# Интеллектуальный анализ и тематическое моделирование проблемно-ориентированного корпуса медицинских текстов

Докладчик: к.т.н., доцент Гаянова М.М.

Научный консультант: д.т.н., проф. Юсупова Н.И.

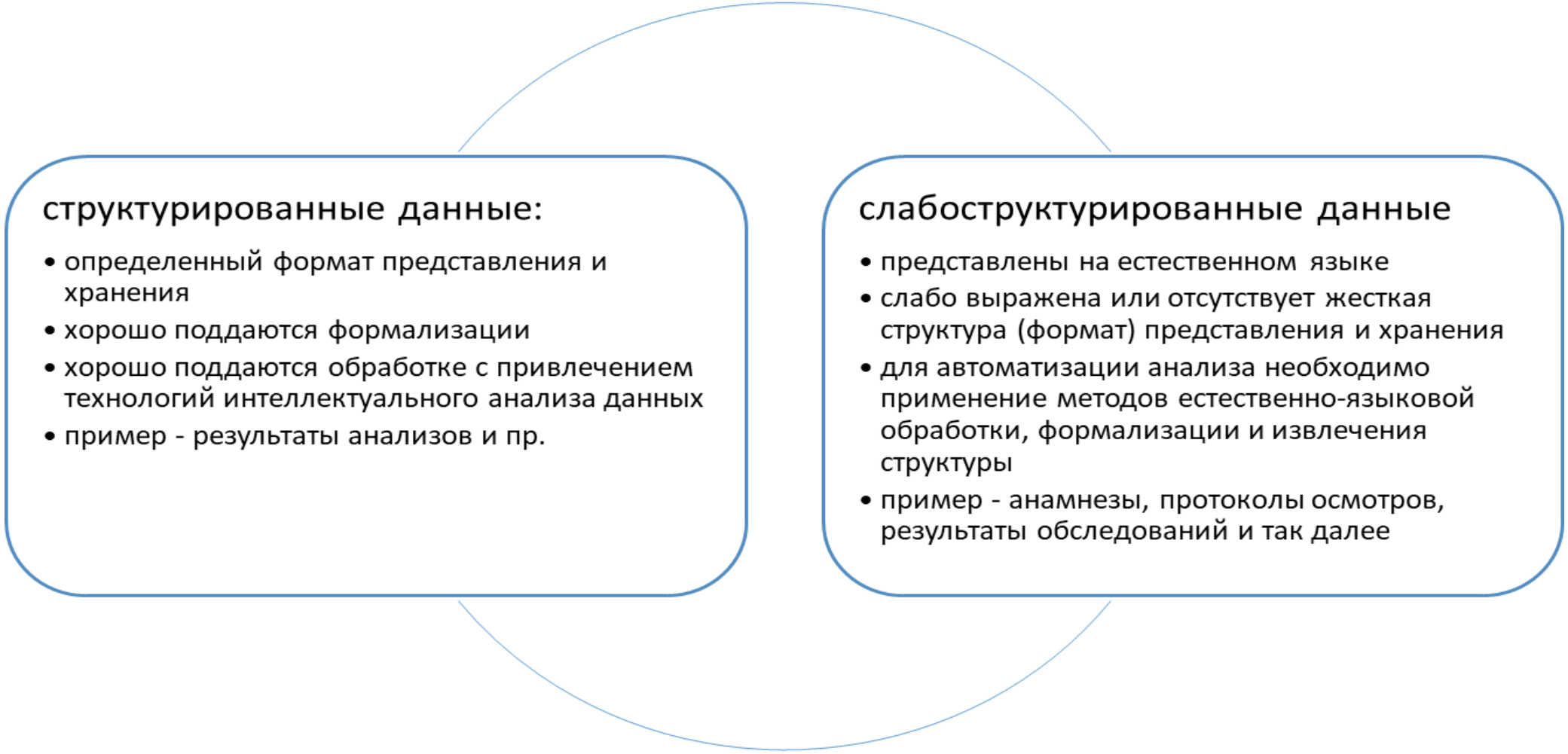
# Цель и задачи

**Цель:** интеллектуальный анализ и тематическое моделирование проблемно-ориентированного корпуса текстовых документов на русском языке на примере научных публикаций медицинской тематики

**Задачи:**

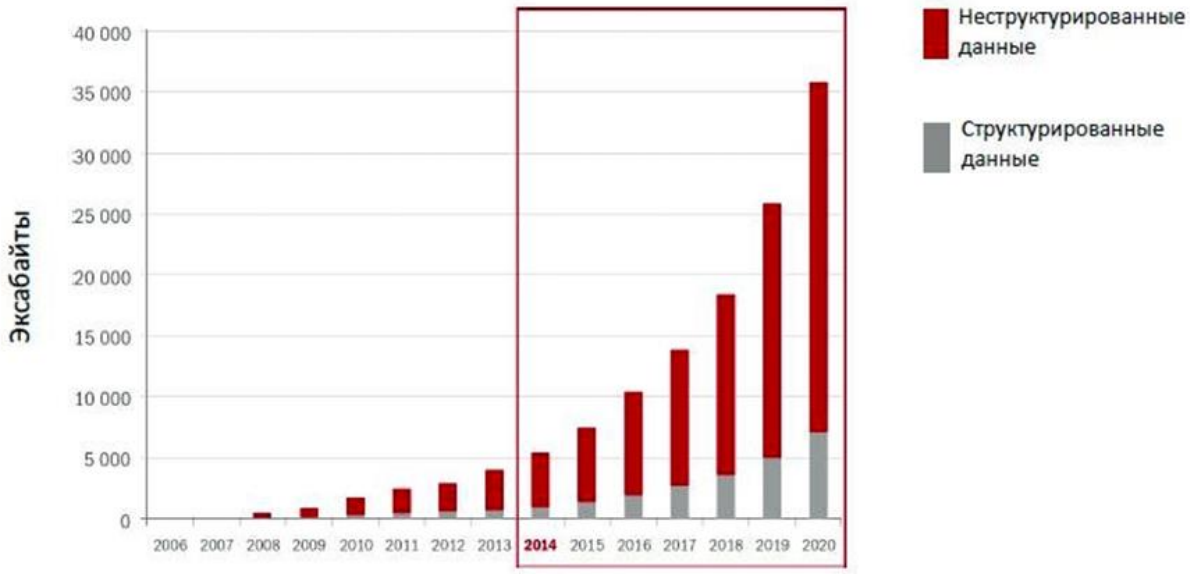
- разработка структурно-функциональной организации системы извлечения знаний и автоматического построения онтологии и графа знаний проблемно-ориентированного корпуса для конкретной предметной области с целью последующего построения системы поддержки принятия решений при анализе клинических рекомендаций;
- анализ корпуса медицинских текстов заданной тематики (пульмонология, история болезней пациентов) в задаче извлечения знаний из неразмеченного корпуса

# Проблема и актуальность исследования

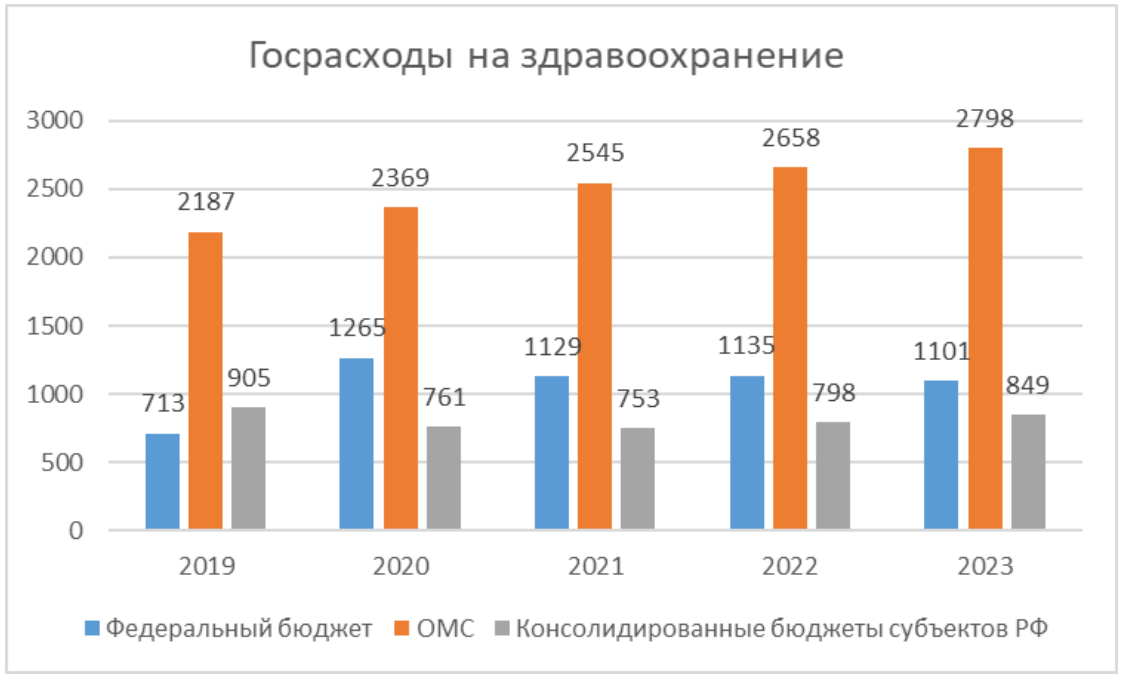


# Проблема и актуальность исследования

## Рост цифровой информации в мире*




*По данным исследовательской компании IDC.



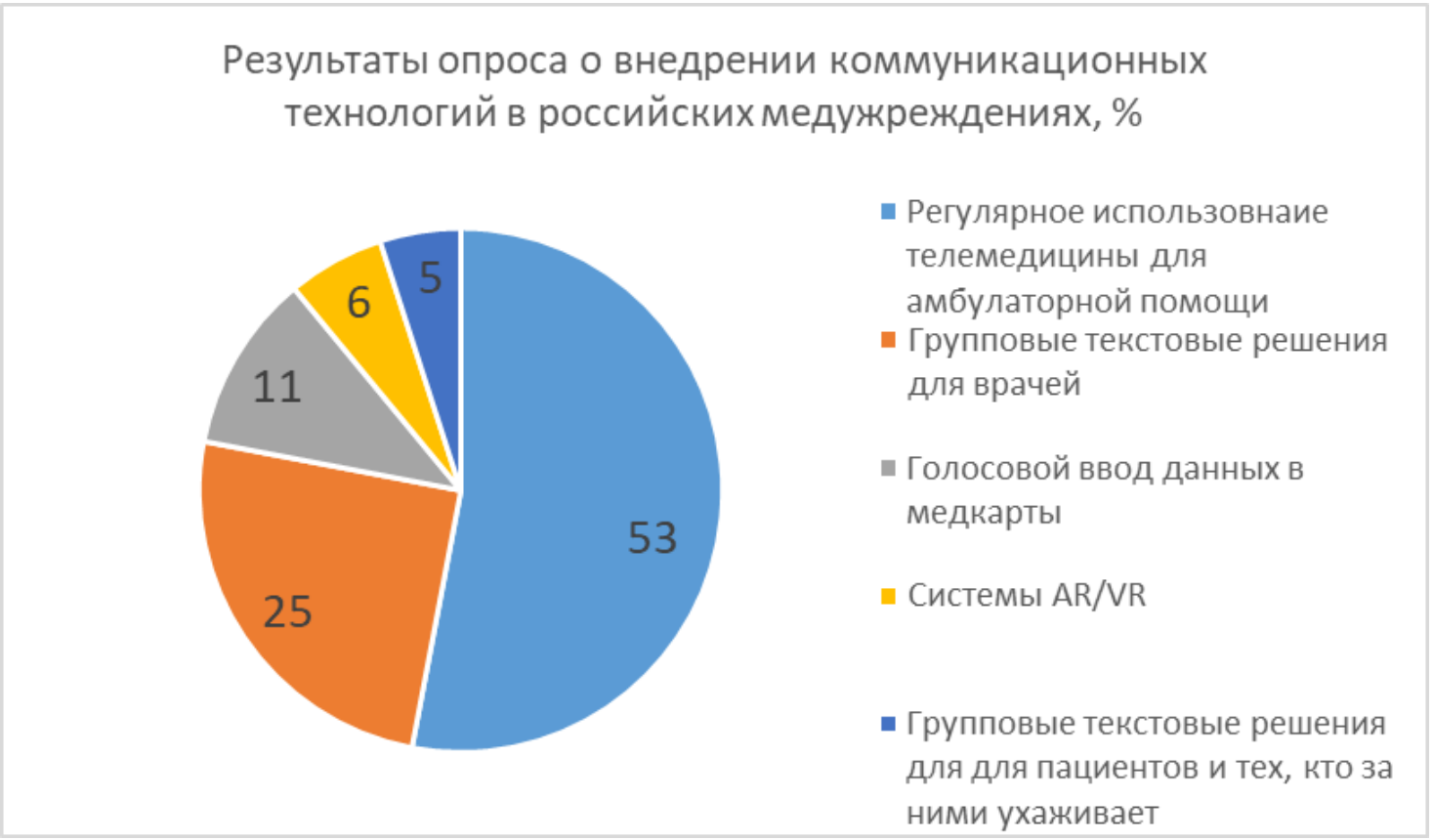

Актуальным является работа с «большими данными» в медицинской практике – оперативный анализ (сбор, хранение, формализация, постоянное обновление, анализ, интерпретация) с целью создания регулярно пополняемых баз – клинических регистров

# Проблема и актуальность исследования



Интеллектуальный анализ клинических текстов и извлечение знаний из накопленных массивов периодически меняющихся данных является одним из перспективных научных направлений на стыке компьютерной лингвистики, машинного обучения и медицины, направленных на решение данной проблемы.

# Проблема и актуальность исследования





На сегодняшний день существует много технологий лингвистического анализа текстов, но анализа текста на уровне только лингвистических правил недостаточно для корректного извлечения фактов из корпуса медицинских документов. Для эффективного извлечения фактов из текста база знаний должна содержать информацию, включающую медицинские онтологии, классификаторы, систематизированные знания в области анатомии, физиологии и патофизиологии человека.

