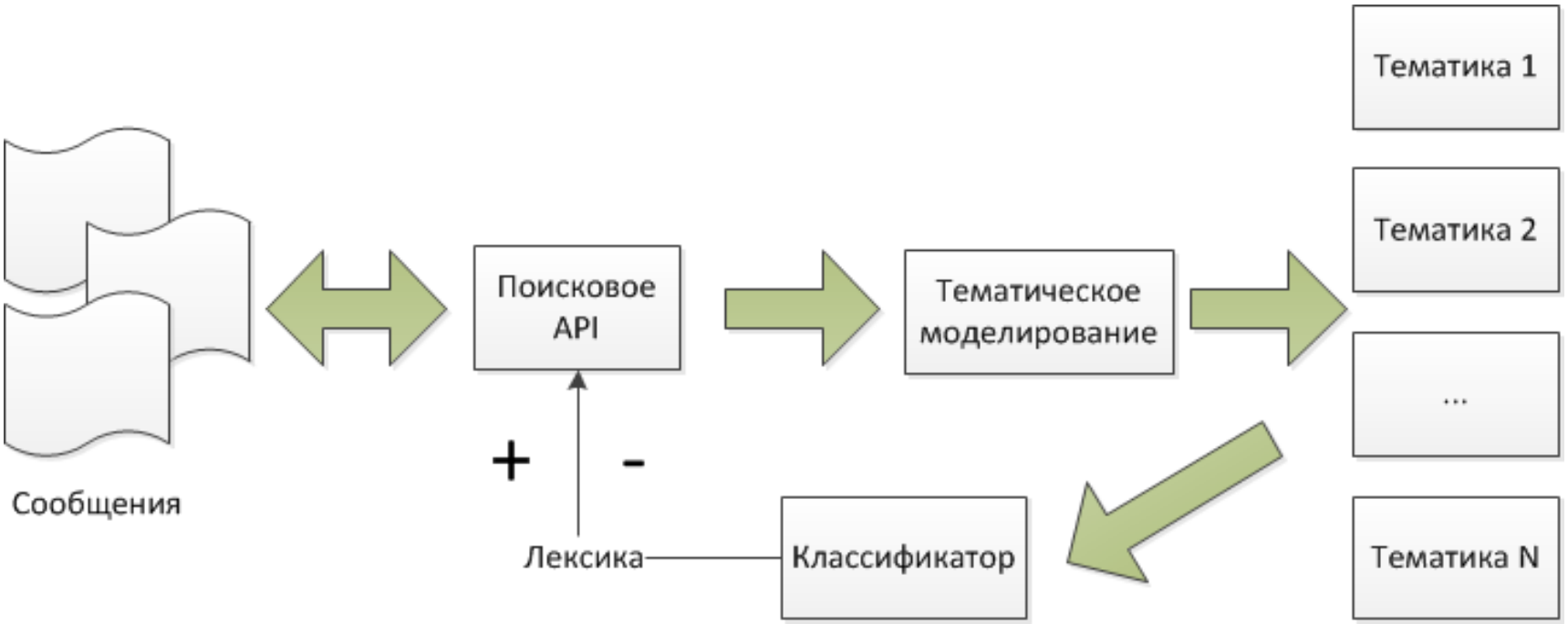


Сбор данных

- Основные источники данных:
  - Автоматизированные системы (импорт, API)
  - Базы документов (научно-технических, юридических, и др.) (импорт)
  - Социальные сети (API)
  - СМИ (RSS)
  - Ресурсы произвольной структуры в Интернете (веб-краулеры)

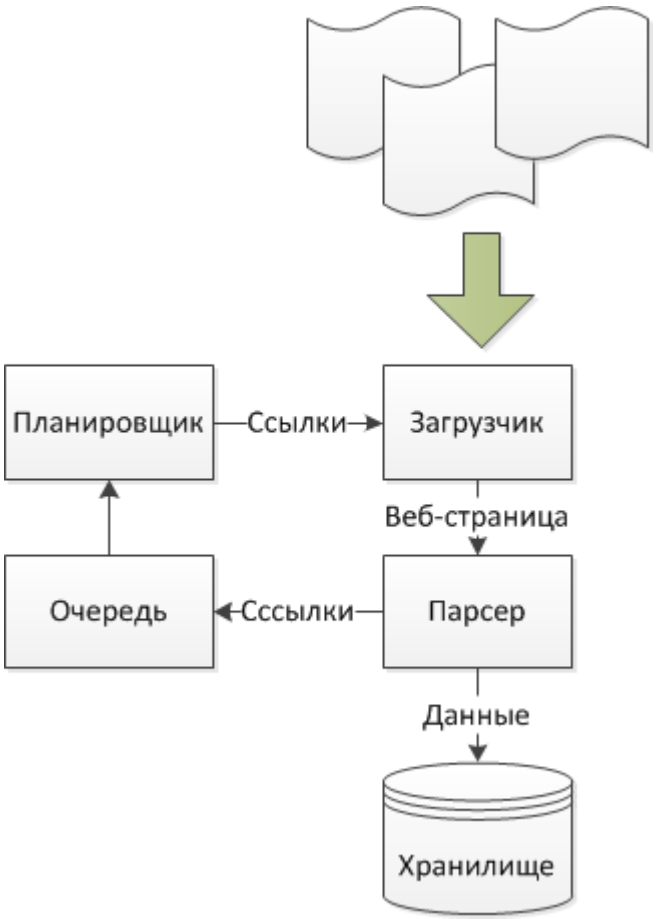
Сфокусированный сбор данных социальных сетей

- Используется поисковое API социальных сетей.
- Выразительная способность языка запросов соц. сетей ограничена.
- Количество запросов ограничено.

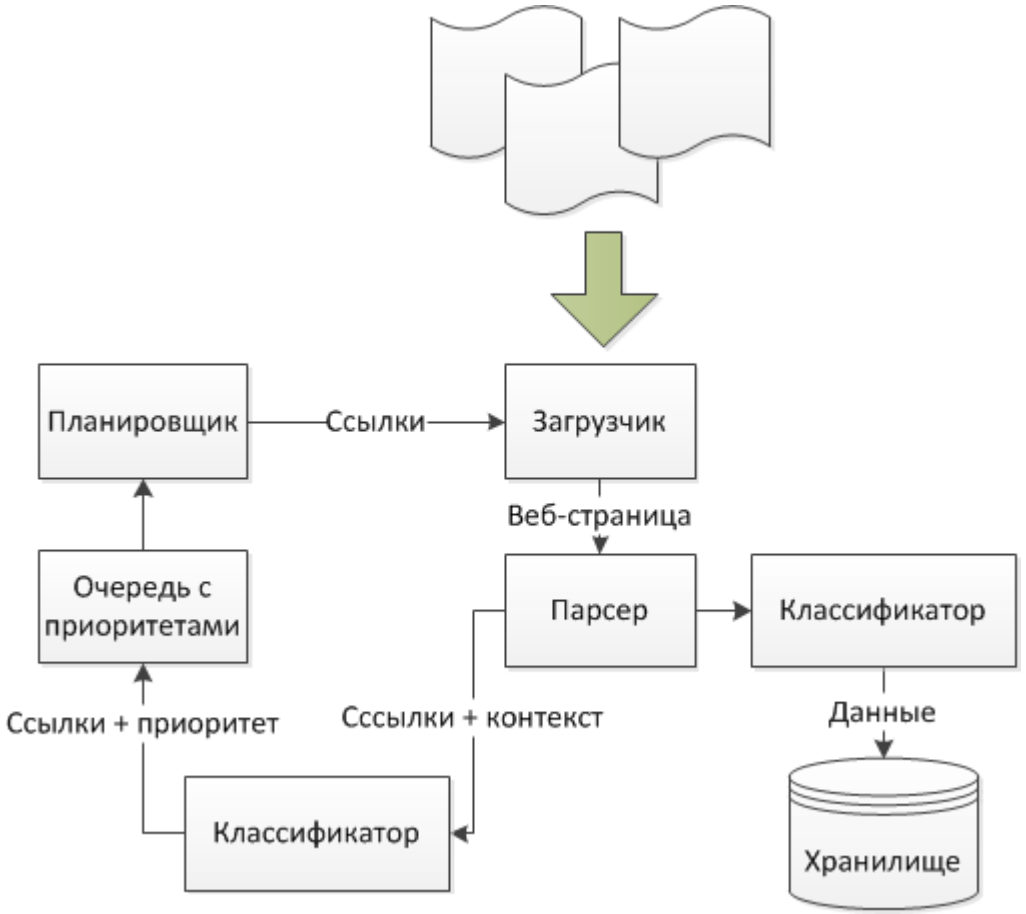


Веб-краулеры

Простой:



Сфокусированный:



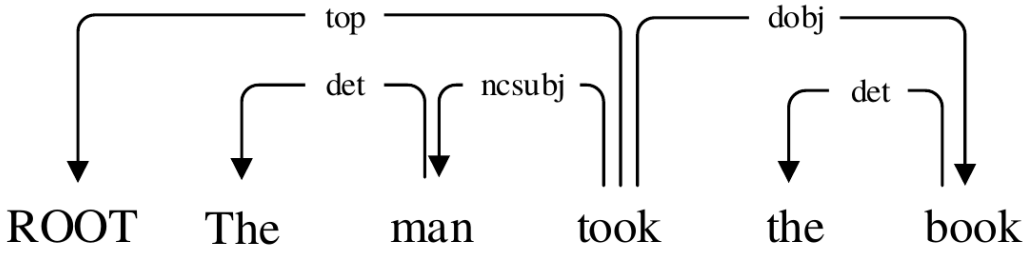
FishSearch, SharkSearch

Лингвистический анализ

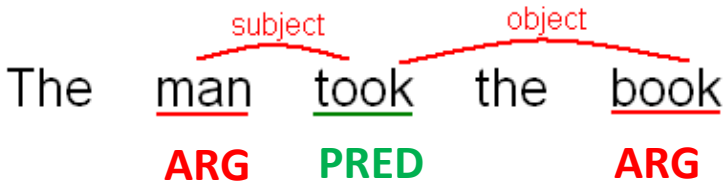
https://github.com/IINemo/isanlp, (Shelmanov et al, 2017)

	The	**man**	**took**	**the**	**book**
Часть речи	Артикль	Существительное	Глагол	Артикль	Существительное
Число		Ед.ч.		Ед.ч.	
Время			Прош.		
Лемма	the	man	take	the	book