Данные с сайта EverCare.ru – новости и события мира телемедицины

# Подходы к автоматизации построения онтологий

1. Распознавание/ извлечение именованных сущностей (Named Entity Recognition/Extraction)

2. Связывание/снятие омонимии сущностей, или семантическое аннотирование (Entity Linking/Disambiguation, Semantic Annotation)

3. Извлечение терминов (Term Extraction)

4. Извлечение ключевых слов/фраз (Keyword/Keyphrase Extraction)

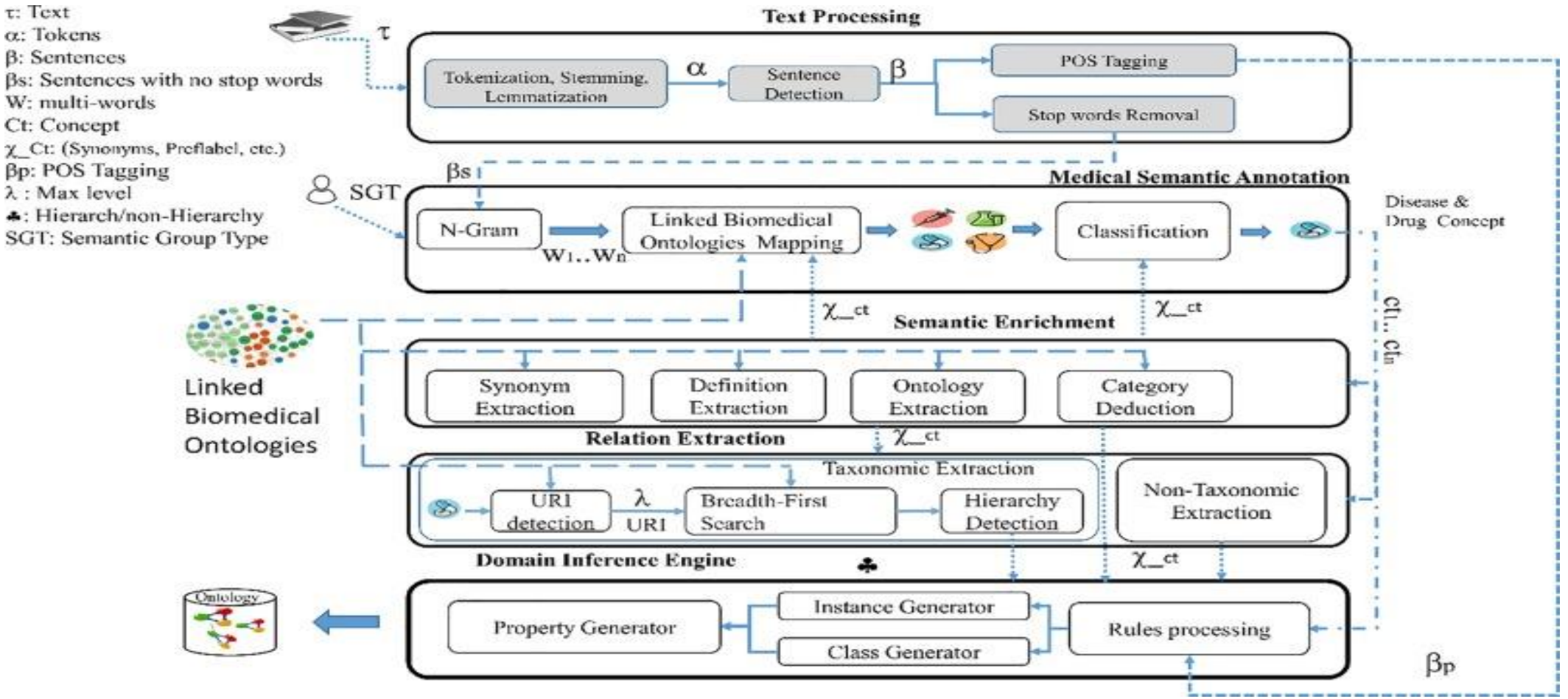
5. Тематическое моделирование/классификация (Topic Modeling, Classification)

6. Маркирование/ идентификация темы (Topic Labeling/ Identification)

7. Извлечение отношений (Relation Extraction)

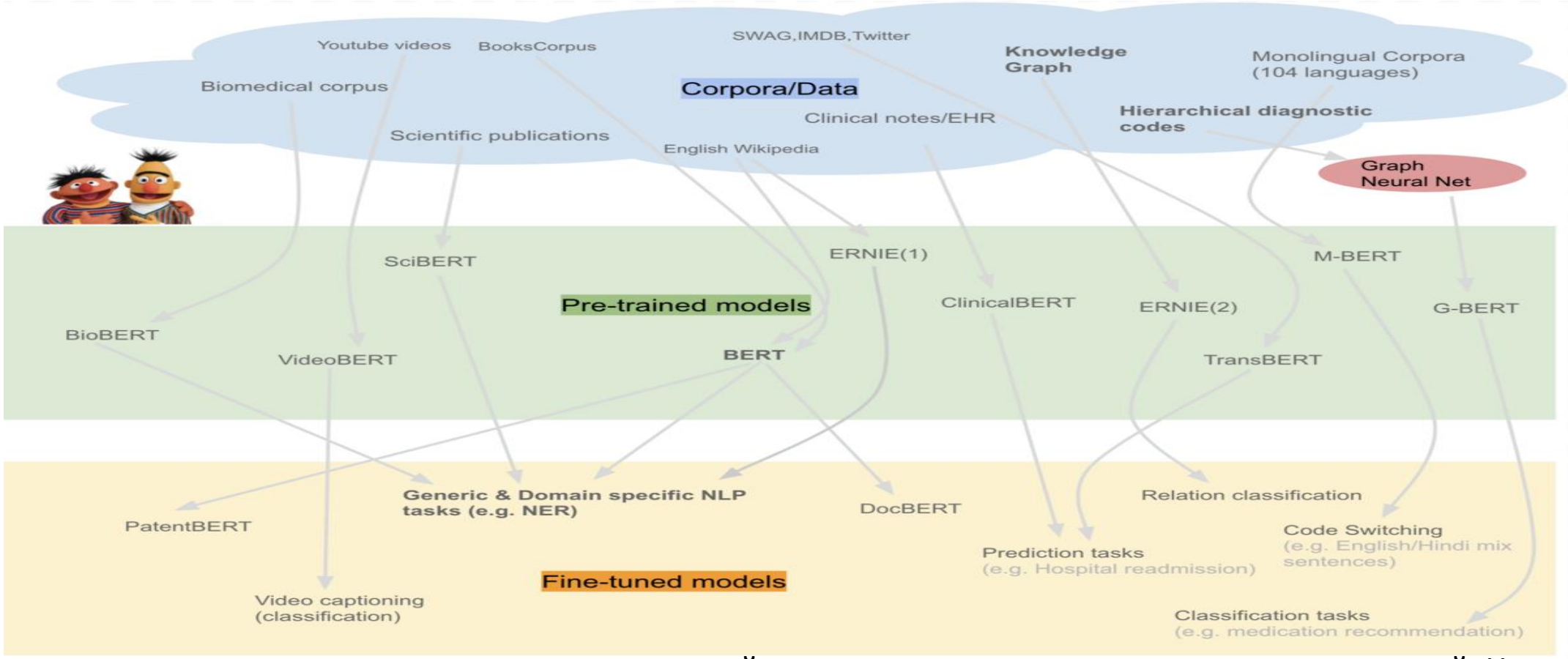
Перспективным является подход по извлечению знаний непосредственно из собираемых данных с помощью интеллектуальных алгоритмов, когда роль человека-эксперта сводится к верификации автоматически построенных онтологических моделей («обучение онтологий»)

# Подходы к автоматизации построения онтологий



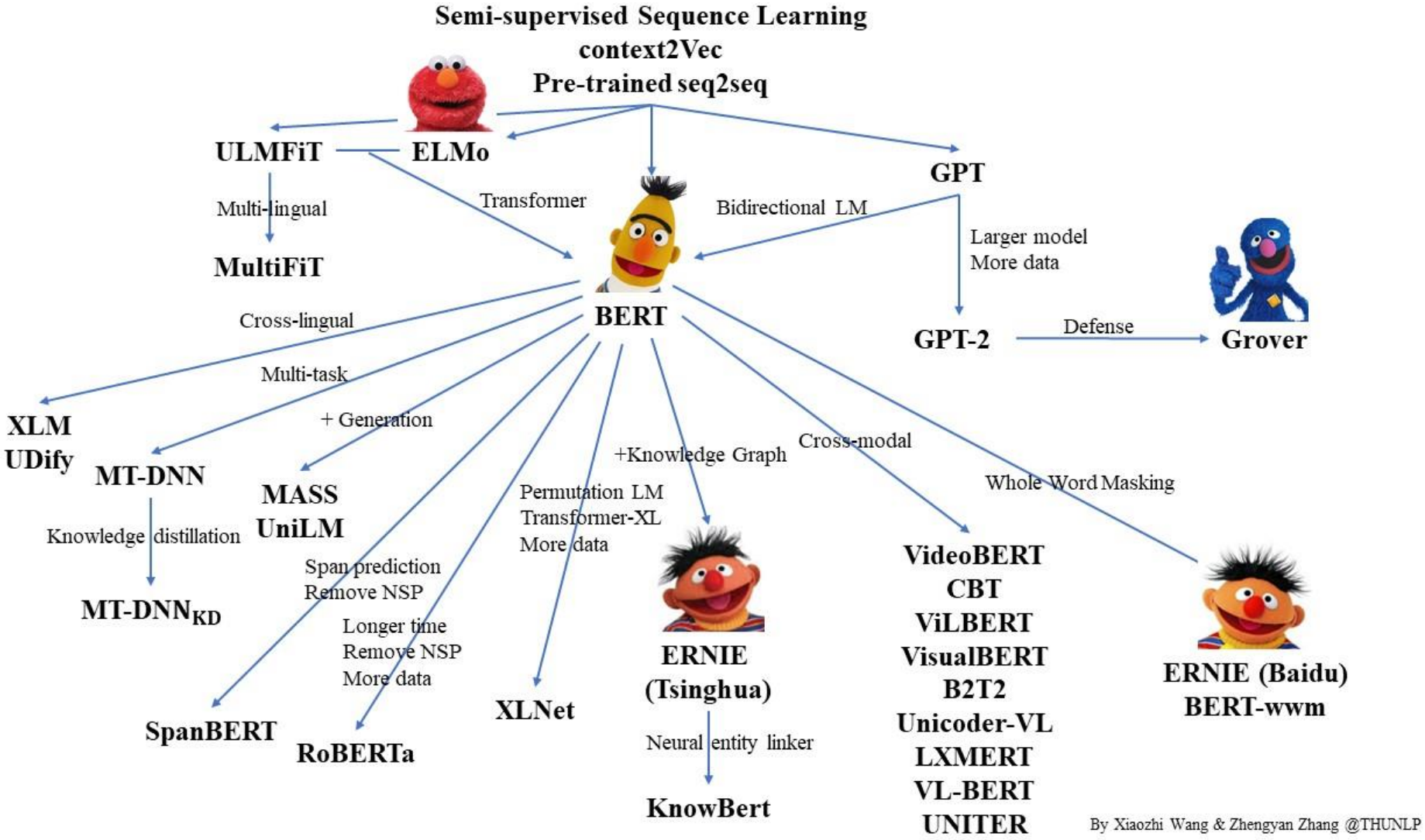

Пример реализованной системы по сбору, предобработке и анализу слабоструктурированных данных в задаче автоматизации построения онтологий и выделения знаний является решение;  предложен финскими исследователями; использован комплекс методов и моделей обработки документов на естественном языке

# Языковые модели в контексте инженерии знаний



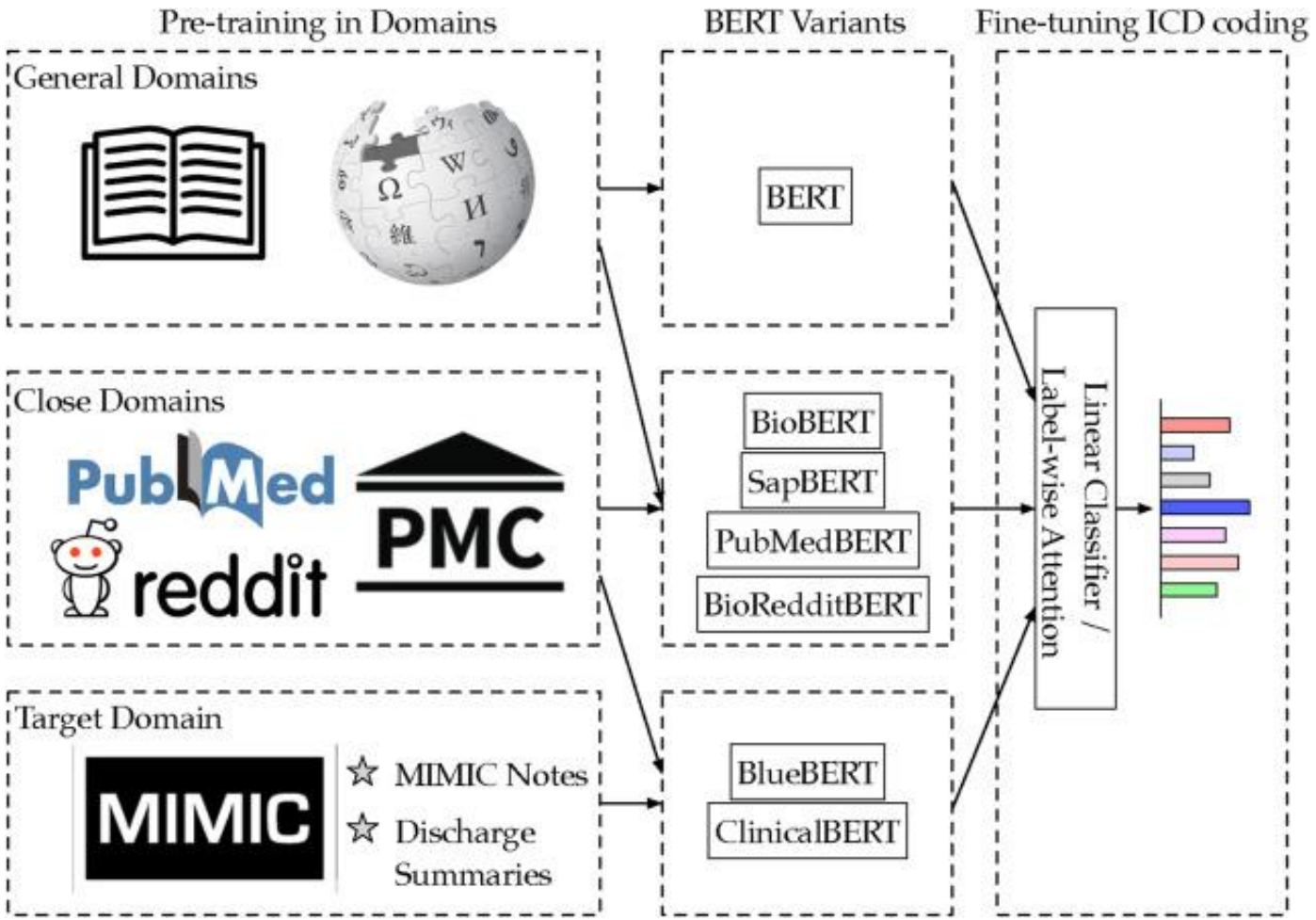

Ключевым этап автоматизации построения онтологий - построение качественных языковых моделей. На сегодняшний день в качестве таких моделей выступают нейросетевые модели на основе технологий глубокого обучения с реализацией механизма внимания – трансформеры семейства BERT. Технологии «переноса обучения» - большая модель обучается на доступном корпусе неспециализированных текстов, а затем дообучается на компактном наборе узкоспециализированных текстов для решения частных задач: классификации, кластеризации, выделения именованных сущностей и отношений. Подобные подходы реализованы для англоязычных корпусов