Морфология



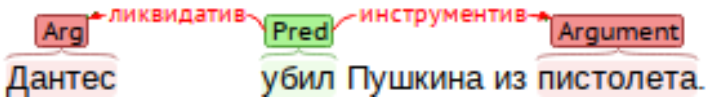
Синтаксис



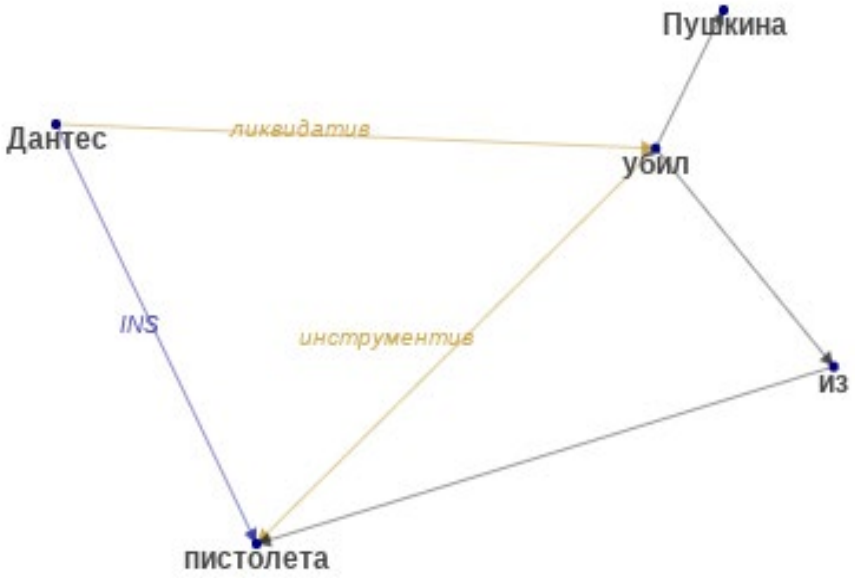
Семантика

Модели семантики текста

Ролевая структура предложения (Ч. Филлмор):

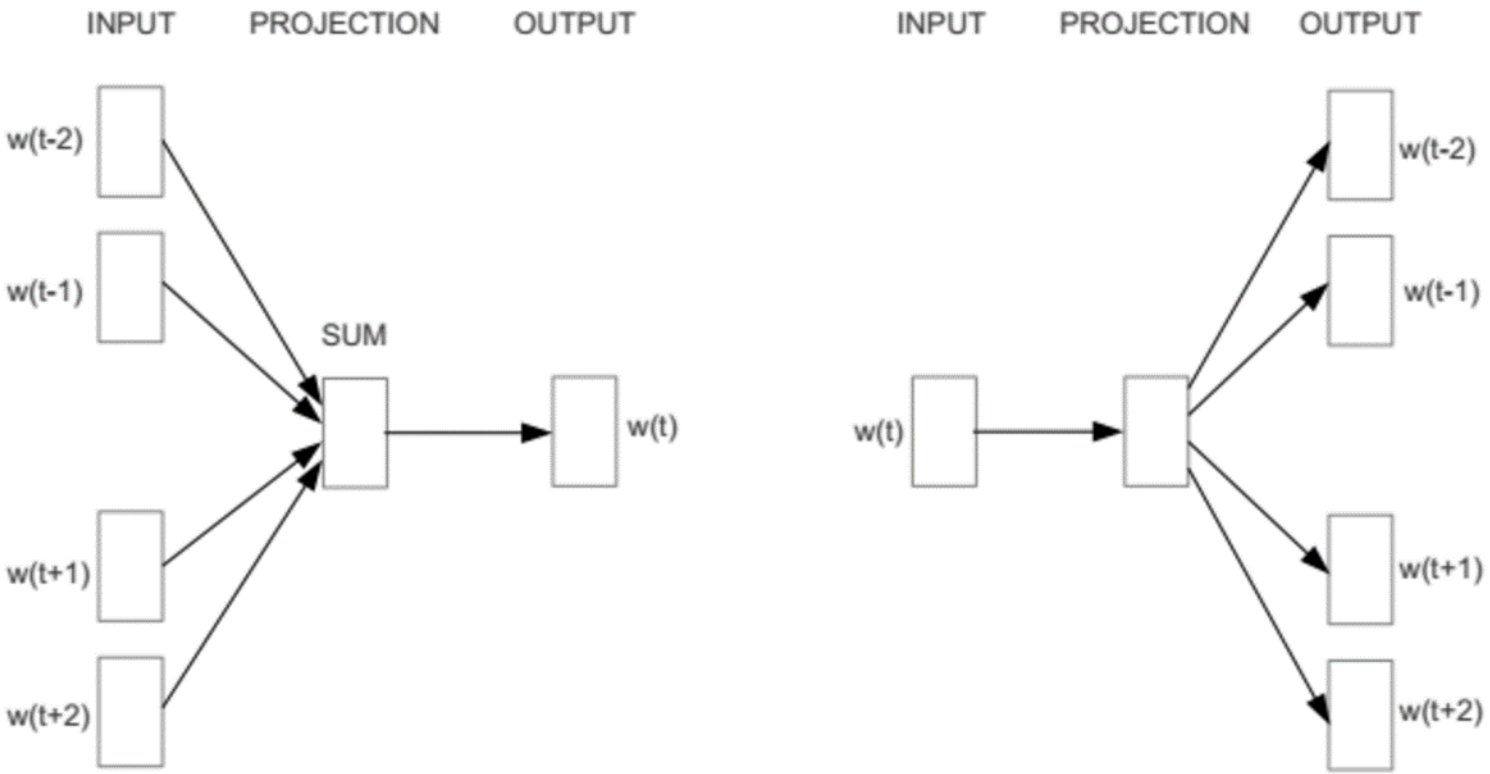


Реляционно-ситуационная модель текста (Osipov G, 2013)



Векторная модель (word embeddings)

(Mikolov T. et al., 2013)



CBOW **Skip-gram**

W_{t-2}	W_{t-1}	W_t	W_{t+2}	W_{t+2}
The	authorities	allocated	1	billion

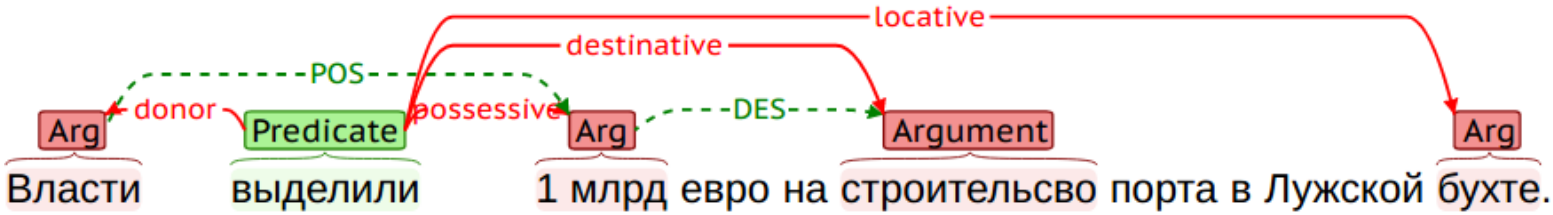
Извлечение информации

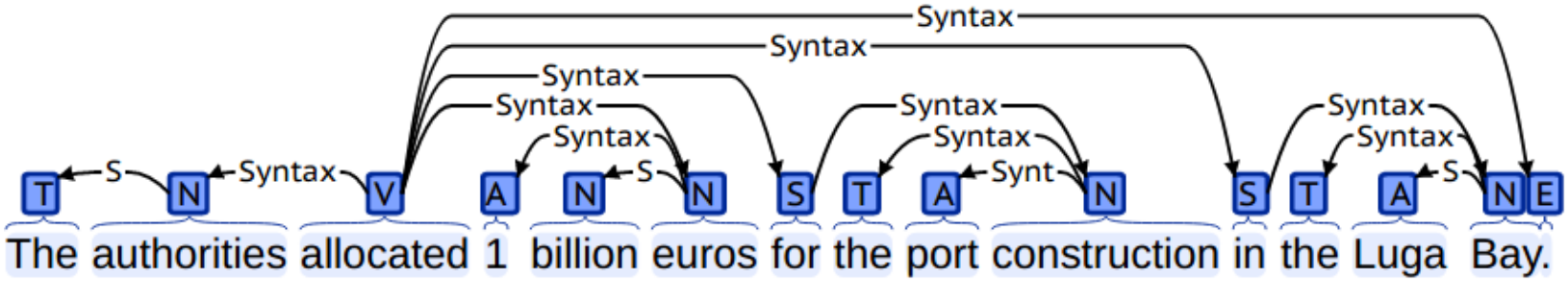
- В качестве признаков используются результаты лингвистического анализа текстов
- Методы:
  - основанные на словарях и правилах;
  - с использованием методов машинного обучения с учителем;
  - открытое извлечение информации.

Извлечение финансово-экономической информации: семантико-синтаксические шаблоны

Шаблон	Пример текстового фрагмента
НФ(«выделить») + * + [ЧР(Числ) + ЧР(Сущ)&КСК(количественное)?] + НФ(«рубль»)	На разработку ресурсов полярного региона в ближайшее время будет выделено почти 100 миллиардов долларов.
НФ(«привлечь») + * + [ЧР(Числ) + ЧР(Сущ)&КСК(количественное)?] + НФ(«доллар») + * + НФ(«инвестиция»)	Мурманская область в 2011 году привлекла 75 млн. долларов иностранных инвестиций.
НФ(«объем»)||НФ(«величина»)|| НФ(«порядок») + * +ЧР(Глаг) + [ЧР(Числ) + ЧР(Сущ)&КСК(количественное)?] + НФ(«рубль»)	За январь-ноябрь 2013 года объем инвестиций в основной капитал составил 41 947,7 млн рублей.
Обозначения: «НФ» – нормальная форма; «ЧР» – часть речи; КСК – категориально-семантический класс; & – логическое и; || – логическое или; * – любой текстовый фрагмент; […] – фрагмент повторяется 1 или более раз; ? – фрагмент необязателен.	