# Языковые модели в контексте инженерии знаний



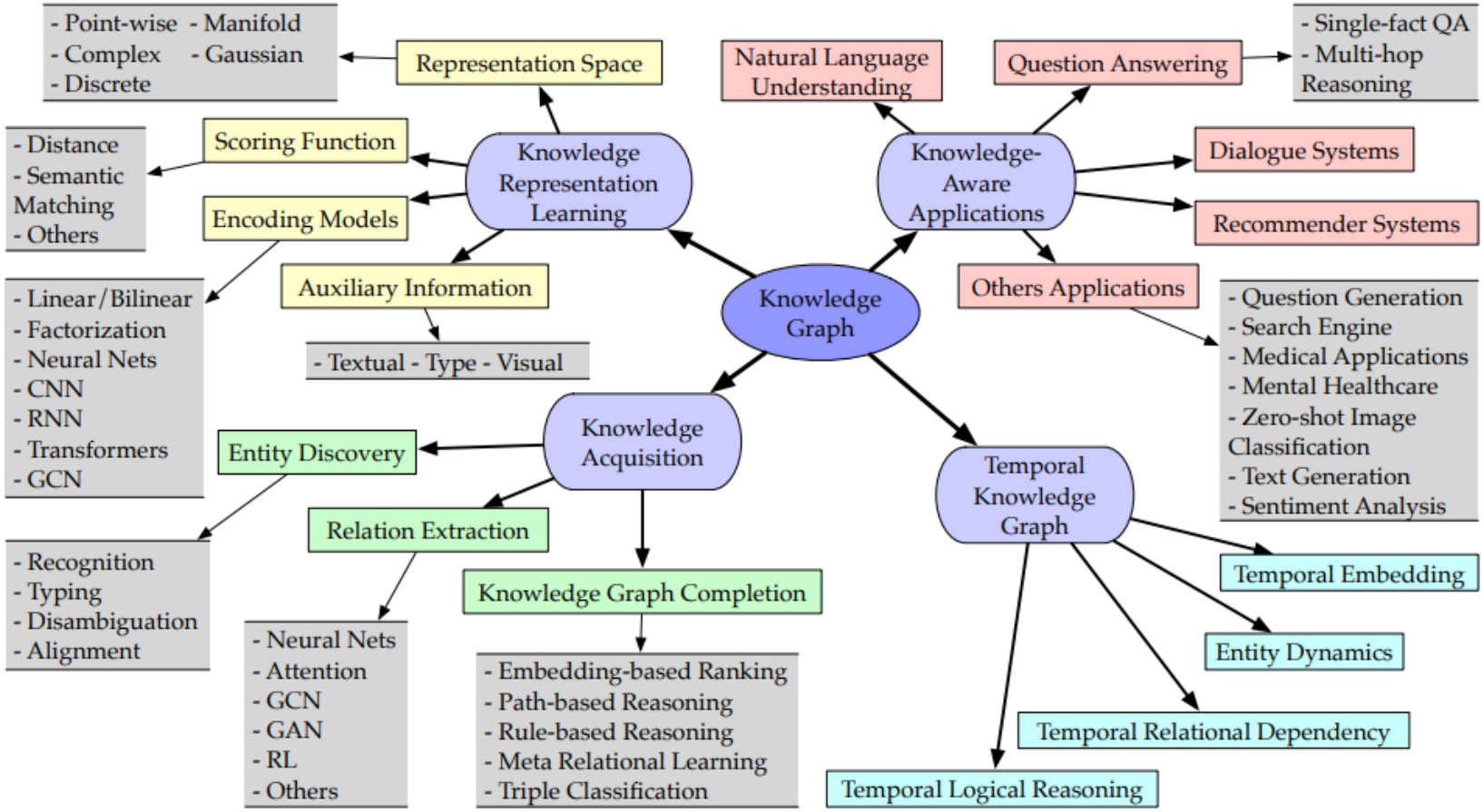

Семейство нейросетевых языковых моделей-трансформеров активно развивается и направлено на снижение требований к вычислительным ресурсам и объемам данных на этапах обучения, а также на повышение качества анализа в конкретных прикладных задачах.

# Языковые модели в контексте инженерии знаний




Примером поэтапного построения нейросетевой языковой модели-трансформера BioBERT является проведенный специалистами корейского университета анализ медицинских научных публикаций на портале pubmed, продемонстрировавшей эффективность подобного подхода в задачах семантической классификации текстов и решении ряда задач по выделению медицинских терминов и названий лекарственных препаратов.

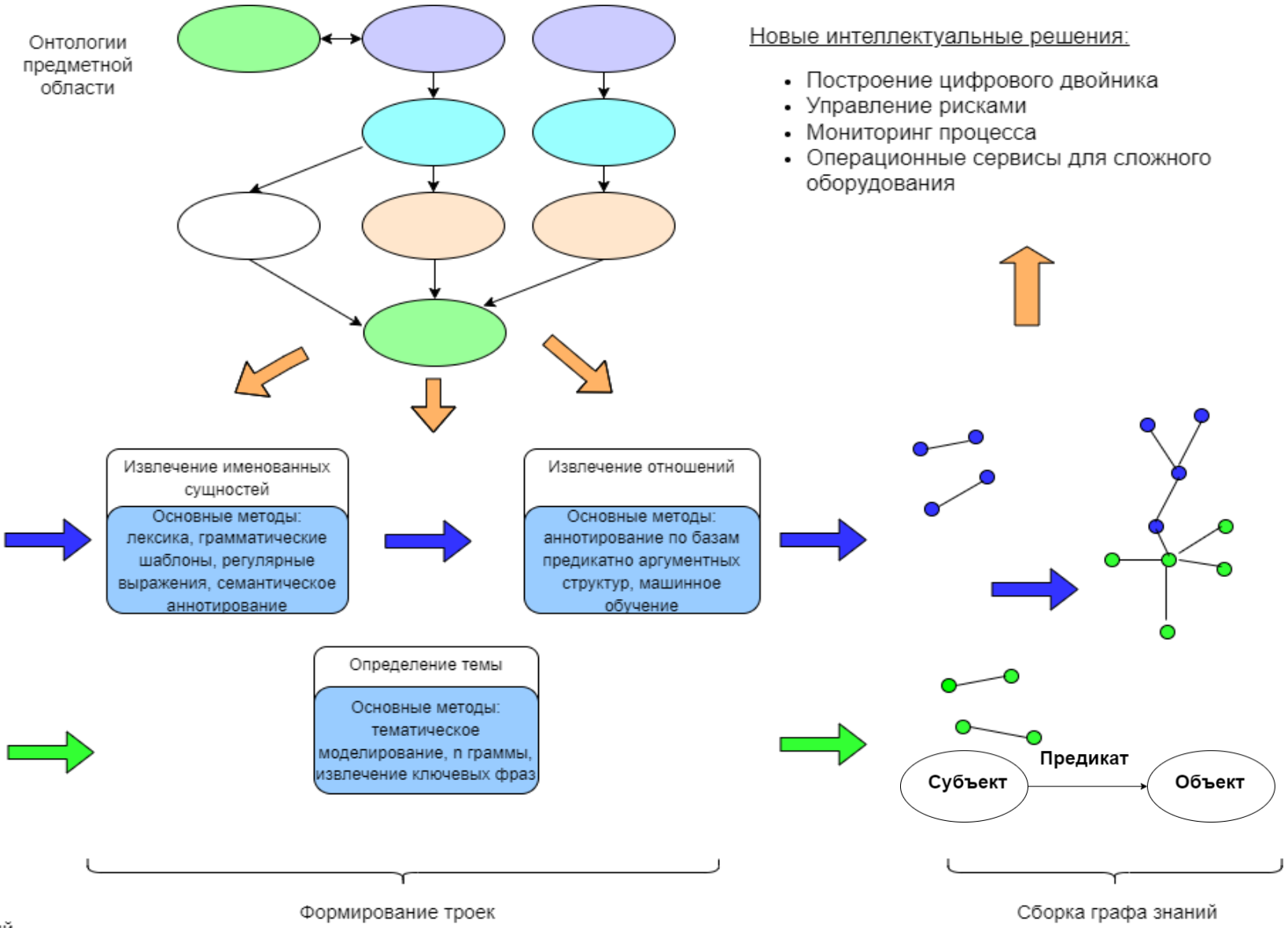
# Граф знаний как инструмент семантического моделирования проблемно-ориентированного корпуса текстов



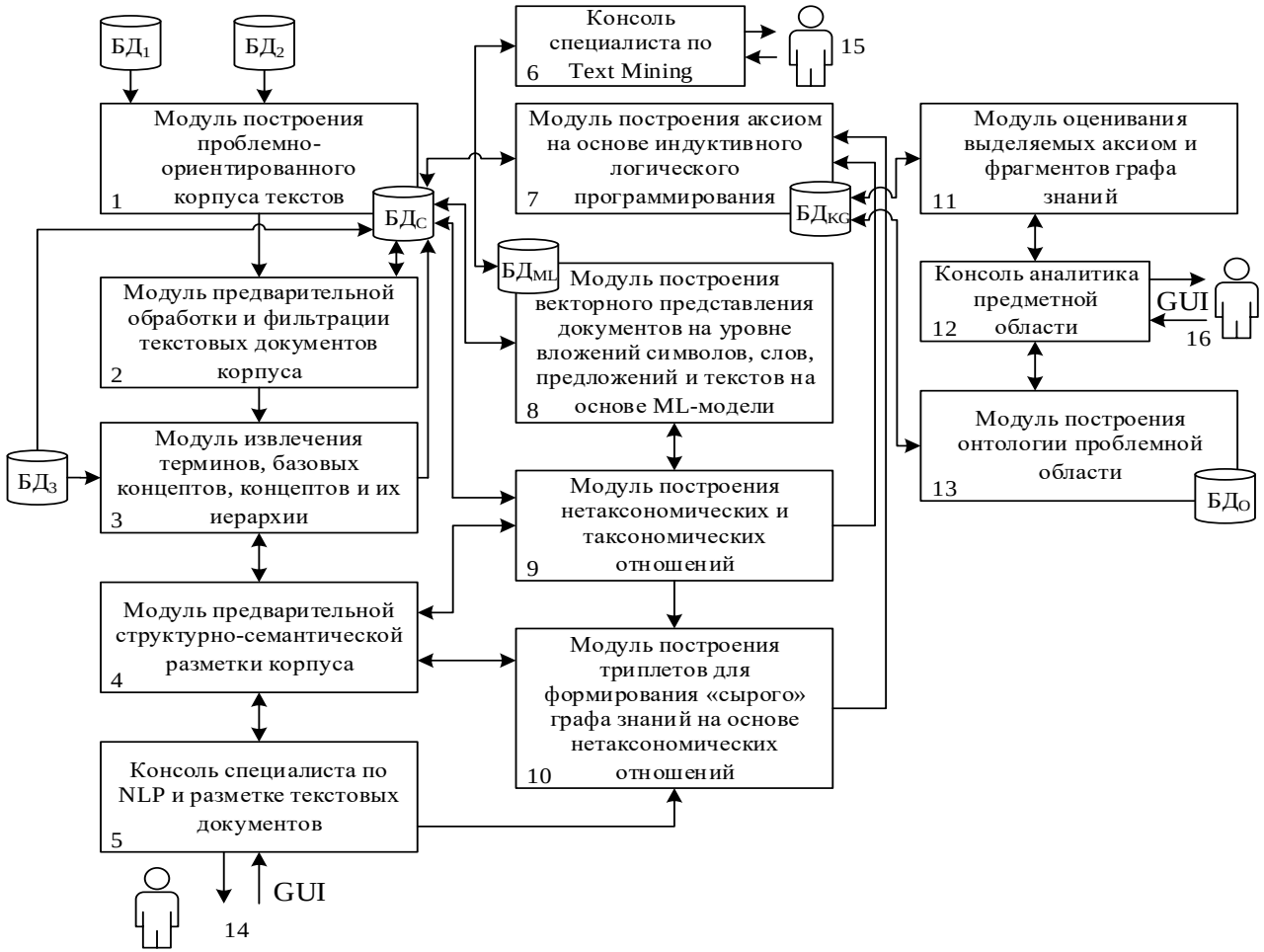

Граф знаний представляет собой структурированное графическое представление семантических знаний и отношений, где узлы в графе представляют сущности, а ребра представляют отношения между ними. Построение графа знаний предполагает извлечение связей из неструктурированного текста с последующим эффективным хранением в граф-ориентированных базах данных.

# Автоматизированный процесс построения графов знаний

Особенностью графа знаний является не только способ представления знаний, но и способ получения новых знаний: «граф знаний собирает и интегрирует информацию в онтологию и применяет подсистему вывода для получения новых знаний».



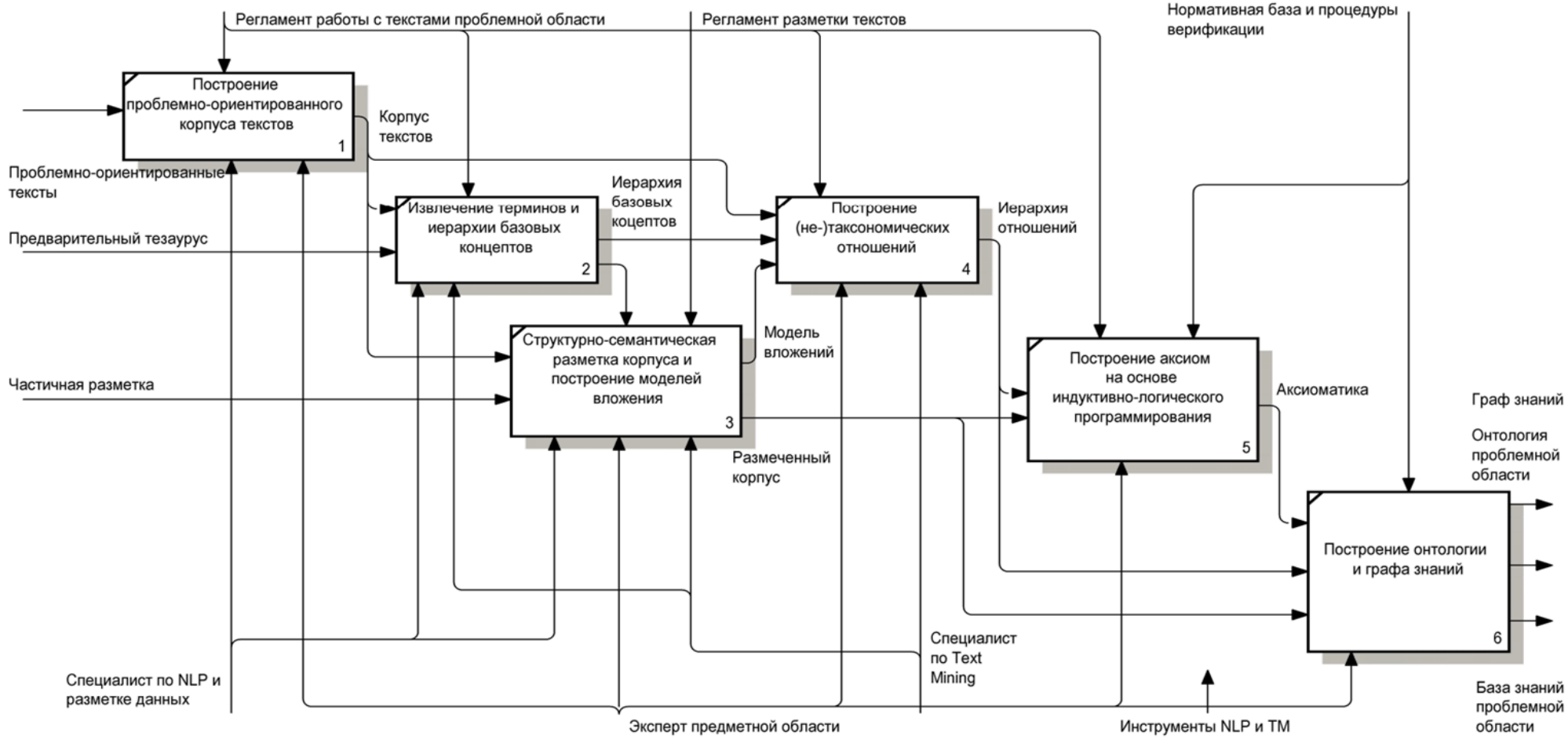
# Структурная схема системы автоматизированного построения онтологий и графов знаний




БД1 — специализированные тексты – клинические описания и истории болезней пациентов, включая: эпикризы, результаты функциональной диагностики, осмотров и рекомендации врачей по лечению;
БД2 — описания клинических рекомендаций по лечению;
БД3 — граф-ориентированная база данных (специалист корректирует предварительно выделенные термины и иерархию концептов);
БДML — построение векторного представления с помощью дообучаемых моделей машинного обучения, под контролем специалиста по анализу и извлечению знаний из корпуса;
БДKG — граф-ориентированная БД, предназначенной для хранения графа знаний;

Отличием системы является применение специализированных нейросетевых моделей (в том числе, для русскоязычных текстов) для автоматизации выделения триплетов, их предобработки и семантической фильтрации.

# Функциональная модель процесса автоматизированного построения онтологии и графа знаний

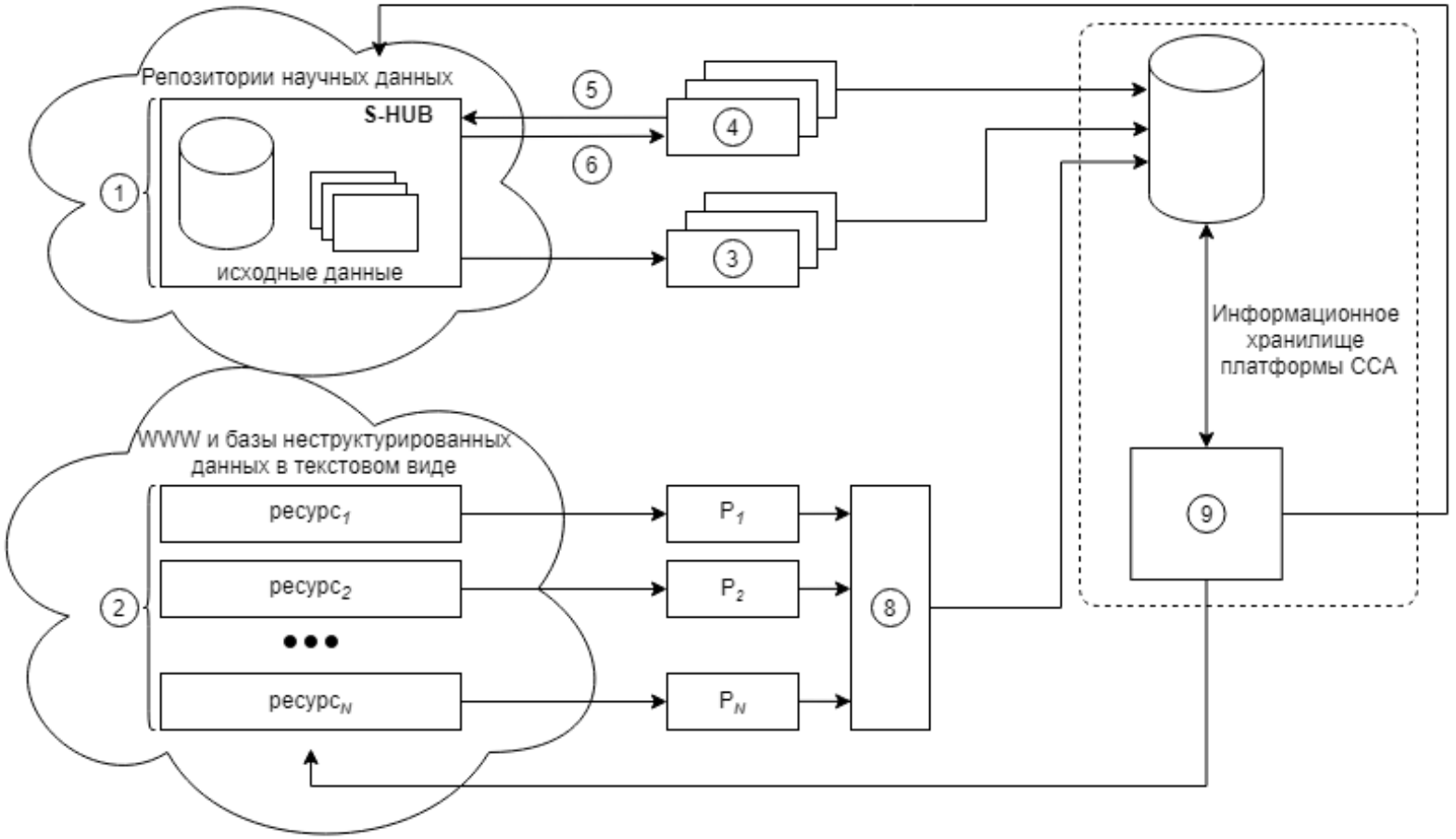


# Классификация проблемно-ориентированного корпуса текстов

Эксперимент 1

Реализован ряд моделей и алгоритмов сбора и обработки корпуса текстов, а также проведения разведочного анализа в рамках разработки прототипа системы построения онтологии и графа знаний. Проведена серия экспериментов для оценки состоятельности, качественных и количественных показателей предлагаемого решения.
Первый эксперимент посвящен модулям сбора, предобработки и построения модели классификации для корпуса научных текстов, включающего 1,5 тыс. документов.

# Схема организации поиска и извлечения текстовых документов с тематических ресурсов




① Структурированное информационное хранилище;

② Неструктурированные данные в WWW в виде электронных информационных ресурсов;

③ Парсеры и грабберы (специализированные программные модули);

④ Применение API для получения метаинформации;

⑤ Запрос на получение метаинформации о документах;

⑥Метаинформация о документах;

⑦Множество специфических парсеров для выделенных электронных информационных ресурсов;

⑧ Агрегатор результатов сбора данных;

⑨ Автоматизированный поисковой модуль (Web crawler +spider);

# Функциональная модель структурно-семантического анализа текстовых данных

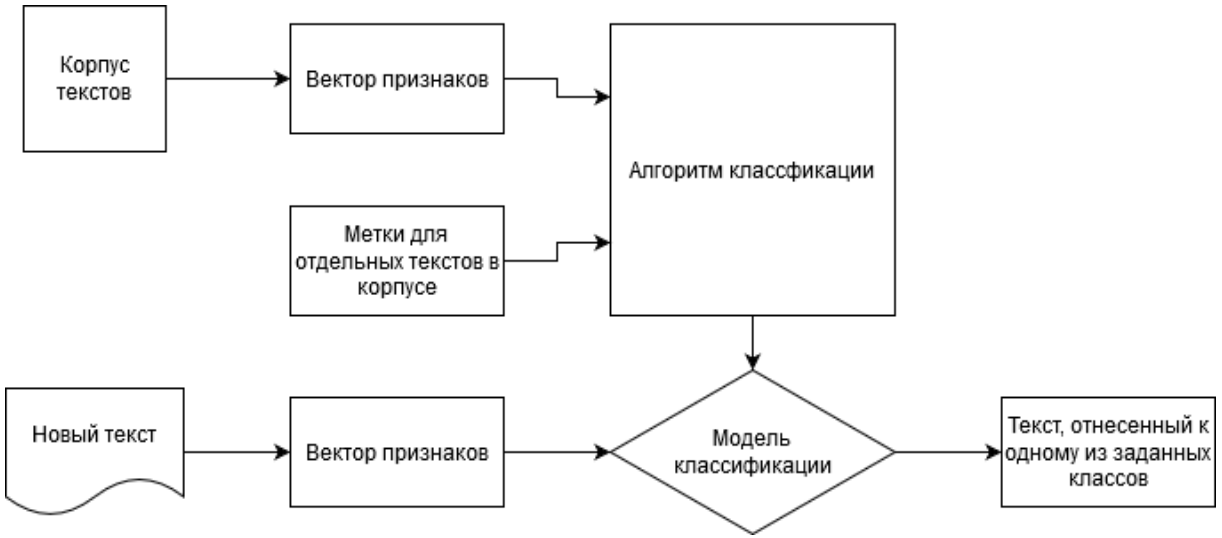

⑦ Информационный массив для поиска литературы;
⑧ Инструменты разметки и аннотирования текстовой документации;
⑨ Список кортежей
⑩ Список кортежей;
⑪ Критерий упорядочивания

① для каждого $w_i$ в $W$ строится упорядоченное по убывания приоритетности для текущего контекста множество смысловых значений $S = \{s_1, \dots, s_p\}$;
② Для каждой пары $w_i$ и $w_j$ близких по смыслу слов строится оценка меры их семантической близости $d$, которая сравнивается с заранее заданным экспертом пороговым значением $\Theta$. При выявлении близких по смыслу слов происходит уточнение их смысловых значений: $d(w_i, w_j) > \Theta\ s_i \cap s_j$. В результате для каждого слова и словосочетания формируется множество S, характеризующее их уточненный смысл;
③ Эксперт предметной области;
④ Специалист по ССА текстов на естественном языке;
⑤ Нормативно-справочные документы и материалы;
⑥ Лексические единицы, которые передают основной смысл тематики документов. $D = \{d_1, \dots, d_m\}$ – множество документов; для $\forall\ d \in D\ \exists\ W$: для каждого документа $d_k$ строится множество $W$ ключевых слов и словосочетаний;

# Методы интеллектуального анализа в задаче классификации текстов на естественном языке

Сравнительный анализ алгоритмов классификации русскоязычных текстов на основе применения методов интеллектуального анализа данных

Обобщенная структура работы классификатора текстов на естественном языке




Результаты сравнения методов классификации документов с использованием корпуса новостей ресурса Lenta.ru

Разработка модели классификатора текстов на естественном языке

1. Загрузка данных из корпуса текстов на русском языке;
2. Предобработка текстовых данных для последующего использования;
3. Векторизация текстовых данных на основе модели «мешок слов» или TF-IDF;
4. Разделение выборки на тренировочную и тестовую;
5. Построение модели классификатора;
6. Оценка результатов модели на обучающей и тестовой выборках;
7. Применение модели в задачах классификации новых данных

| Классификатор | Точность | Полнота |
|---|---|---|
| LinearSVC | 0,88 | 0,87 |
| RandomForestClassifier | 0,87 | 0,87 |
| MultinominalNB | 0,92 | 0,86 |
| BernoulliNB | 0,89 | 0,7 |
| NearestCentroid | 0,5 | 0,43 |
| KneighborsClassifier | 0,76 | 0,49 |

# Автоматизированное построение онтологии и графа знаний корпуса клинических рекомендаций

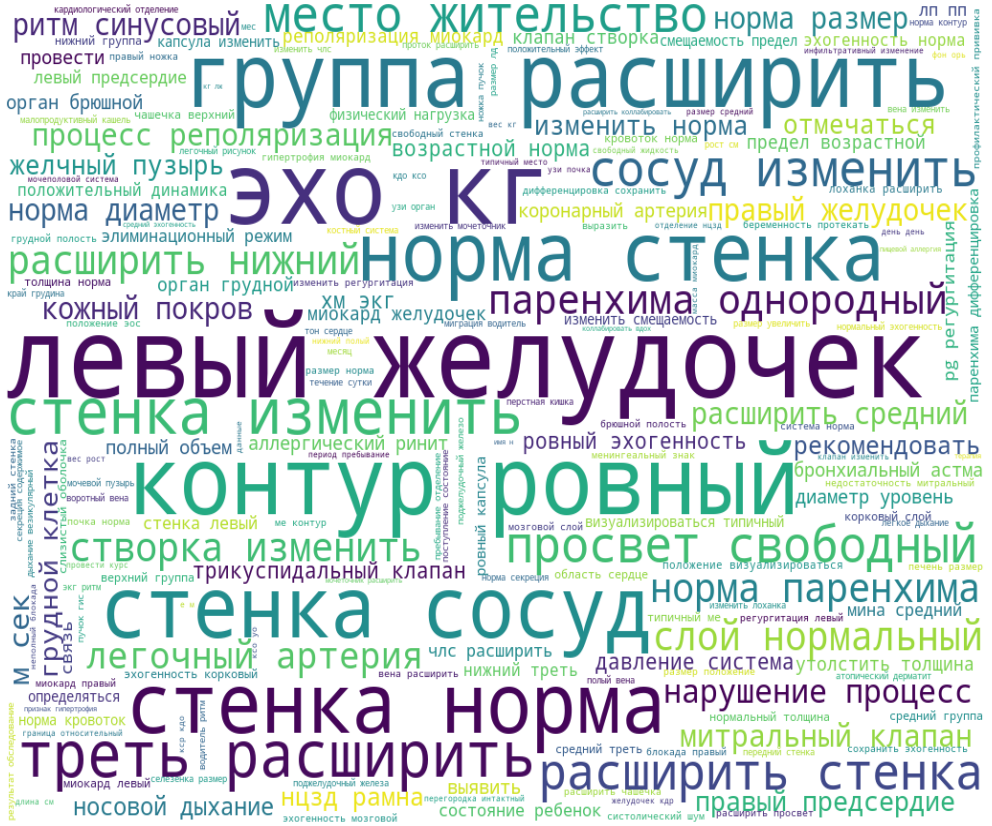
## Эксперимент 2

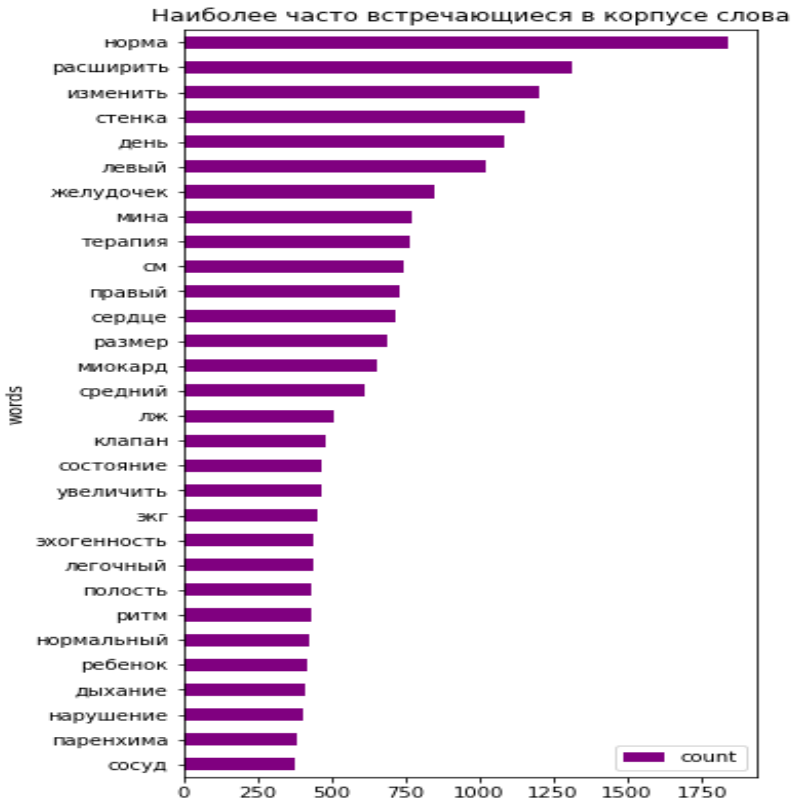
Исходный корпус текстов проблемной области для второго эксперимента предоставлен исследователями Научного центра здоровья детей и Институтом системного анализа Федерального исследовательского центра «Информатика и управление».