Извлечение финансово-экономической информации: семантико-синтаксические шаблоны





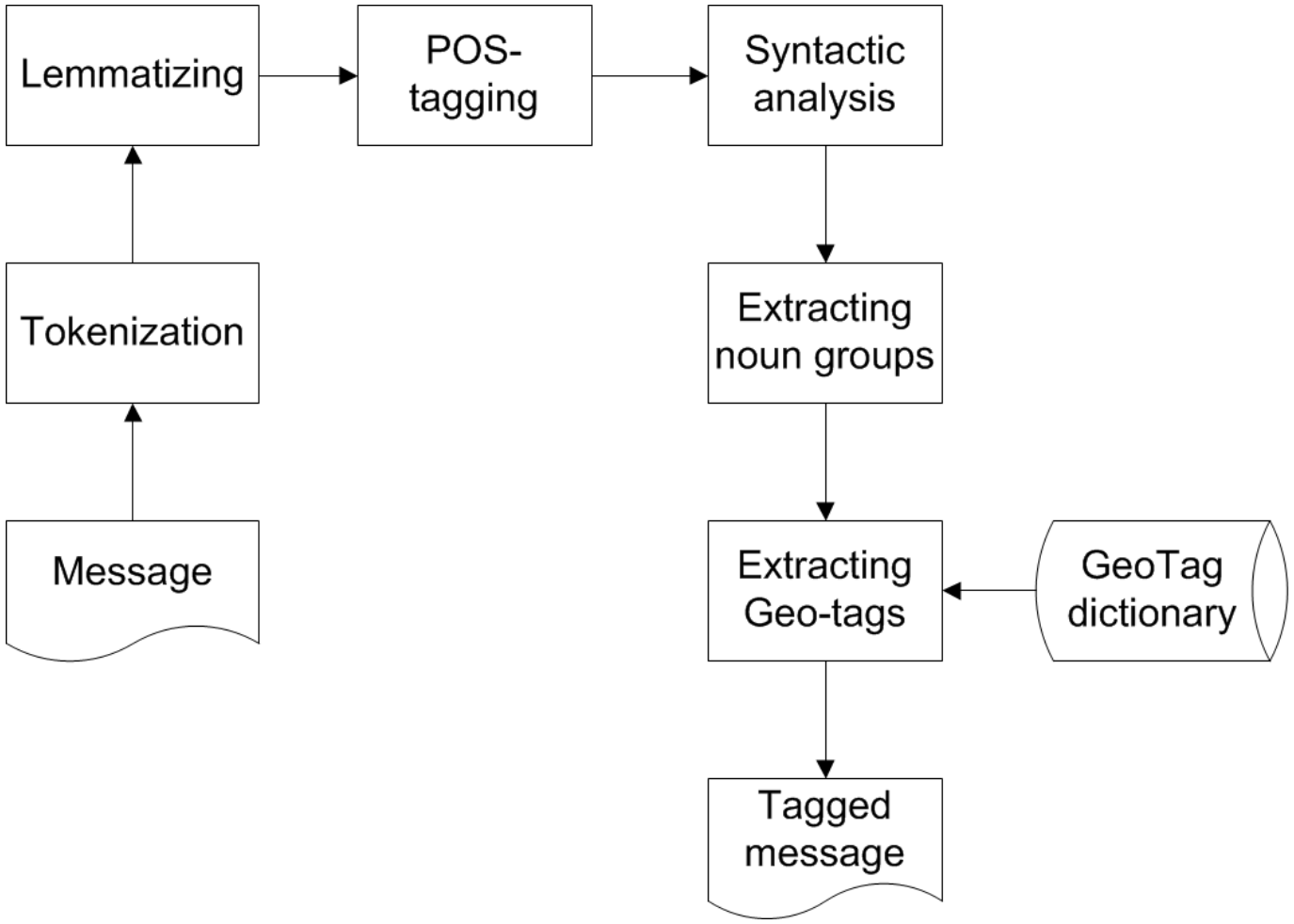
Извлечение именованных сущностей:

лексико-синтаксические шаблоны для выявления болезней и симптомов (Baranov , 2016)

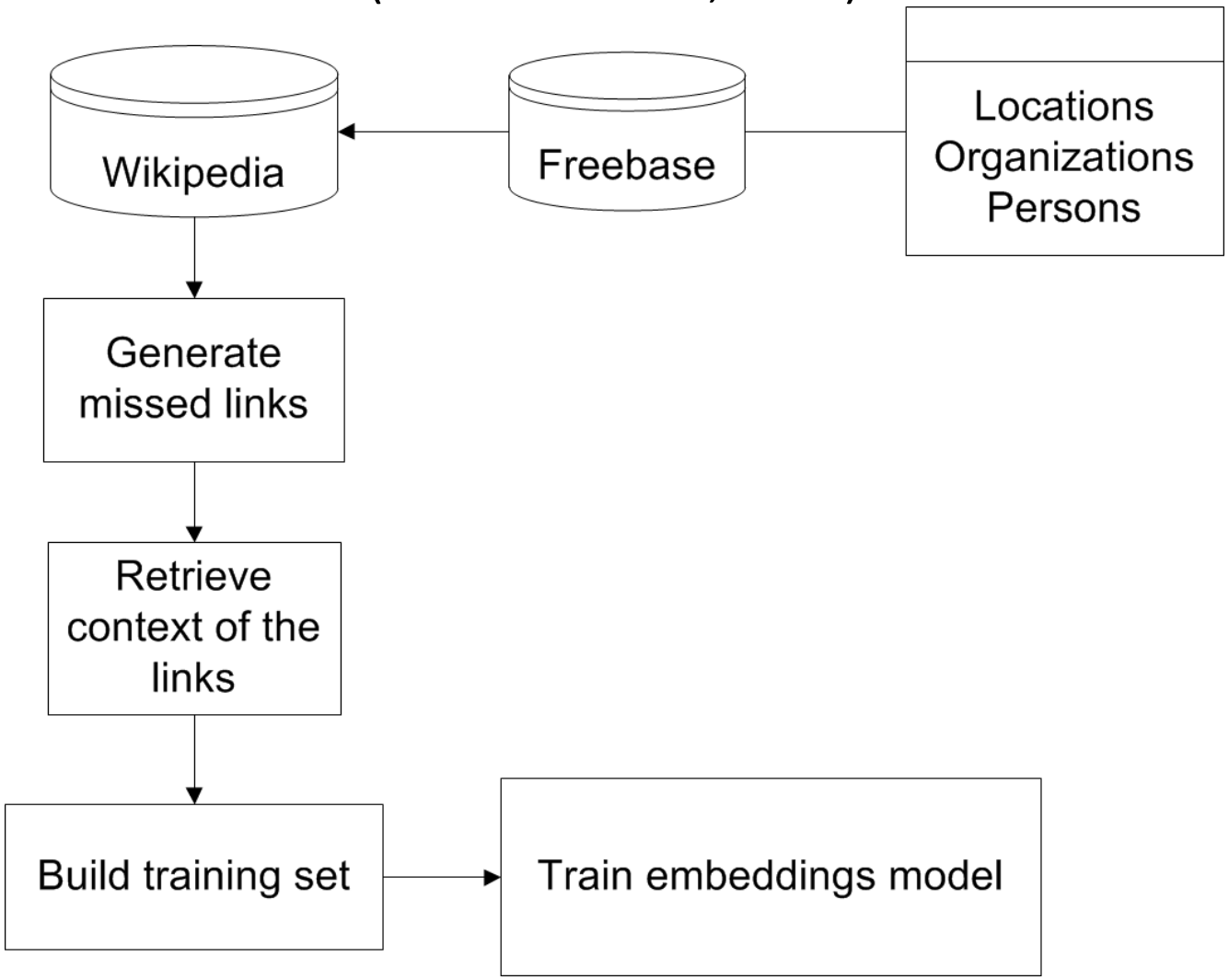
- Реализован поиск шаблонов в синтаксическом дереве
- Шаблоны для обнаружения отрицаний:
  - “не” синтаксически связана с описанием симптома или болезни
  - “не” синтаксически связана с предикатным словом
  - “нет” связана с упоминанием болезни или симптома
  - название симптома или болезни связано с предикатным словом, имеющим значение отрицания, например “отсутствует”
  - “нет” следует после упоминания болезни или симптома

Извлечение именованных сущностей: Газетер

Лингвистический анализ + Словарь (GeoNames.org, MarineTraffic)



Извлечение именованных сущностей: Polyglot
(Al-Rfou R. et al, 2015)

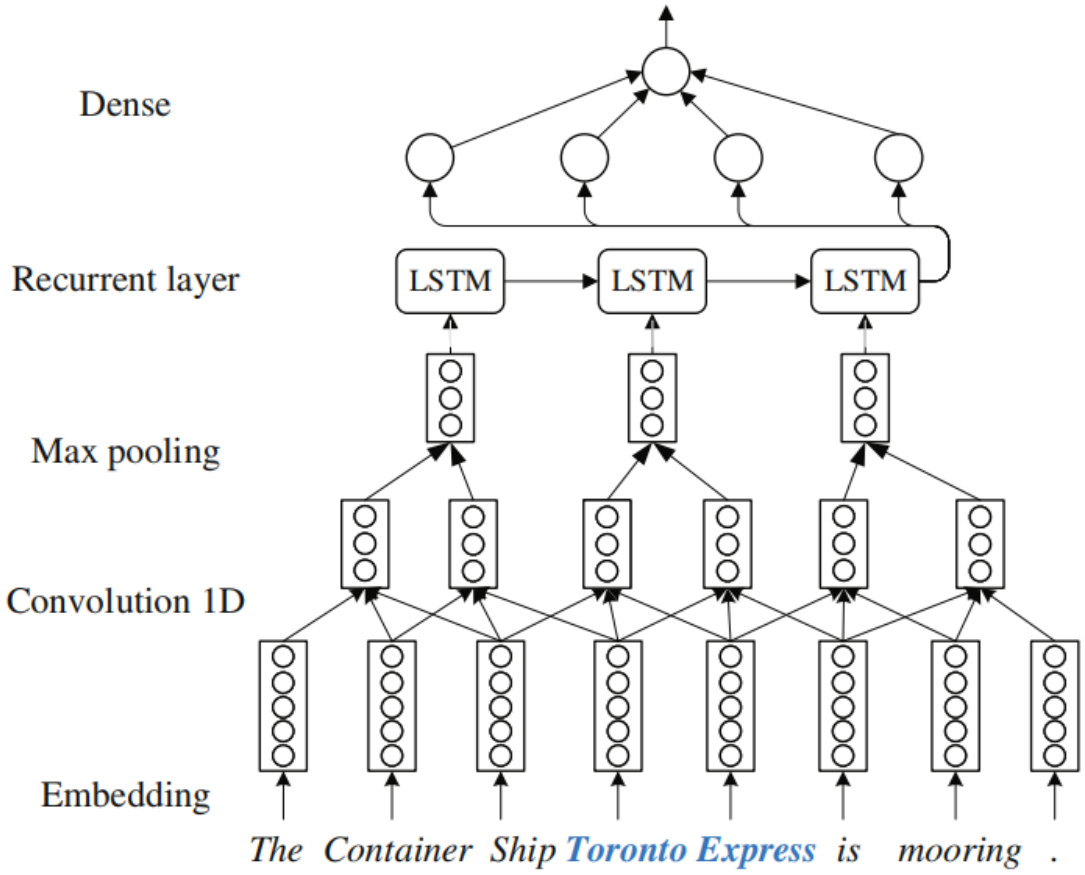


Finland says new Arctic railway should lead to Kirkenes.

{I-LOC Kirkenes, I-LOC Arctic, I-ORG Finland .}

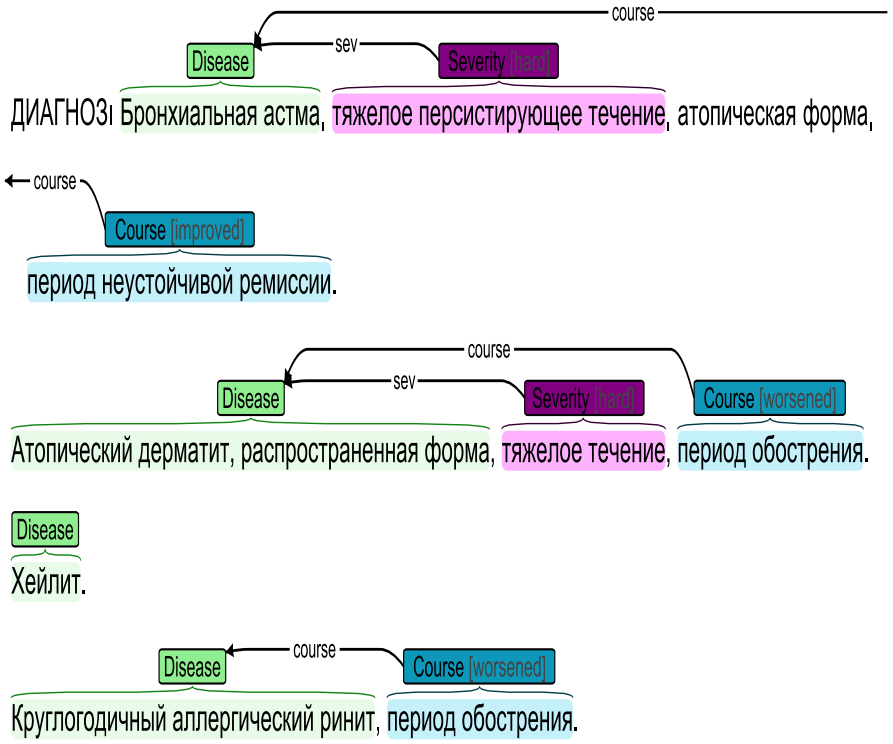
Извлечение именованных сущностей: машинное обучение с учителем для выделения наименований судов (Devyatkin et al, 2016)

- Около 2 тыс. Размеченных сообщений Twitter
- Нейронная сеть C-LSTM
  - Сверточный слой
  - Рекуррентный слой



Извлечение именованных сущностей:

машинное обучение с учителем для выявления течения и тяжести заболеваний



Около 45 тыс. токенов. Более 7,600 размеченных сущностей и более 4,000 размеченных атрибутов и отношений

Связанные ресурсы на английском языке:

- ShARe (Mowery D. L., 2014)
- SHARPn (Pradhan S., 2015)

Анонимизация: удалены имена, изменены даты

Доступен для исследовательских целей: http://nlp.isa.ru/datasets/clinical

Размечено с помощью инструмента Brat – http://brat.nlplab.org

Выявление течения и тяжести заболеваний (Baranov , 2016)

- Линейный SVM, random forest, AdaBoost.
- Использовался подход «скользящего окна»
- Признаки:
  - Морфологические признаки: леммы и части речи токенов;
  - Наличие синтаксической связи классифицируемого токена с наименованием заболевания;
  - Расстояние в тексте между токеном и наименованием заболевания.

Сопоставление подходов

Извлечение наименований кораблей

Анализатор	P	R	F_1
Газетер	0,41	1,00	0,58
Нейронная сеть	**0,90**	**0,92**	**0,91**

Извлечение географических локаций

Анализатор	P	R	F_1
Polyglot	0,78	0,57	0,66
Газетер	0,78	0,74	0,76
Polyglot + газетер	0,76	**0,82**	**0,79**

Формирование ассоциативных правил, выражающих шаблонные комбинации извлеченных сущностей

Алгоритмы:
Apriori (Agrawal R. et al., 1994)
DHP (J. Park, M. Chen and P. Yu, 1995) и др.

- *('ГИПЕРСЕНСИБИЛИЗАЦИЯ', 'АСТМА БРОНХИАЛЬНАЯ', 'РИНИТ') => ('Монтелукаст')*
- *('ГИПЕРСЕНСИБИЛИЗАЦИЯ', 'АСТМА БРОНХИАЛЬНАЯ') => ('Будесонид')*
- *('ГИПЕРСЕНСИБИЛИЗАЦИЯ', 'РИНИТ') => ('АСТМА БРОНХИАЛЬНАЯ')*
- *('КАШЕЛЬ', 'АСТМА БРОНХИАЛЬНАЯ') => ('РИНИТ')*
- *('АСТМА БРОНХИАЛЬНАЯ') => ('Флутиказон')*
- *('IGE повышен', 'РИНИТ') => ('КАШЕЛЬ')*
- *('IGE повышен') => ('Уровень базофилов повышен')*
- *('IGE повышен', 'ГИПЕРСЕНСИБИЛИЗАЦИЯ ') => ('РИНИТ')*

Тональность текста и психолингвистические показатели эмоциональной напряженности

Тексты, написанные здоровыми людьми в состоянии эмоционального напряжения, содержат индикаторы неблагополучия, которые отсутствуют в текстах тех же авторов, написанных в другое время (раньше и позже эмоциогенной ситуации)

- Массовое повышение индикаторов эмоционального напряжения свидетельствует об эмоциональном заражении - процессы группирования на основе общности аффекта

ПСИХОЛИНГВИСТИЧЕСКИЕ МАРКЕРЫ

(Vybornova, 2011)

Разработаны методы вычисления значений маркеров:

- 1. Количество слов в предложении.
- 2. Коэффициент определенности действия.
- 3. Количество глаголов в пассивном залоге.
- 4. Средняя длина слова.
- 5. Отношение количества инфинитивов к общему числу глаголов.
- 6. Количество безличных глаголов.
- 7. Количество местоимений.
- 8. Коэффициент Трейгера отношение количества глаголов к количеству прилагательных.
- 9. Количество глаголов несовершенного вида.

ПСИХОЛИНГВИСТИЧЕСКИЕ МАРКЕРЫ

- 10. Количество местоимений 1-го лица множественного числа.
- 11. Количество инфинитивов.
- 12. Отношение числа глаголов и существительных к числу прилагательных и наречий.
- 13. Количество глаголов первого лица, единственного числа, прошедшего времени.
- 14. Количество предложений в тексте.
- 15. Средний размер предложения.
- 16. Отношение количества глаголов будущего времени к общему количеству глаголов.
- 17. Количество местоимений 3-го лица множественного числа.

➢ Значения маркеров вычисляются непосредственно на основе результатов лингвистического анализа.

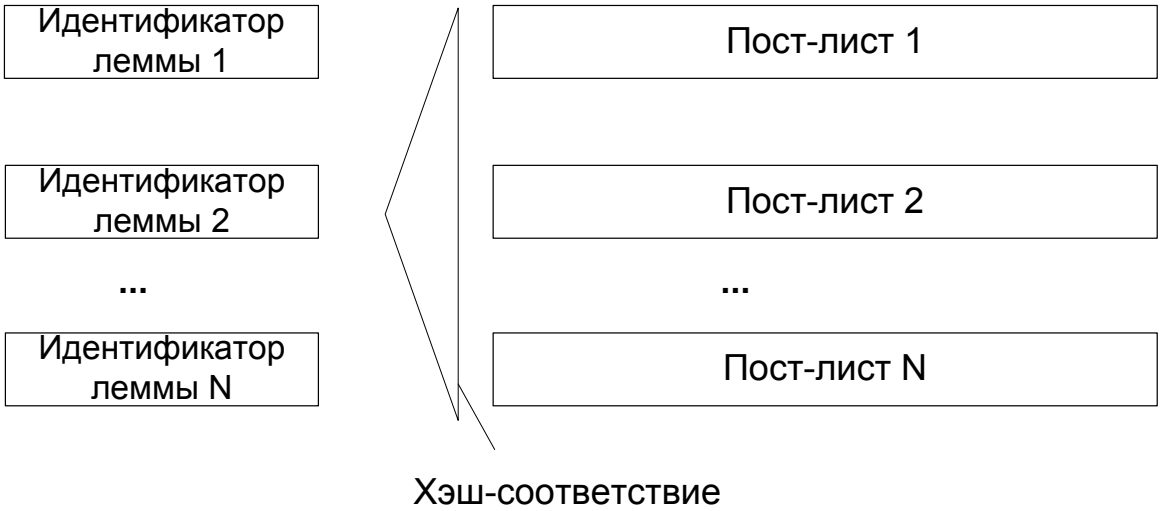
26

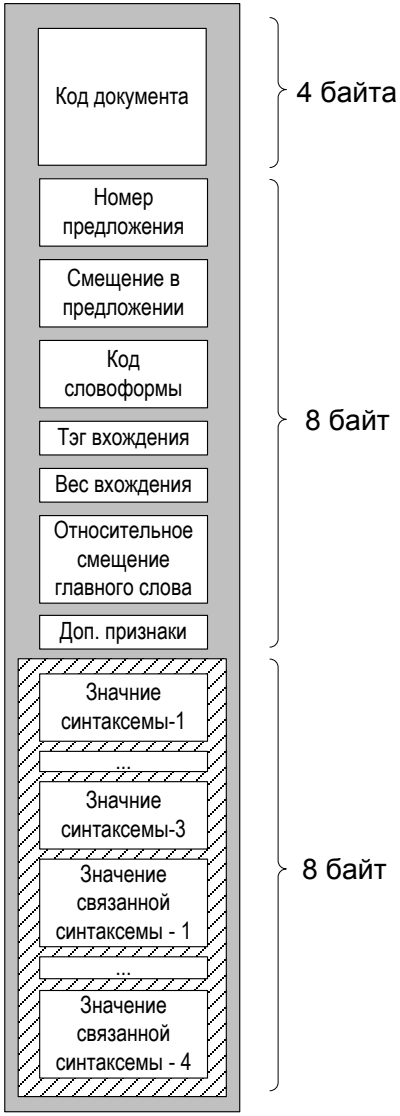
Поиск

- Полнотекстовый поиск с ранжированием (Abbas A, 2014).
- Поиск по запросу и по документу.
- Фасетный поиск.
- Учет синтаксиса (фразовый поиск) и семантики запроса.