Корпус включает 162 документа обезличенных историй болезни пациентов педиатрического центра с болезнями органов дыхания, с аллергическими, нефрологическими и ревматическими болезнями.

Рассмотрено построение конвейера предобработки текстовых данных и извлечения именованных сущностей с помощью нейросетевых моделей общего назначения, не обучавшихся на специализированных текстах

# Структура корпуса тестов клинических рекомендаций




«Облако слов» является удачным инструментом разведочного анализа, позволяющим визуально оценить частотность распределения ключевых слов и словосочетаний. Последующий количественный анализ диаграмм вхождения позволяет уточнить стоп-словарь и необходимость корректировки разметки

# Структура конвейера NLP

| Этап | Шаги | Действия | Инструменты |
|---|---|---|---|
| Предобработка | Символьная фильтрация | Удаление нерелевантных символов, HTML-тегов | Набор регулярных выражений |
| | Токенизация | Разбивка текста на токены с помощью предобученной для русского языка нейросетевой модели | Razdel (фреймворк Natasha), Spacy, Stanza, nltk |
| | Фильтрация нерелевантных токенов | Удаление ссылок, нерелевантных сокращений | Регулярные выражения |
| Нормализация | Лемматизация | Приведение слов в исходную форму с помощью предобученной нейросетевой модели | Morph (фреймворк Natasha), pymorphy2, spacy |
| Постобработка | Частеречная фильтрация | Остаются только существительные, глаголы, прилагательные, наречия, местоимения | Morph (фреймворк Natasha) |
| | Извлечение именованных сущностей | Разметка тегами выделенных типов именованных сущностей | Natasha, spacy |
| | Фильтрация на основе стоп-словарей | Фильтрация нерелевантных лемм с помощью составного стоп-словаря, включающего наиболее часто встречающиеся слова корпуса текстов | NLTK-russian, english |
| | Формирование документа-строки | Объединение лемм в нормализованную строку-документ | |

# Автоматизация извлечения именованных сущностей

Одной из обозначенных в ходе анализа проблем была названа трудоемкость выделения именованных сущностей и необходимость глубокой иерархической разметки корпуса текстов. Далее оценивается применение распространенных для русскоязычного домена моделей выделения именованных сущностей на основе нейронных сетей (рекуррентных и глубоких архитектур), представленных в фреймворках Natasha и spacy, общего назначения. Модели позволяют предварительно разметить корпус для дальнейшей верификации и уточнения разметки, а также построения специализированных моделей (оценка F1-меры для модели общего назначения – 20,4% по сравнению с вариантом использования словаря – 16,7%).

| Параметр | Характеристика | | | |
|---|---|---|---|---|
| Фреймворк | Natasha | Spacy | Базовый вариант I | Модель II |
| Количество выделенных именованных сущностей | 773 | 1164 | - | - |
| Общее пересечение | 335 | | - | - |
| Оценка F1 меры выделения именованных сущностей | 20,4 % | 19,1 % | 16,7 % | 81,9 % |
| Пример выделяемых именованных сущностей (с меткой типа) | 'Квинке': 'PER', 'ГКС': 'ORG', 'Институте иммунологии': 'ORG', 'ОВЛД': 'ORG', 'АБ': 'ORG', 'ЗОД': 'ORG', 'ФВД': 'ORG', 'НИИ педиатрии': 'ORG', 'И.И. Балаболкина': 'PER', 'Зев': 'PER', 'НЦЗД РАМН': 'ORG', | 'АЛТ': 'ORG', 'Цитофлавин': 'PER', 'НИИ паразитологии': 'ORG', 'МБТ': 'ORG', 'Великий Новгород': 'LOC', 'Смекта': 'PER' | - | - |

# Теги частеречной разметки

| Тег-POS | Описание | Пример |
|---|---|---|
| ADJ | adjective, имя прилагательное | большой, старый, зеленый, непонятный |
| ADV | adverb, наречие | очень, завтра, вниз |
| AUX | auxiliary, вспомогательный глагол | есть, будет |
| NOUN | noun, имя существительное | девушка, кошка, земля, |
| NUM | numeral, имя числительное | 1, 20200, один, двадцать восемь, IV, MMXIV |
| PRON | pronoun, местоимение | я, ты, он, она, я, себя, кто-то |
| PROPN | proper noun, имя собственное | РФ, Людвиг Витгенштейн |
| VERB | verb, глагол | бежать, бежит, ест |

В качестве разрешенных тегов частеречной разметки  и последующего построения триплетов «объект»-«действие»-«субъект» использован кортеж разрешенных частей речи. Выбор тегов зависит от специфики текстового корпуса и может быть иным

# Иллюстрация фрагментов исходных, префильтрованных и нормализованных текстов корпуса

| | Фрагменты исходных текстов | Префильтрованные тексты | Нормализованные тексты |
|---|---|---|---|
| 0 | НАХОДИЛСЯ НА ЛЕЧЕНИИ С 9.10.41 ПО 20.10... | находился на лечении с по г диагно... | [находиться, лечение, 20.10.41, г, диагноз, по... |
| 1 | ВОЗРАСТ: 9 лет (16.01.2021).\nНАХОДИЛСЯ НА ЛЕЧ... | возраст лет находился на лечении с ... | [возраст, год, находиться, лечение, 26.02.2031... |
| 2 | ДИАГНОЗ: Бронхиальная астма, атопическая, легк... | диагноз бронхиальная астма атопическая ... | [диагноз, бронхиальный, астма, атопический, ле... |
| 3 | ДИАГНОЗ: Бронхиальная астма, атопическая форма... | диагноз бронхиальная астма атопическая ... | [диагноз, бронхиальный, астма, атопический, фо... |
| 4 | ДИАГНОЗ: Бронхиальная астма, тяжелое персистир... | диагноз бронхиальная астма тяжелое пер... | [диагноз, бронхиальный, астма, тяжелый, персис... |

# Выделение триплетов

Универсальные синтаксические отношения для построения триплетов

| Отношение | Пояснение отношения |
|---|---|
| nsubj | Именное подлежащее, которое является синтаксическим подлежащим. |
| nsubj:pass | Именная группа, которая является синтаксическим подлежащим пассивного предложения. |
| obj | Именная группа, обозначающая объект, на который воздействуют или который претерпевает изменение состояния или движения. |
| obl | Отношение используется для именных (существительное, местоимение, именное словосочетание), функционирующих как неосновной (косвенный) аргумент или дополнение. |
| nmod | Отношение используется для номинальных зависимостей другого существительного или именной фразы и функционально соответствует атрибуту или дополнению родительного падежа. |
| nummod | Числовой модификатор существительного — это любая числовая фраза, которая служит для изменения значения существительного с помощью количества. |

Фрагмент исходной базы триплетов

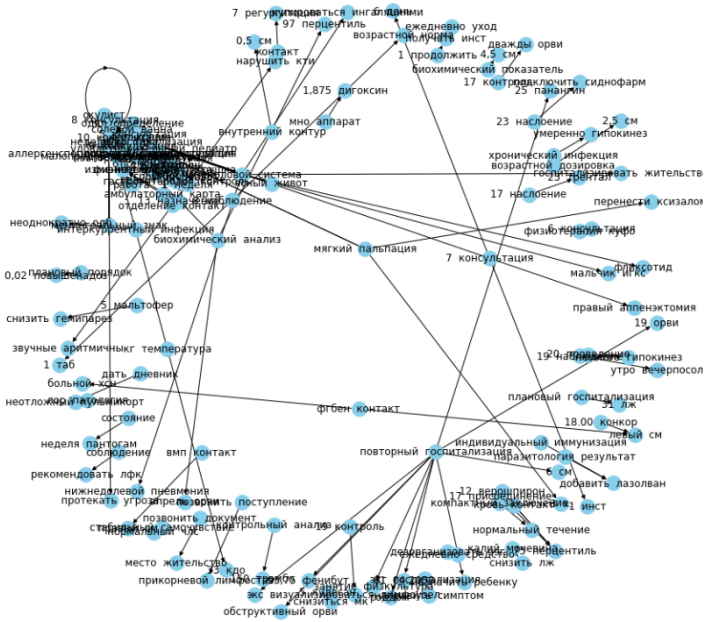
| Полное предложение | subject | verb | object | Subj (нормальная форма) | Obj (нормальная форма) |
|---|---|---|---|---|---|
| ЭПИКРИЗ Ребенок поступил в отделение впервые с... | ЭПИКРИЗ | поступил | отделение | эпикриз | отделение |
| НАХОДИЛСЯ НА ЛЕЧЕНИИ с 26 02 2031 по 4 02 2... | ДИАГНОЗ | НАХОДИЛСЯ | ЛЕЧЕНИИ | диагноз | лечении |
| ЭПИКРИЗ Ребенок поступил в отделение впервые с... | ЭПИКРИЗ | поступил | отделение | эпикриз | отделение |
| ЭПИКРИЗ Ребенок поступил в отделение повторно ... | ЭПИКРИЗ | поступил | отделение | эпикриз | отделение |
| ЭПИКРИЗ Мальчик поступил в клинику впервые с ... | Мальчик | поступил | клинику | мальчик | клинику |

# Фрагмент базы триплетов после фильтрации

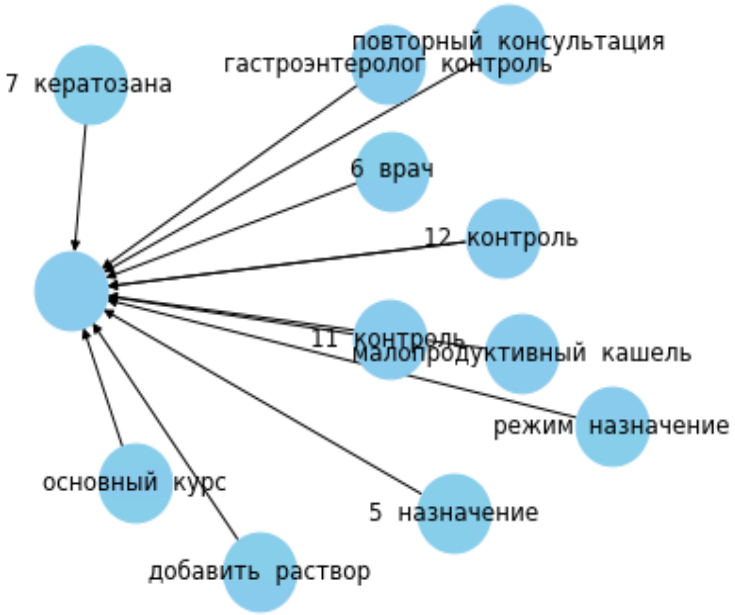
Иллюстрация фрагментов исходных, префильтрованных и нормализованных текстов корпуса

| | Subj (нормальная форма) | Obj (нормальная форма) | verb | Полное предложение |
|---|---|---|---|---|
| 1 | проба | фтизиатром | запрещена | Проба запрещена фтизиатром |
| 2 | орви | форме | болел | Дважды болел ОРВИ в легкой форме в марте 2054... |
| 3 | талия | ушка предсердие | сглажена | Талия сердца сглажена за счет ушка левого пред... |
| 4 | галотерапия | условиях галокамера | Рекомендовано | Консультация врача физиотерапевта Рекомендова... |
| 5 | бронх | уровня ветвь | прослежены | Бронхи прослежены до уровня субсегментарных ве... |
| 6 | бронх | уровня ветвь | прослежены | Бронхи прослежены до уровня субсегментарных ве... |
| 7 | бронх | уровня ветвь | прослеживаются | Бронхи прослеживаются до уровня субсег ментарн... |
| 11 | девочка | улучшением | выписана | Девочка выписана с улучшением |
| 12 | мальчик | терапию стационар | получал | Мальчик регулярно получал терапию в стационаре... |
| 18 | лимфоузел | сторон | структурны | Лимфоузлы структурны с двух сторон |
| 19 | мониторирование | стационарных | проведено | Исследование проведено на аппарате BPLab Мони... |
| 21 | жалоба | стабильным | оставалось | За период пребывания в отделении состояние реб... |
| 25 | преднизолон | снижением доз | добавлен | Терапия в отделении Дигоксин 0 000035 мг х 2 ... |
| 32 | голова | размере | увеличена | Голова резко увеличена в размере |
| 38 | вено-венозный | признаками эпитализация | кондуит | Кавапульмональный анастомоз проходим диаметро... |
| 42 | терапия | показаниям | назначена | Повторная госпитализация в НЦЗД РАМН при стаби... |
| 54 | эпикриз | отделение | поступила | ЭПИКРИЗ Девочка поступила в отделение впервые ... |
| 63 | терапия | объеме | продолжена | Терапия продолжена в прежнем объеме Следующа... |
| 64 | терапия | объеме | оставлена | Терапия оставлена в прежнем объеме После вып... |