Классический инвертированный индекс

- Инвертированный индекс реализуется в виде хэш-таблицы
- Ключ – идентификатор леммы
- Значение – «пост-лист» (упорядоченный набор **элементов данных** (ЭД), представляющих информацию о словоупотреблениях в текстах)
- Некоторые поля ЭД могут быть «пустыми» (иметь значение по умолчанию)
- Поиск информации решается как задача «слияния упорядоченных списков»

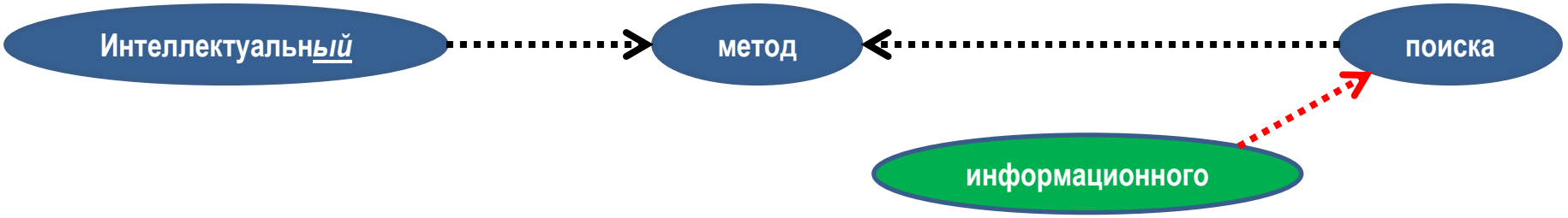




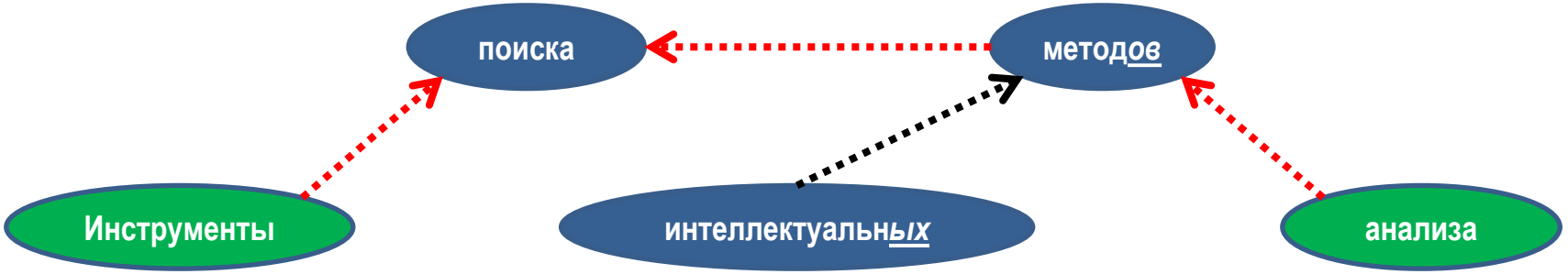
Оценка сходства текстов с учётом синтаксических структур

1. Интеллектуальные методы поиска…

2. Интеллектуальные методы информационного поиска…

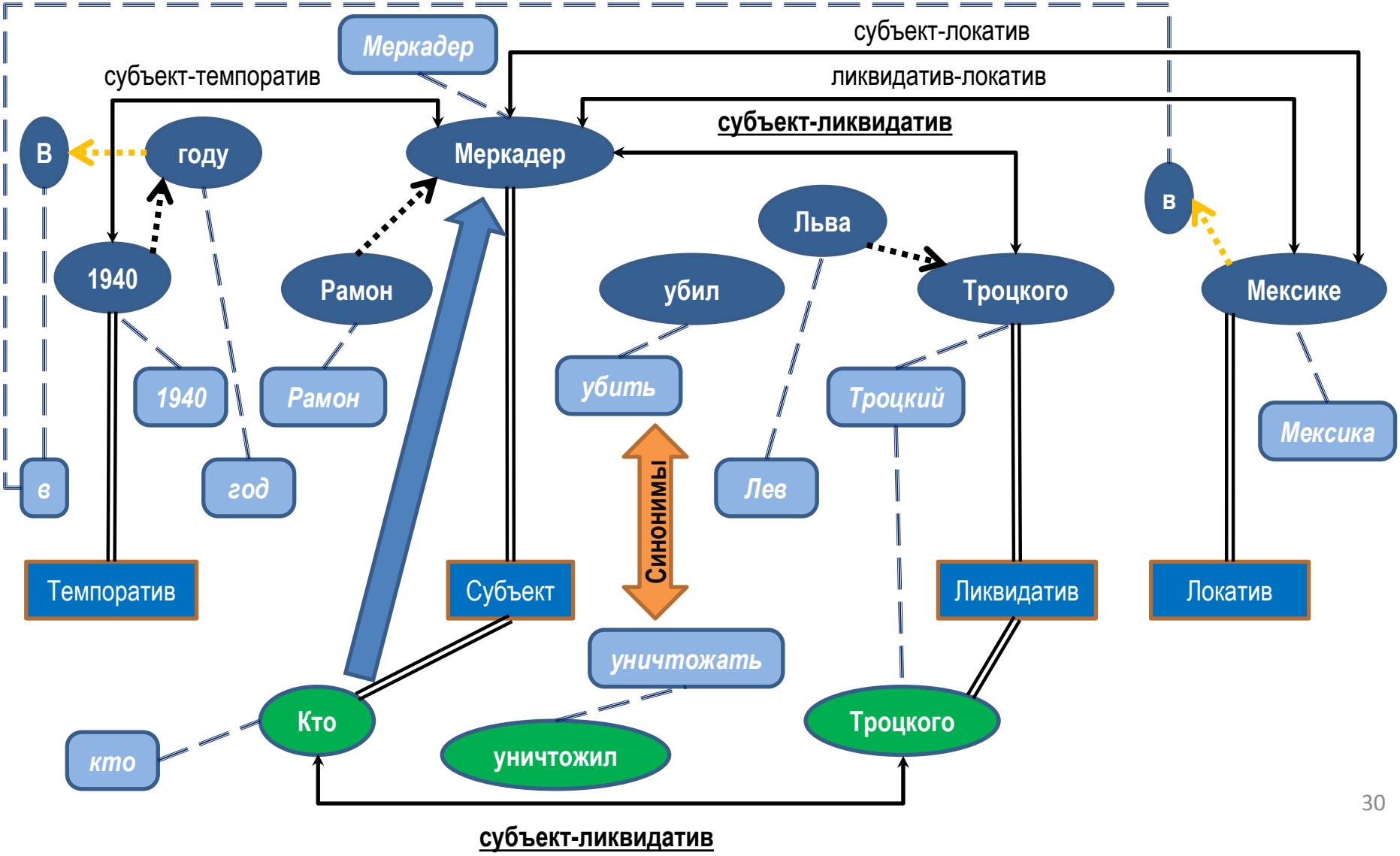


3. Инструменты поиска интеллектуальных методов анализа…



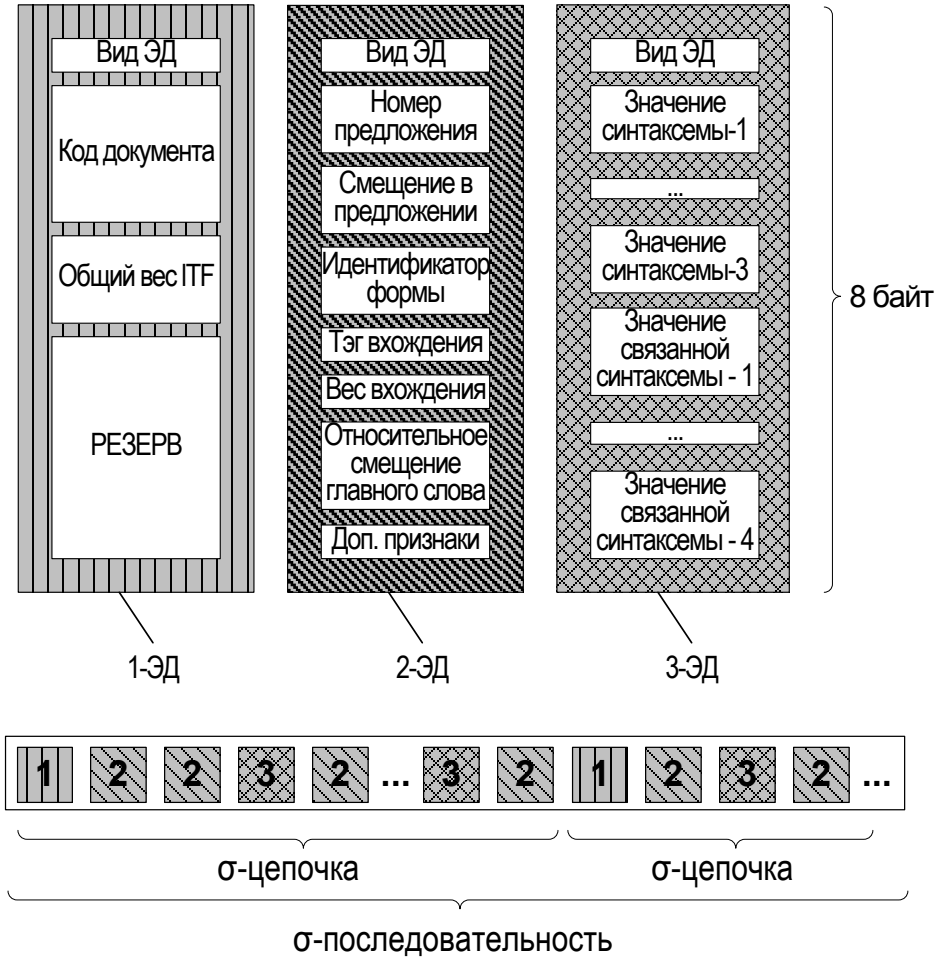
Пример оценки сходства текстов на основе сопоставления графовых структур

В 1940 году Рамон Меркадер убил Льва Троцкого в Мексике. | **Кто уничтожил Троцкого?**



Реляционно-ситуационный инвертированный индекс (Соченков 2013)

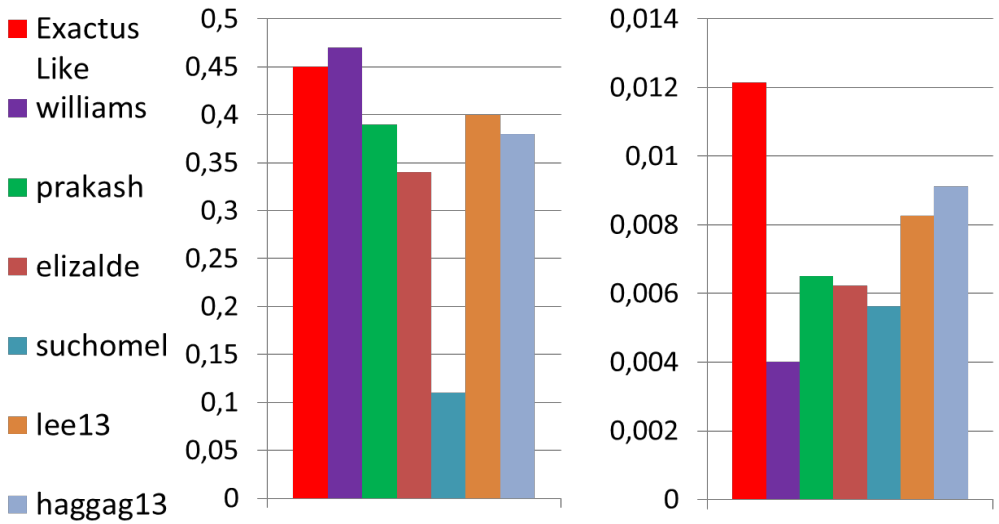
- Элементы данных (σ-ЭД) нескольких типов (1-ЭД, 2-ЭД, 3-ЭД) представляют информацию о словоупотреблениях в текстах документов
- σ-ЭД упорядочены в «пост-листах» определённым образом – σ-последовательности
- 3-ЭД опциональны (если поля данных в этих ЭД имеют значения по умолчанию)



Качество информационного поиска

Соревнования методов и систем:

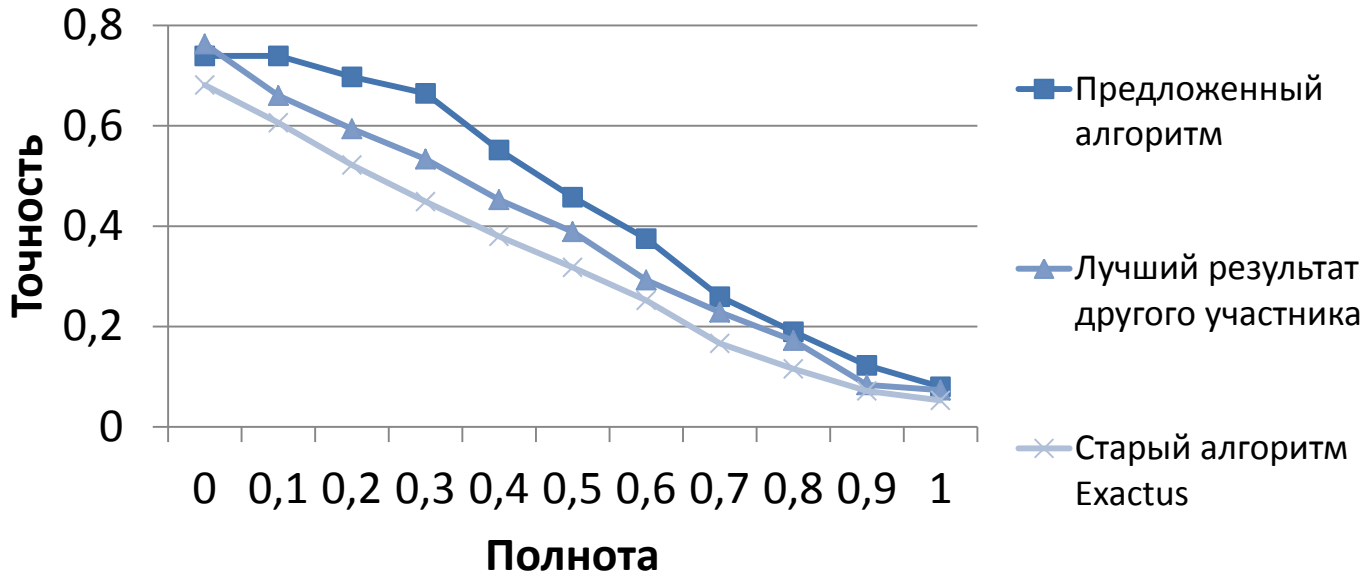
- ✓ РОМИП,
- ✓ SemEval,
- ✓ Dialog-21,
- ✓ TREC,
- ✓ CLEF



Эксперименты по оценке качества патентного поиска при экспертизе заявки по существу:

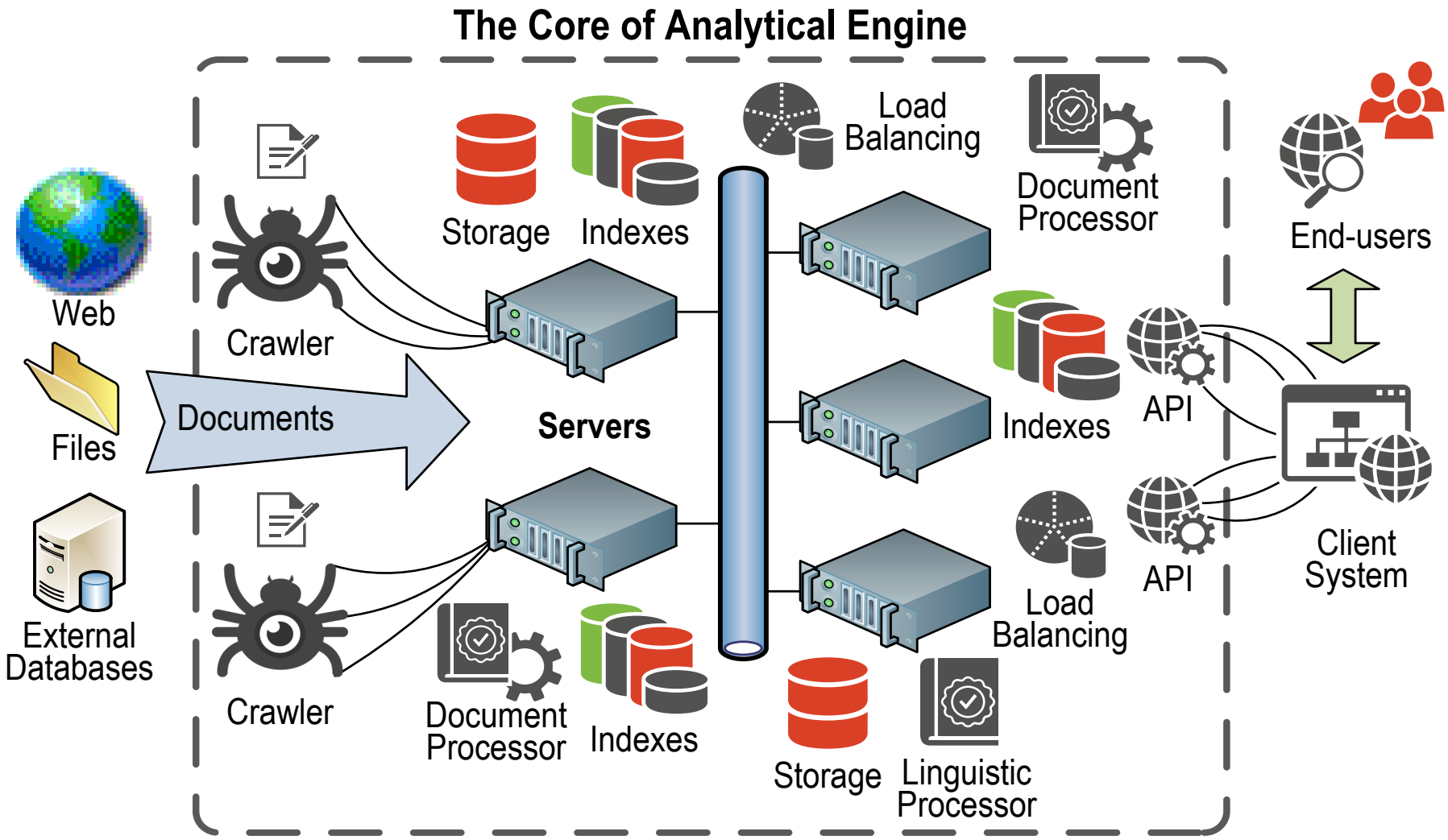
Метод	nDCG	Полнота	D-10	D-20	Медиана
Базовый метод ФИПС	0,141	0,414	0,198	0,262	200
TextAppliance	0,411	0,751	0,451	0,502	20

Оценка качества по методике РОМИП



ХАРАКТЕРИСТИКА	ЗНАЧЕНИЕ
Глубина пула оцениваемых документов (для каждого запроса)	100
Найдено релевантных документов: F_{rel}	1769
Найдено неоценённых документов: $F_{n/a}$	5079
Всего найдено документов: N_F	9100
$F_{n/a} / N_F$	0,56
$F_{rel} / F_{n/a}$	0,35
Обобщённая полнота на глубине пула:	0,49
Средняя точность	0,45
nDCG	0,69

Архитектура поисковой машины Exactus Expert



Выявление новых событий (Devyatkin et al, 2018)

- Небольшие временные периоды (1 день)
- Многомодальная тематическая модель с аддитивной регуляризацией (на основе библиотеки BigARTM)
- Каждый токен имеет модальность (т.е. тип или класс: дата, локация, персона итп.) (Lanina, 2017)
- Сопоставление упоминаний событий в различных источниках данных (например Facebook и Twitter)
- Дивергенция Дженсона-Шеннона между распределениями лексики в тематиках

$$JSD(\Phi_t||\Phi_s) = \frac{1}{2}\big(KL(\Phi_t||M) + KL(\Phi_s||M)\big),$$
$$M = \frac{1}{2}(\Phi_t + \Phi_s).$$

Выявление новых событий: результаты

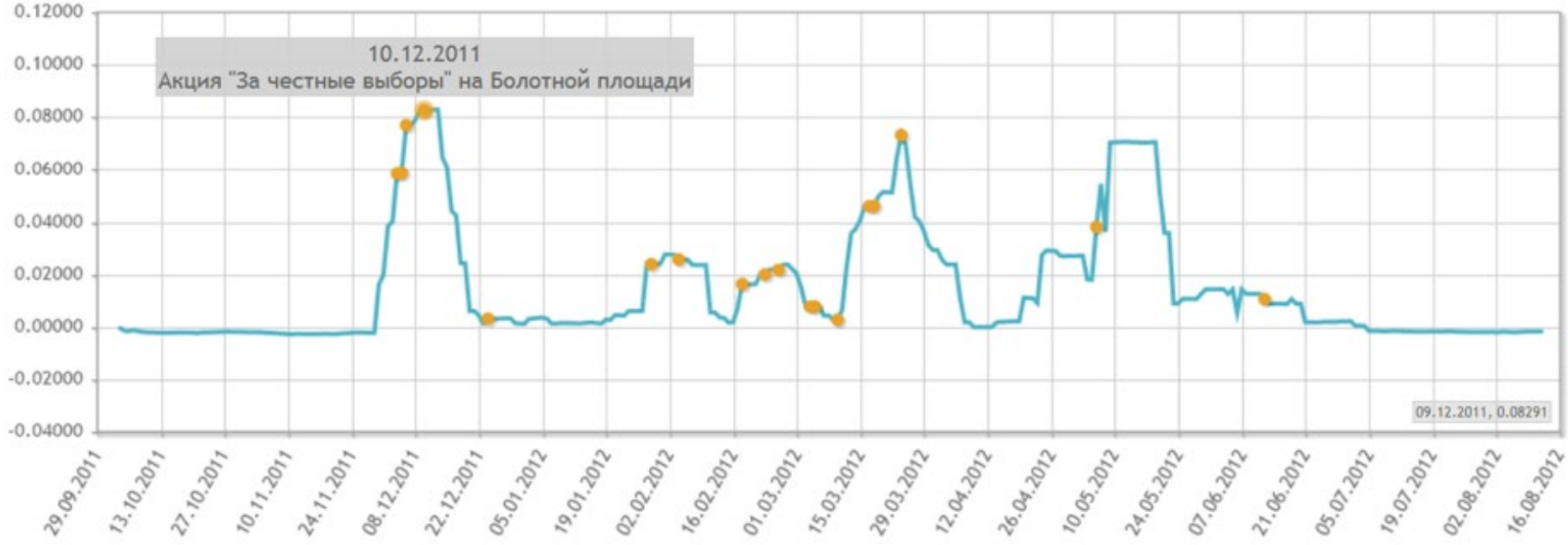
60 тыс. сообщений в Twitter с 1 по 12 апреля 2018