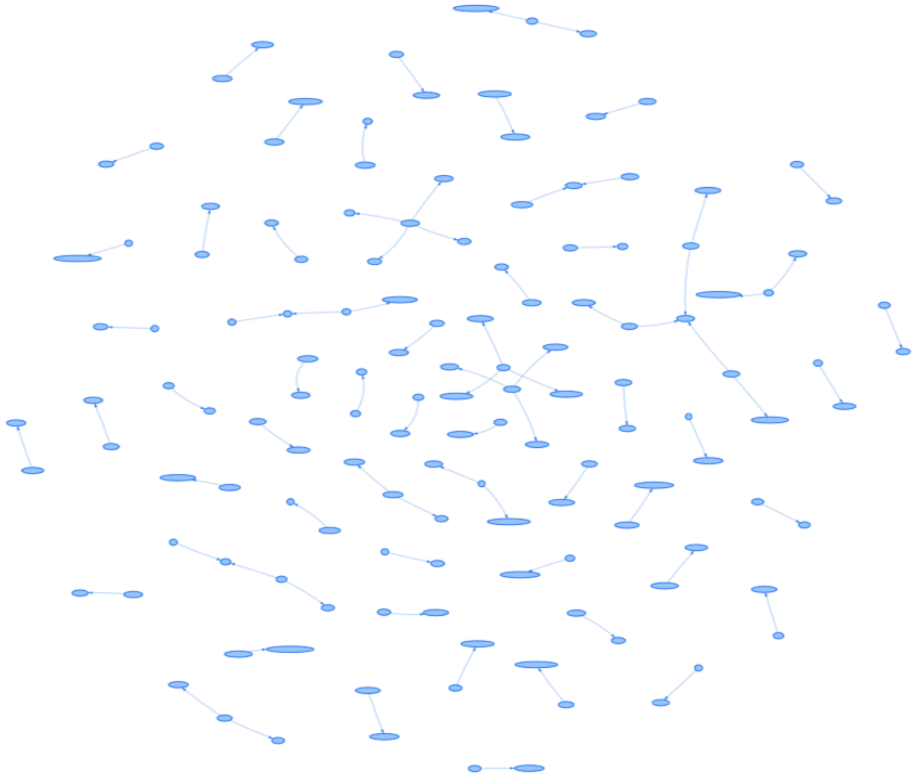
# Вариация взаимной привязки триплетов



Вариация взаимной привязки триплетов

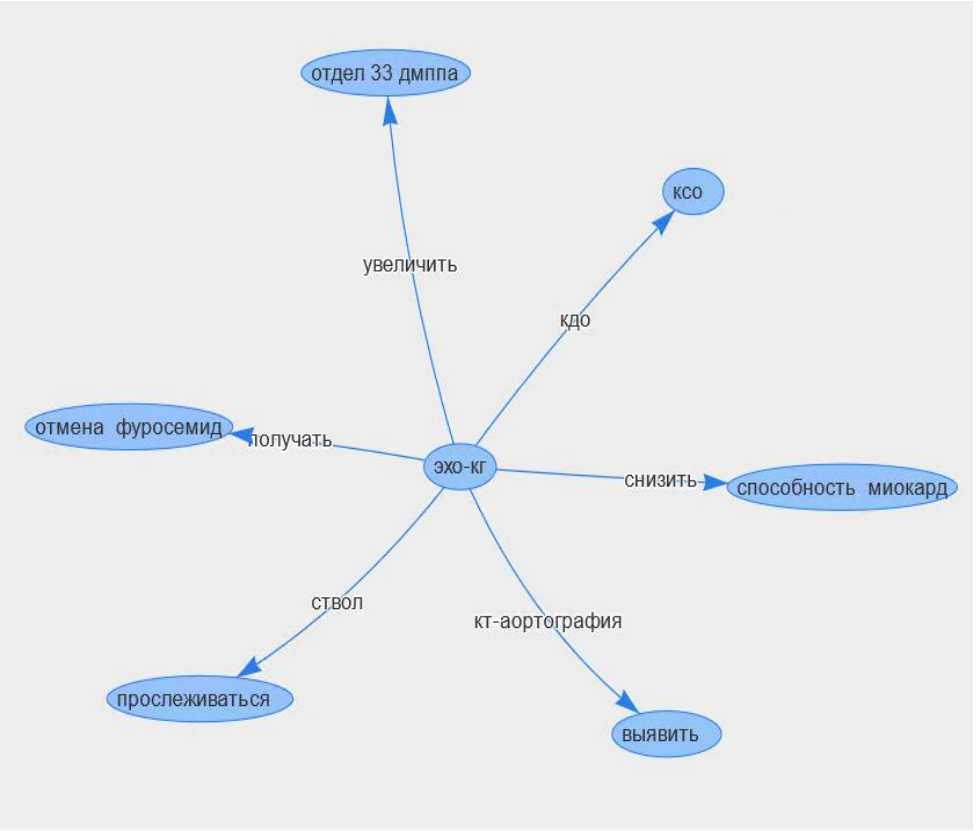


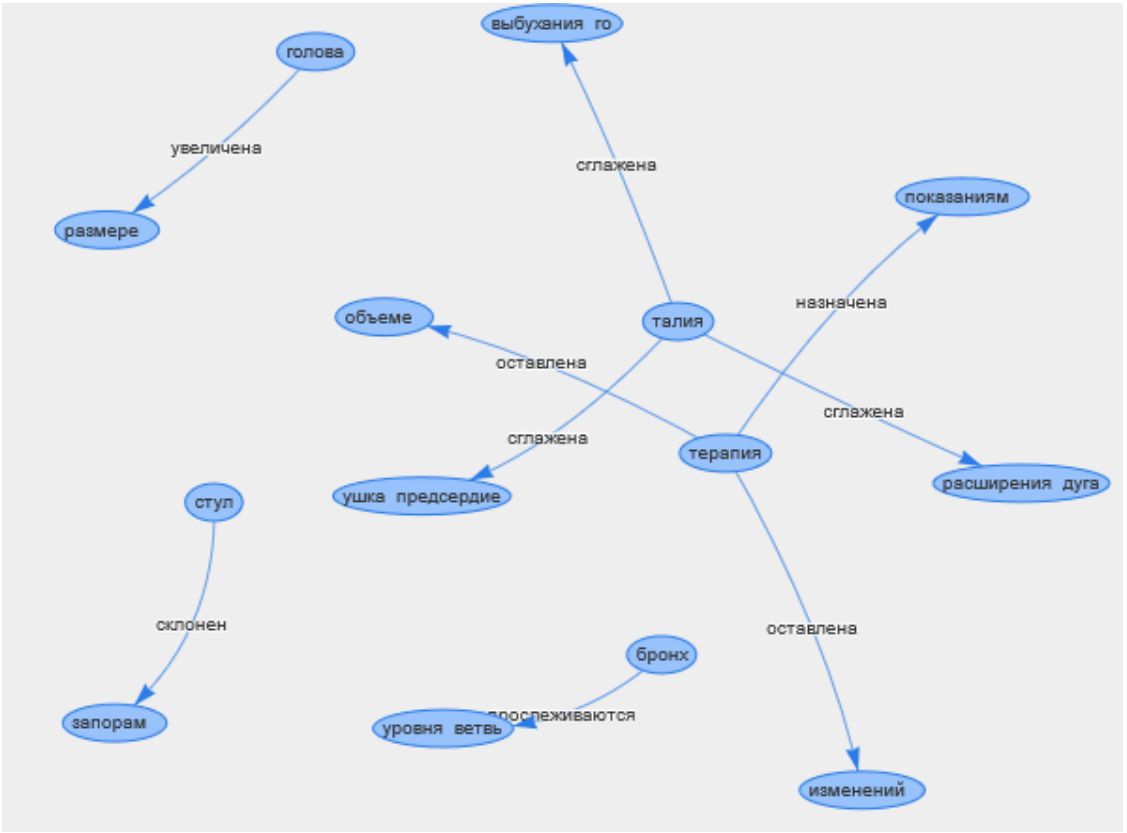

Подграф с типом ребра «наблюдение»



Фрагмент структуры построенного графа знаний

# Фрагмент автоматически извлеченного графа знаний




Фрагменты автоматически извлеченного графа знаний

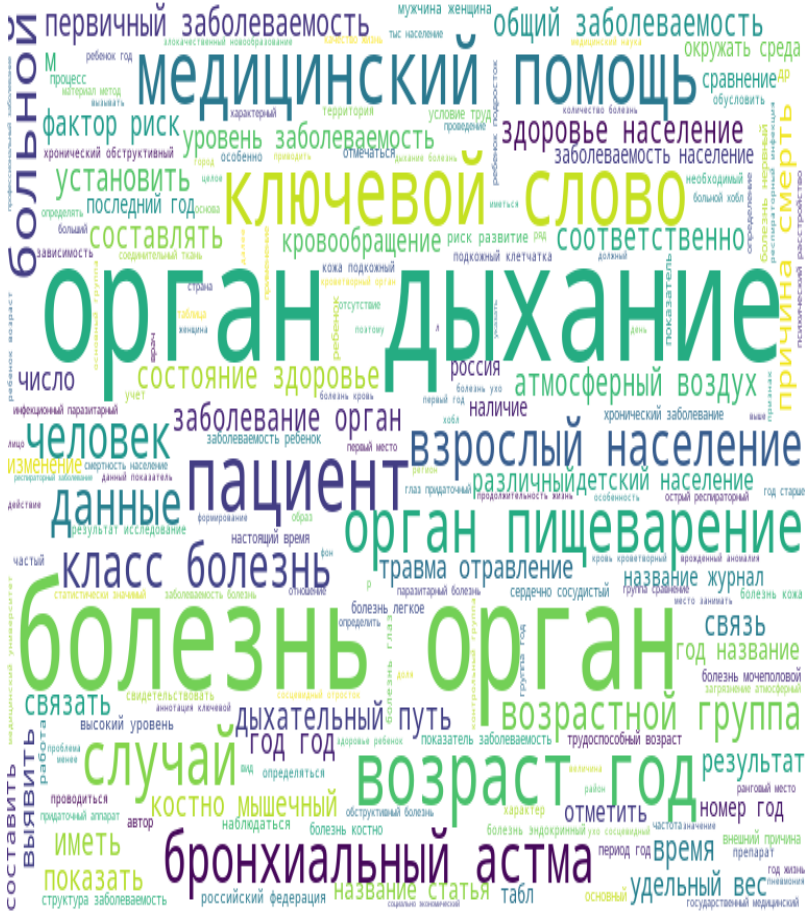
# Тематическое моделирование и построение фрагмента графа знаний корпуса медицинских научных текстов

## Эксперимент 3

Эксперимент 3 основан на сборе и разведочном анализе текстовых документов на русском языке на примере научных публикаций в рецензируемых изданиях по тематике описания различных аспектов болезней органов дыхания, доступных на платформе elibrary.ru, за период с 1997 по 2022 г.

Проведен эксперимент на корпусе медицинских текстов заданной тематики (более 1 тыс. публикаций) без предварительной разметки с целью выявления основных направлений публикаций.

# Проблемно-ориентированный корпус (статистика)




Статистика корпуса

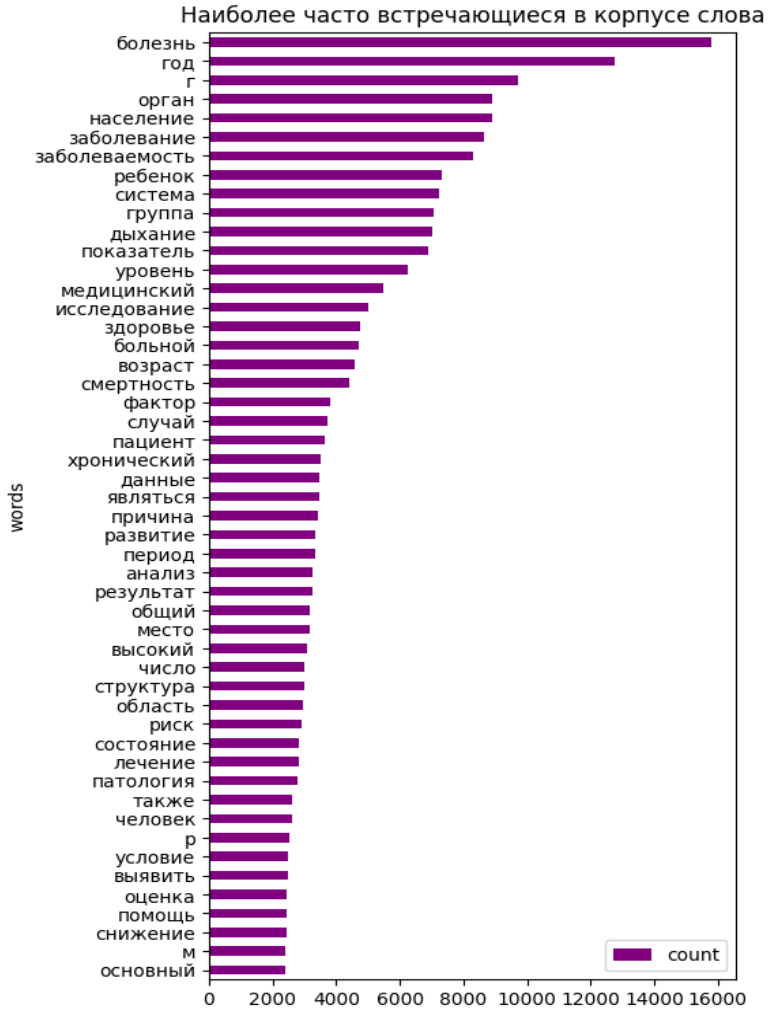
Max words: 15085
Min words: 349
Avg words: 3160.2875175315567
Avg sentences 108.22300140252455
Avg unique words: 1058.4726507713885
Avg unique lemmas: 843.8597475455821

Summary stats:

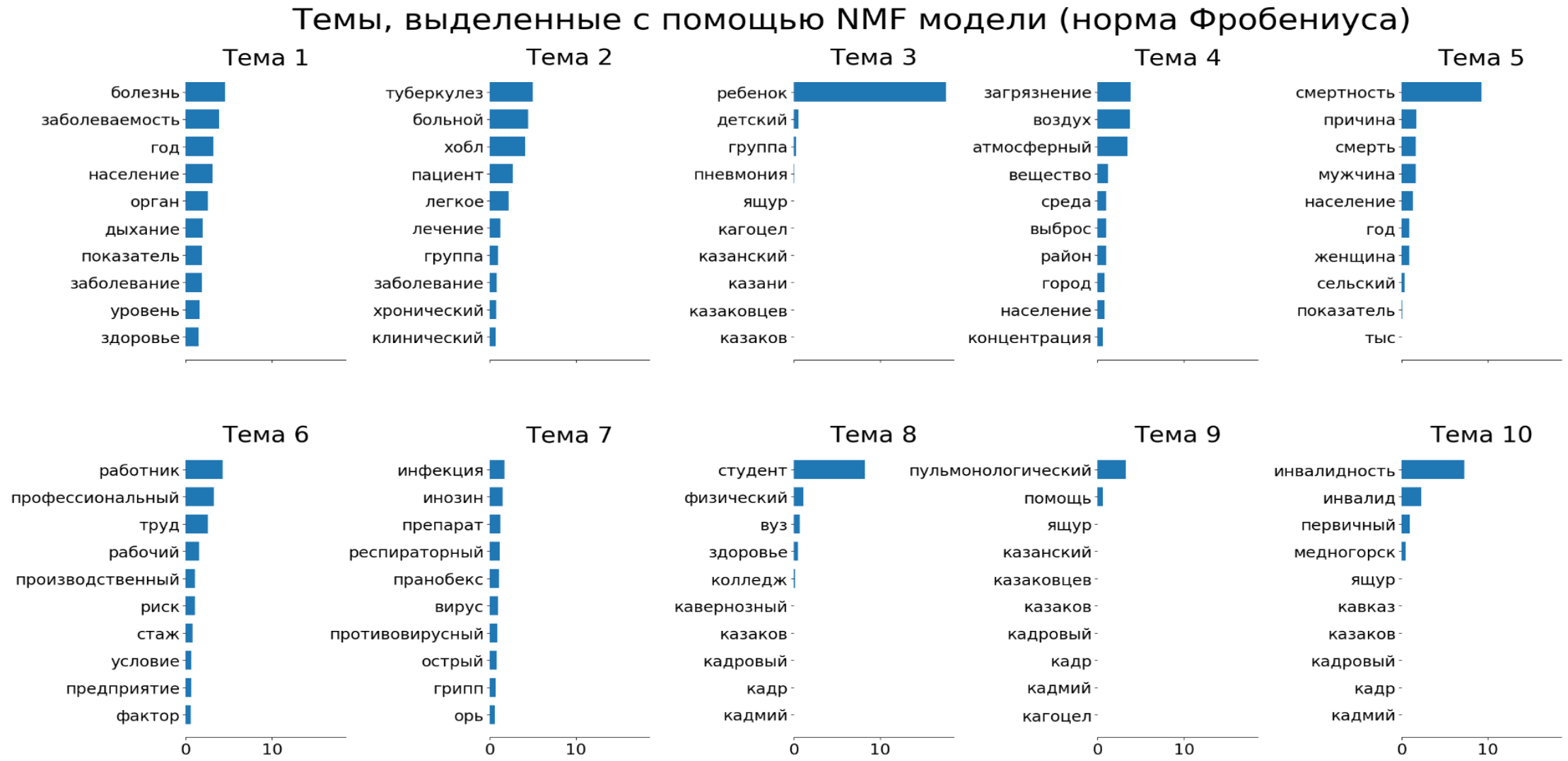
Max words: 381
Min words: 0
Avg words: 50.711079943899016
Avg sentences 2.5680224403927068
Avg unique words: 34.96774193548387
Avg unique lemmas: 32.52594670406732

Intersection stats:

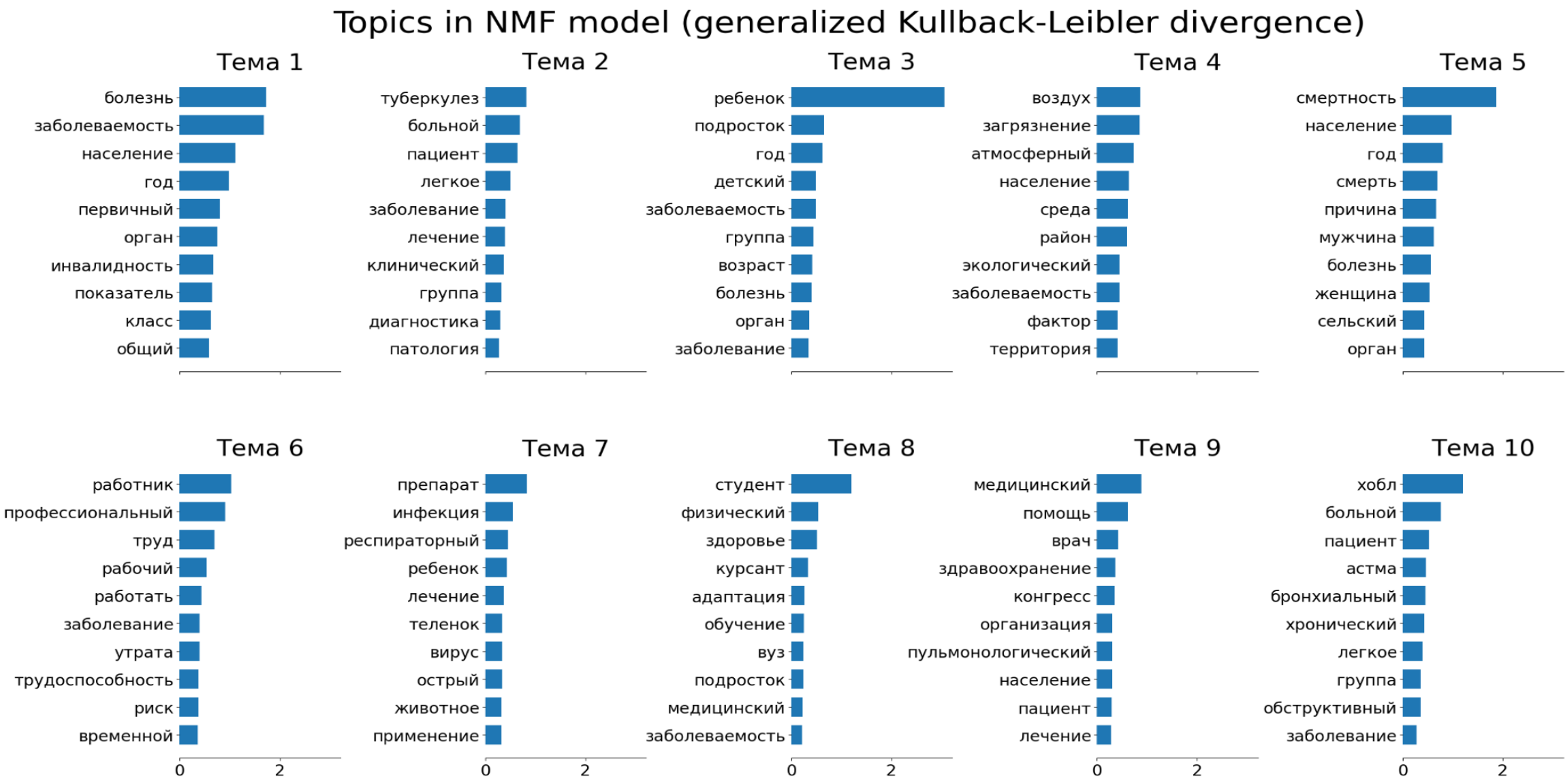
Avg common unique words: 34.96774193548387
Avg common unique lemmas: 32.52594670406732
Avg common words: 99.86%
Avg common lemmas: 99.86%
Avg common bigrams: 48.11%
Avg common trigrams: 48.11%



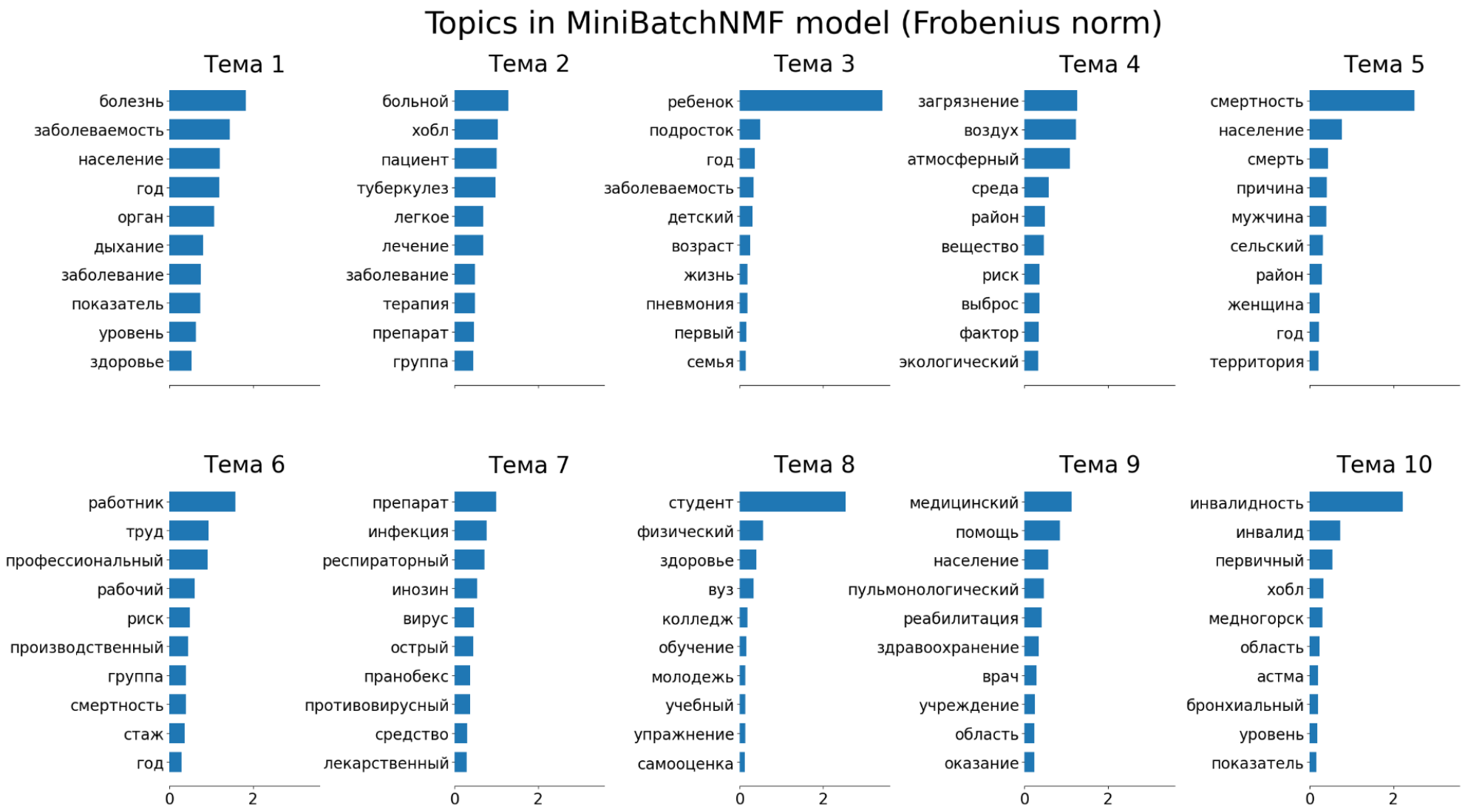
# Извлечение тем документов: NMF + норма Фробениуса



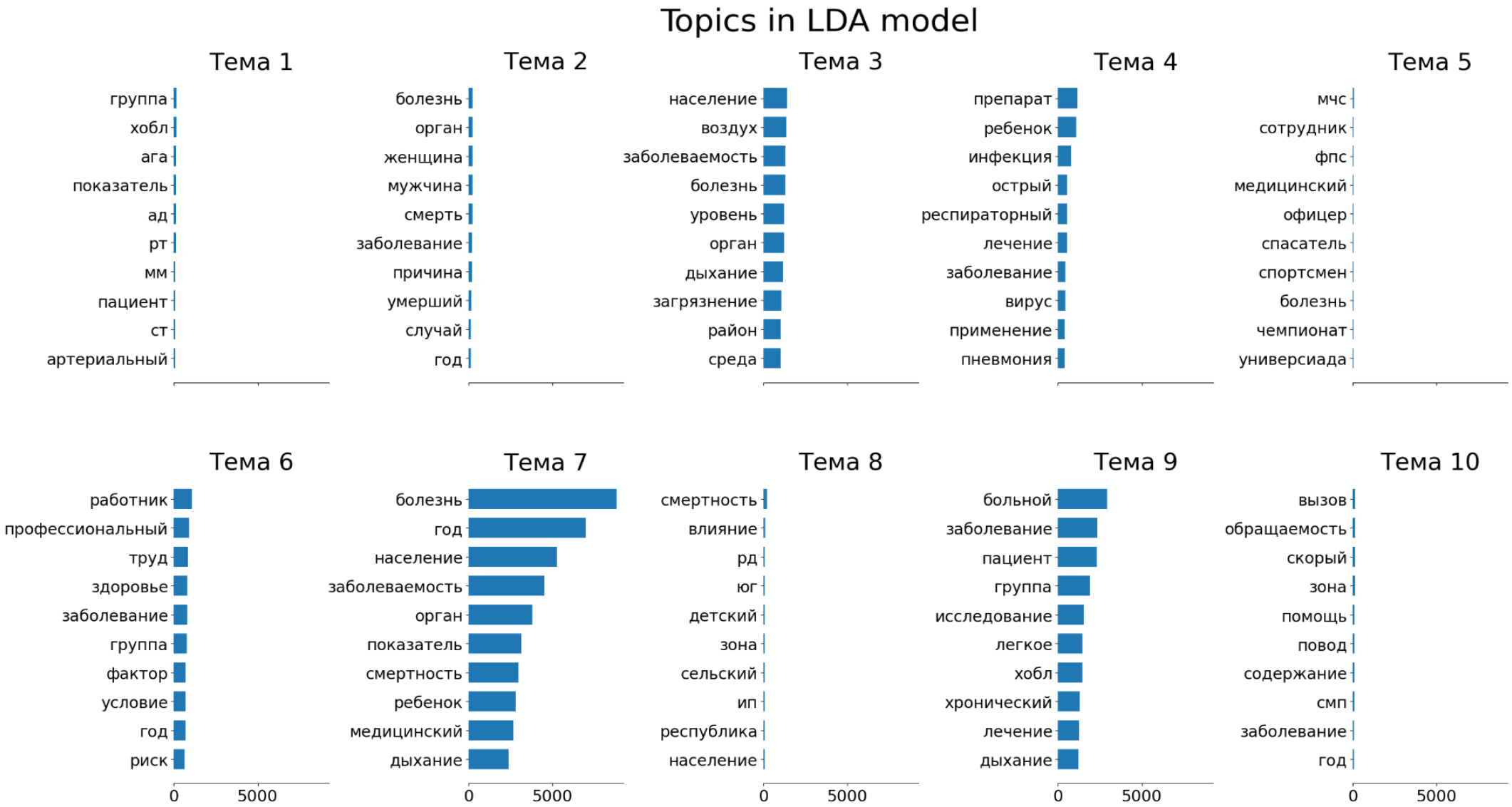
# Извлечение тем документов: NMF + обобщенная дивергенция Кульбака-Лейблера