Точность, %

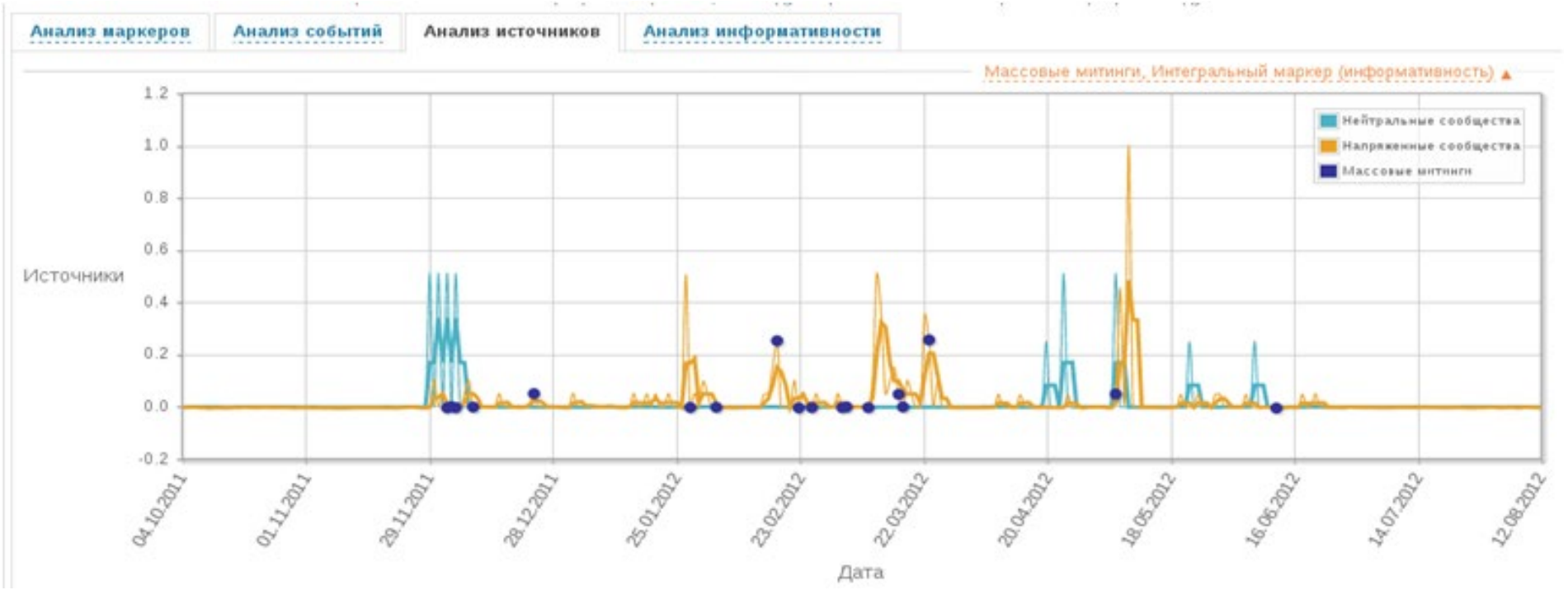
Шаг	Базовый метод (LDA)	Многомодальная модель
Все события	63.3	**93.3**
Новые события	71.4	**80.0**

Пример: динамика эмоциональной напряженности сетевых сообществ

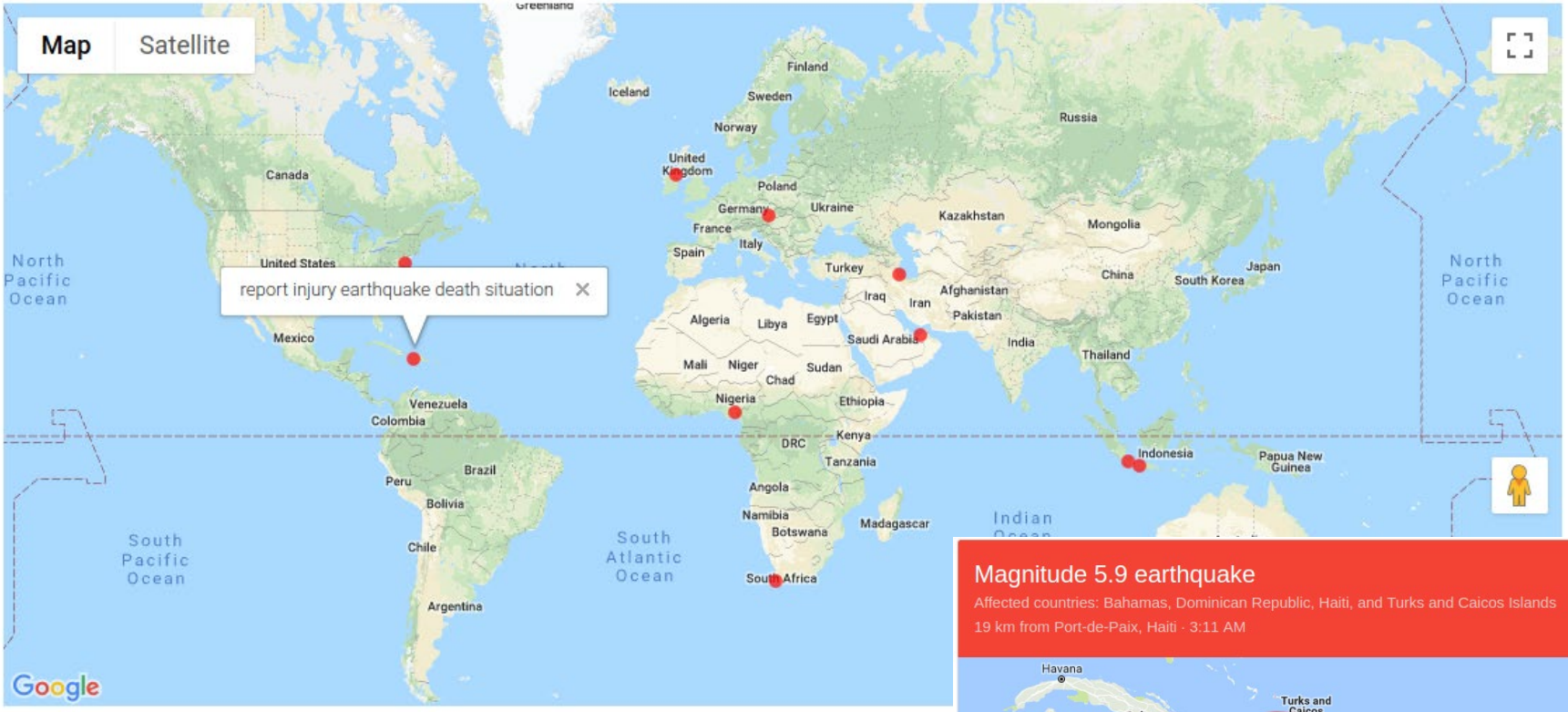
- Значения интегрального маркера напряженности по сообщениям Живого Журнала 2011-2012 гг.



Пример: сравнение степени эмоциональной напряженности сетевых сообществ



Пример: мониторинг чрезвычайных происшествий



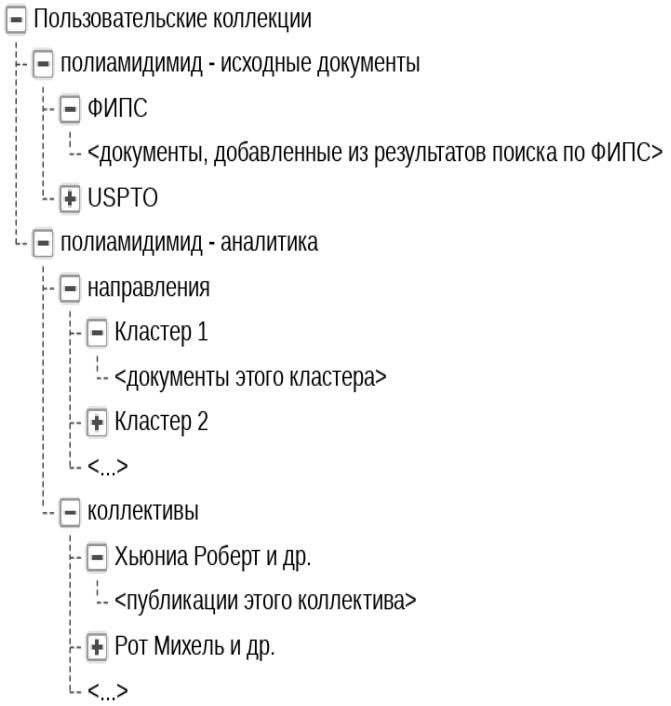
Твиты о теме:

Jacqueline Charles
@Jacquiecharles

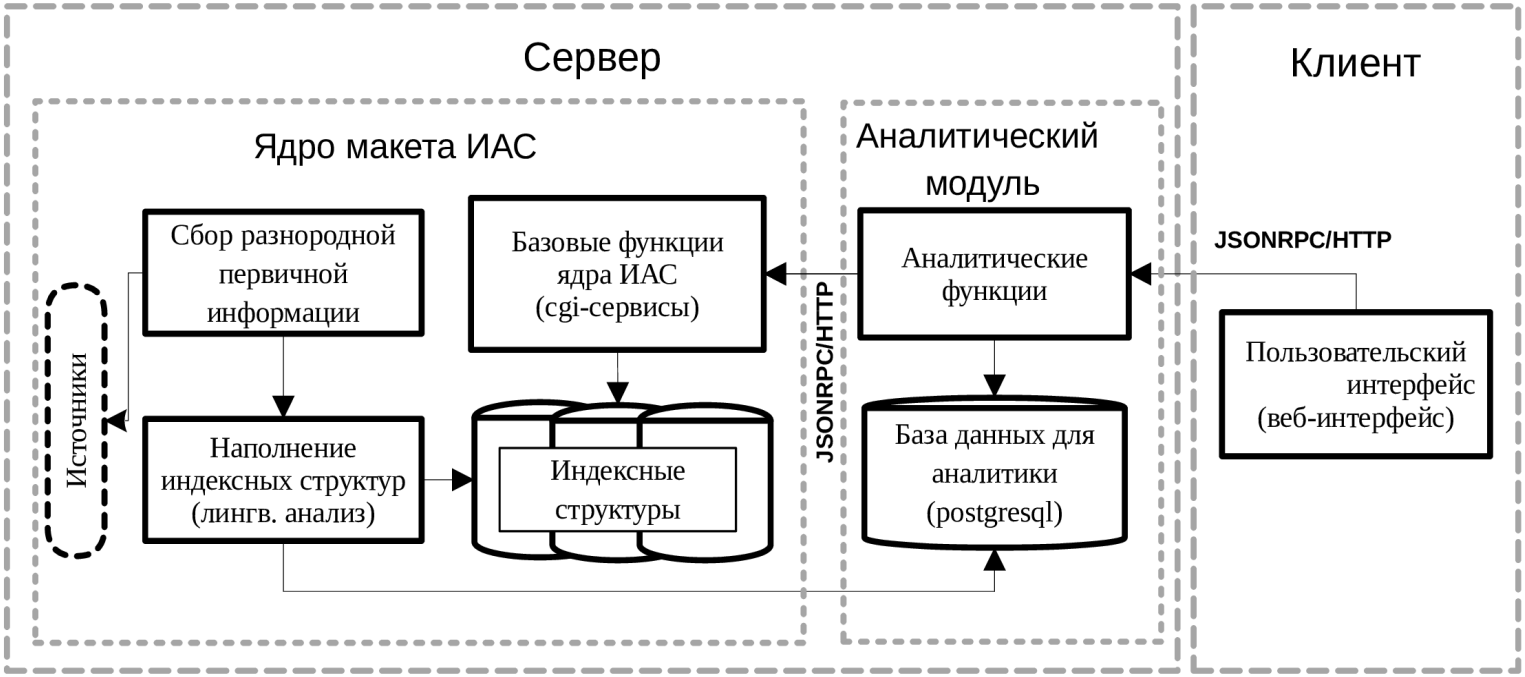
Latest situation report from @Pwoteksyonsivil on preliminary deaths & injuries on #Haiti #earthquake reports 10 confirmed dead, 135 injured . There were also two more fatalities confirmed during the night before report was issued. miamiherald.com/news/nation-wo…

Пример: информационно-аналитическая система «Приоритеты»

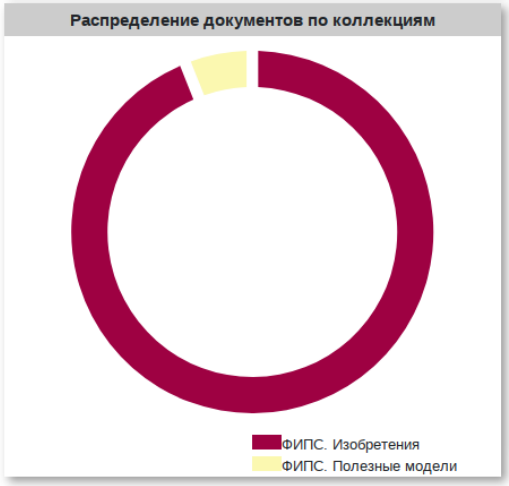
- Более 10 млн. научно-технических документов.
- Полнотекстовый поиск с учетом синтаксиса и семантики.
- Поиск тематически похожих документов.
- Построение резюме документа.
- Построение ключевых слов документа или коллекции.
- Поиск текстовых заимствований.
- Пользовательские коллекции.
- Кластеризация коллекций, выявление авторских коллективов.
- Агрегированная статистика в графической форме.

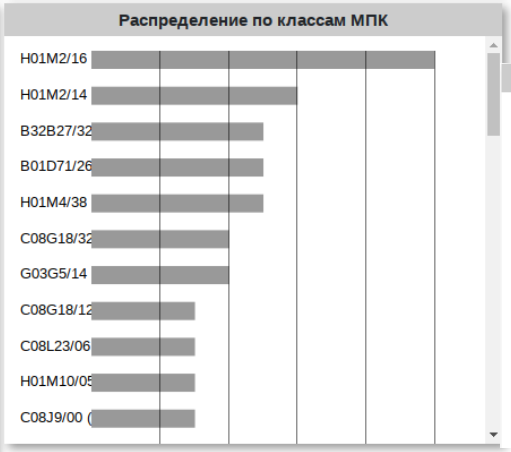


Пример: информационно-аналитическая система «Приоритеты»

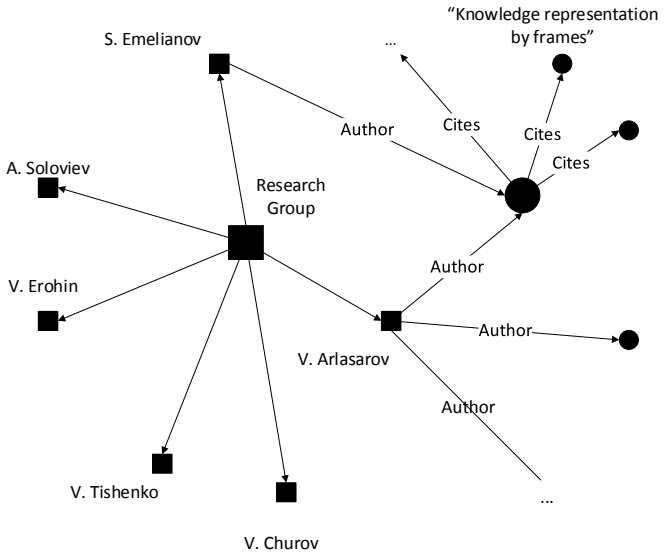


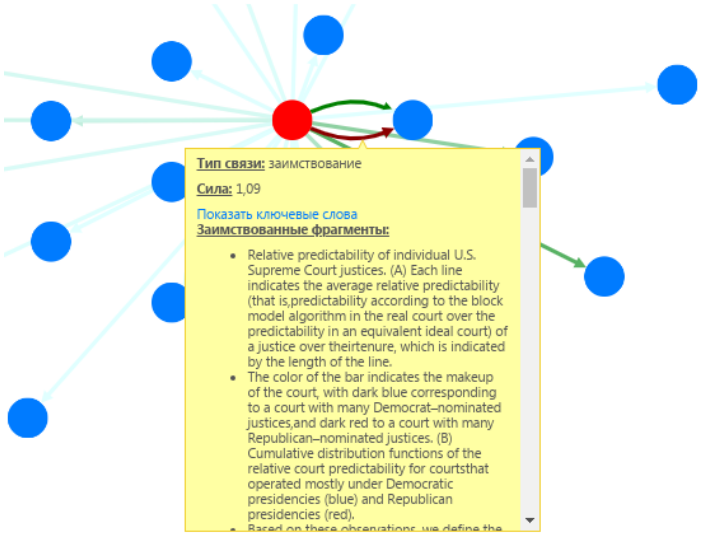
Пример: информационно-аналитическая система «Приоритеты»





Пример: информационно-аналитическая система «Приоритеты»





Направления развития

1. Совершенствование методов сбора данных, лингвистического анализа, извлечения информации и информационного поиска.
2. Создание и развитие размеченных корпусов для обучения методов анализа текстов.
3. Разработка методического обеспечения систем поддержки принятия решений.
4. Разработка методов и подходов к верификации результатов работы пользователя.

Контакты

Девяткин Дмитрий Алексеевич

ФИЦ ИУ РАН

devyatkin@isa.ru

Источники

- Shelmanov A. O., Smirnov I. V. Methods for semantic role labeling of Russian texts //Computational Linguistics and Intellectual Technologies. Proceedings of International Conference Dialog. – 2014. – Т. 13. – №. 20. – С. 607-620.
- Nivre J., Hall J., Nilsson J. Maltparser: A data-driven parser-generator for dependency parsing //Proceedings of LREC. – 2006. – Т. 6. – С. 2216-2219.
- Al-Rfou R. et al. Polyglot-NER: Massive multilingual named entity recognition //Proceedings of the 2015 SIAM International Conference on Data Mining. – Society for Industrial and Applied Mathematics, 2015. – С. 586-594.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119, 2013.
- Anastasia Ianina, Lev Golitsyn, and Konstantin Vorontsov. Multi-objective topic modeling for exploratory search in tech news. Conference on Artificial Intelligence and Natural Language, pages 181–193, Springer, 2017.
- Osipov G. et al. Relational-situational method for intelligent search and analysis of scientific publications //Proceedings of the Integrating IR Technologies for Professional Search Workshop. – 2013. – С. 57-64.
- Соченков И. В., Суворов Р. Е. Сервисы полнотекстового поиска в информационно-аналитической системе (Часть 1) //Информационные технологии и вычислительные системы. – 2013. – №. 2. – С. 69.
- Ананьева М. И. и др. Автоматическое извлечение финансово-экономической информации из текстов на русском языке //Труды Института системного анализа Российской академии наук. – 2018. – Т. 68. – №. 1. – С. 23-30.
- Vybornova O. et al. Social tension detection and intention recognition using natural language semantic analysis: On the material of Russian-speaking social networks and Web forums //Intelligence and Security Informatics Conference (EISIC), 2011 European. – IEEE, 2011. – С. 277-281.
- Baranov A. A. et al. Technologies for Complex Intelligent Clinical Data Analysis //Vestnik Rossiiskoi akademii meditsinskikh nauk. – 2016. – №. 2. – С. 160-171.
- Devyatkin D., Shelmanov A. Text Processing Framework for Emergency Event Detection in the Arctic Zone //International Conference on Data Analytics and Management in Data Intensive Domains. – Springer, Cham, 2016. – С. 74-88.
- Gennady Osipov, Ivan Smirnov, Ilya Tikhomirov, Ilya Sochenkov, and Artem Shelmanov. Exactus expert”search and analytical engine for research and development support. In Novel Applications of Intelligent Systems, pages 269–285. Springer, 2016.
- Mitchell, M.L., Mulherin, J.H.: The Impact of Public Information on the Stock Market. Journal of Finance 49 (3) (1994) 923–950
- Abdullah S. S., Rahaman M. S., Rahman M. S. Analysis of stock market using text mining and natural language processing //Informatics, Electronics & Vision (ICIEV), 2013 International Conference on. – IEEE, 2013. – С. 1-6.
- Agarwal R. et al. Fast algorithms for mining association rules //Proc. of the 20th VLDB Conference. – 1994. – С. 487-499.
- Abbas A., Zhang L., Khan S. U. A literature review on the state-of-the-art in patent analysis //World Patent Information. – 2014. – Т. 37. – С. 3-13.