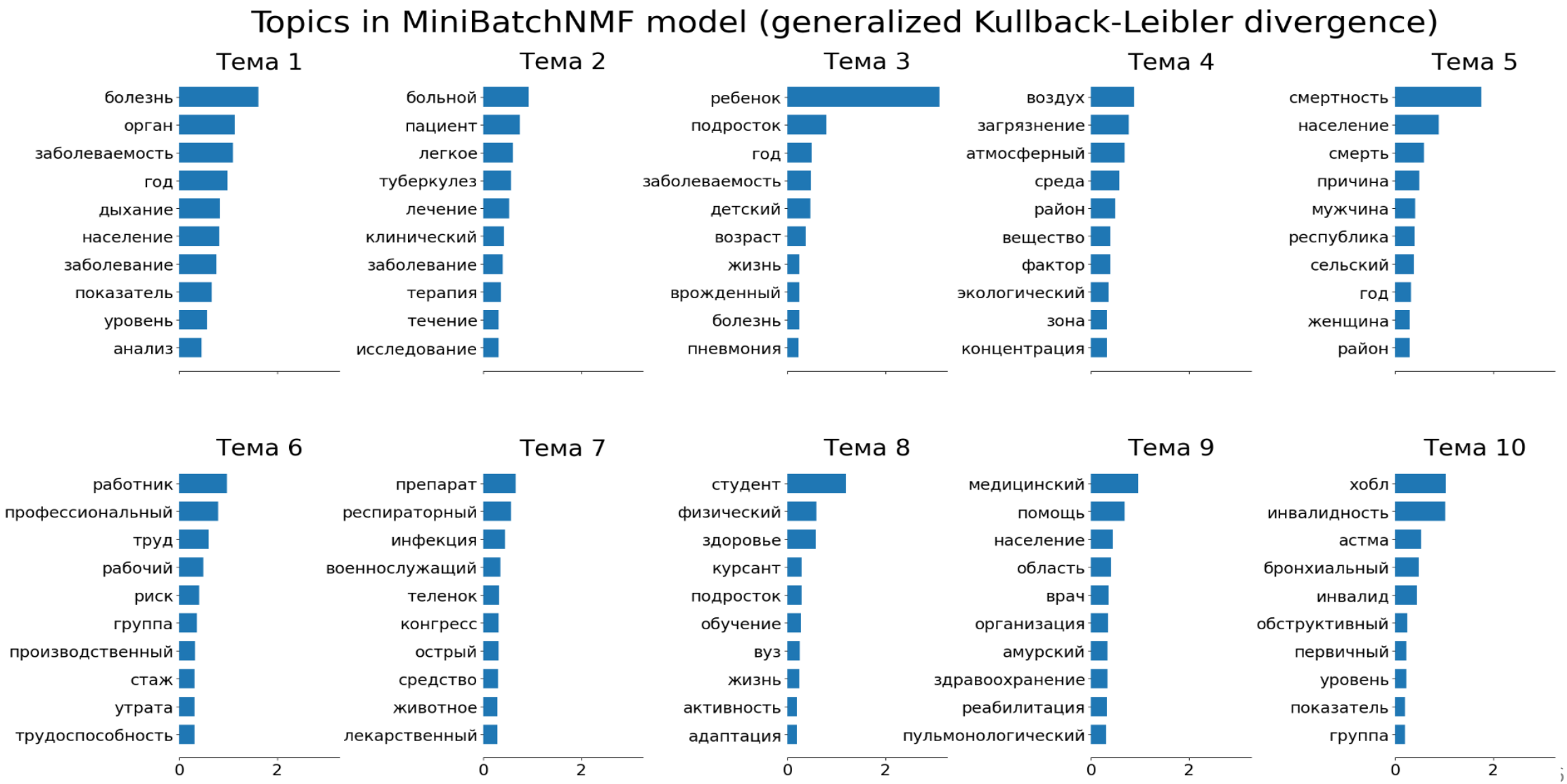
# Извлечение тем документов: пакетная NMF + норма Фробениуса



# Извлечение тем документов: LDA



# Извлечение тем документов: пакетная NMF + обобщенная дивергенция Кульбака-Лейблера (наилучший результат)

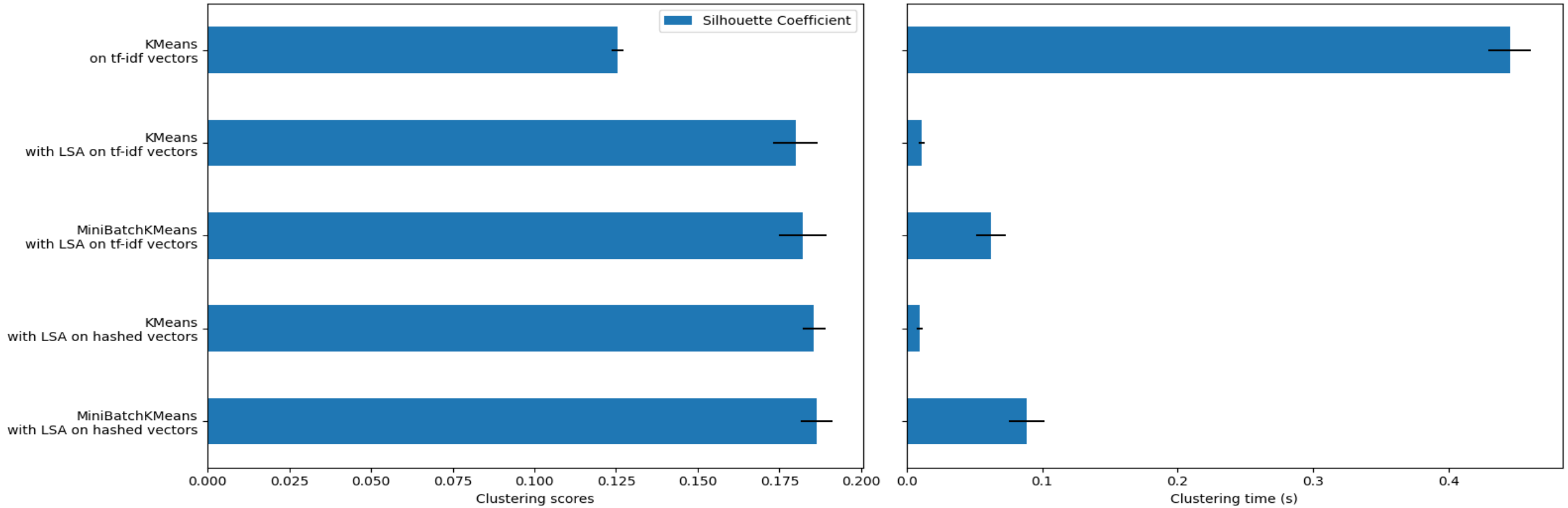


# Кластеризация текстовых документов на основе представления матрицы «терм-документ» (TF-IDF) (альтернативный вариант выявления смысловых групп)

```
[  2 896   2  19   4 365 128   4   2   4]
[  2 828  91  15   4   8   2 317 155   4]
[  1 156 775  83  30  19 345   4   2  11]
[  2 814 387   8  31  31 145   4   2   2]
[ 83 330 762   3   2   2   2  74 146  22]
[102 674   2   3 184 168 152   7 121  13]
[  2 832 143   5  10   2   4  31   2 395]
[613   4 108 315   1  34   6 246  15  84]
[ 91 785   2 358 154   2   2  25   4   3]
[  1 606 259   6  31   8   8   2 471  34]
```

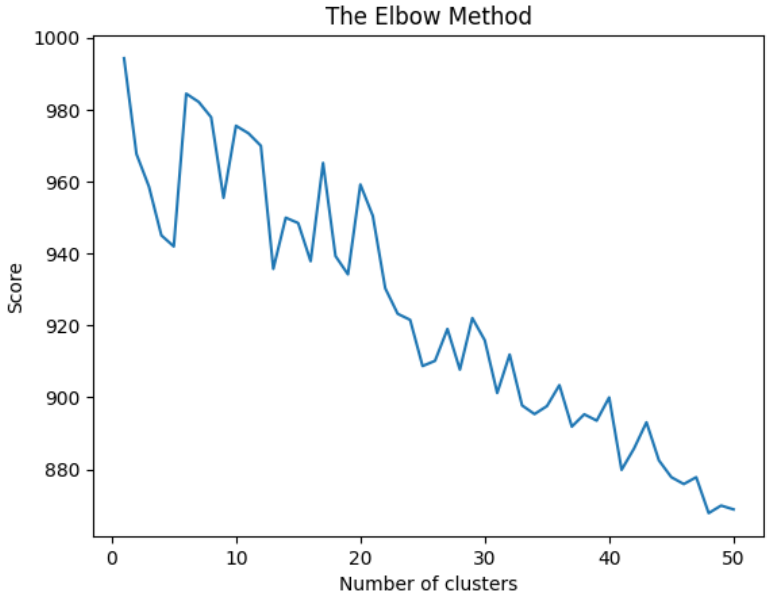
Cluster 0: теленок конгресс военнослужащий медицинский животное респираторный область курсант скот лечение
Cluster 1: препарат ребенок лечение лекарственный терапия инфекция кашель респираторный применение средство
Cluster 2: студент заболеваемость вуз здоровье колледж обучение медицинский самооценка физкультурный молодежь
Cluster 3: студент физический здоровье активность упражнение занятие вуз нагрузка состояние развитие
Cluster 4: больной хобл пациент легкое туберкулез группа хронический лечение клинический патология
Cluster 5: ребенок заболеваемость группа подросток возраст детский здоровье пневмония развитие жизнь
Cluster 6: работник профессиональный труд рабочий заболеваемость группа риск производственный стаж медицинский
Cluster 7: воздух загрязнение атмосферный население заболеваемость среда район вещество здоровье фактор
Cluster 8: инвалидность инвалид первичный население медногорск область возраст реабилитация группа 2011
Cluster 9: население заболеваемость смертность медицинский причина подросток общий структура первичный область

# Кластеризация текстовых документов на основе матрицы «терм-документ» (TF-IDF)




В задаче кластеризации проблемой также является понижение размерности признакового пространства с помощью LSA и выбор устойчивой реализации алгоритма кластеризации – например, пакетный режим.
Наилучшие результаты показал пакетный вариант кластеризации для пространства признаков, полученного с помощью LSA из хешированного представления матрицы «терм-документ». В качестве метрики качества кластеризации использован критерий «ширина силуэта».

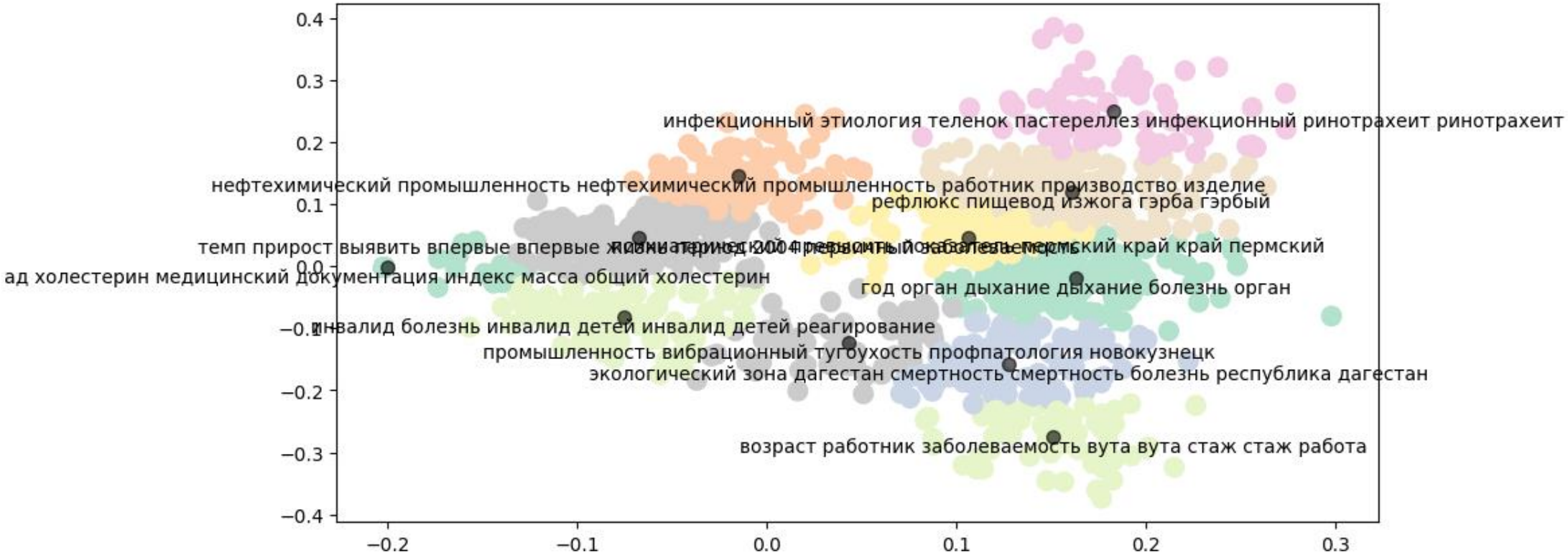
# Кластеризация текстовых документов на основе матрицы «терм-документ» (TF-IDF)




Изменение метрики качества кластеризации в зависимости от количества заданных кластеров показано на графике. Однозначно выбрать оптимальное количество кластеров затруднительно, однако, согласовав тенденцию с результатами тематического моделирования принято решение остановится на 10 кластерах

Cluster 0: общий холестерин ад холестерин индекс масса медицинский документация медицинский карта эффективность профилактический модификация неинфекционный заболевание имя мечников документация
Cluster 1: болезнь орган дыхание год орган дыхание заболевание заболеваемость население болезнь орган показатель уровень
Cluster 2: нефтехимический промышленность нефтехимический работник производство промышленность изделие нервный сердечно тракт орган заболеваемость временной временной утрата утрата трудоспособность утрата
Cluster 3: экологический зона дагестан республика дагестан смертность смертность болезнь рд равнина сельский предгорие горный природно антропогенный
Cluster 4: теленок пастереллез инфекционный этиология инфекционный ринотрахеит ринотрахеит ринотрахеит вирусный сельскохозяйственный предприятие парагрипп животное сальмонеллез животноводство
Cluster 5: инвалид болезнь детей инвалид детей инвалид реагирование речевой симптоматика младший возрастной психологический задержка реабилитация
Cluster 6: вута заболеваемость вута стаж стаж работа возраст работник работник горнорудный группа характерный основный группа зависимость стаж контрольный группа
Cluster 7: пермский край психиатрический пермский край превысить показатель существенный отличие болезнь дыхательный выразить тенденция мочеполовой кожа пищеварение орган отделение
Cluster 8: гэрба гэрбый пищевод рефлюкс изжога аспирационный ночной желудочный содержимое рвота соляный
Cluster 9: новокузнецк вибрационный промышленность профпатология тугоухость коллегия период отмечаться угольный кемеровский кемеровский область металлургия
Cluster 10: первичный заболеваемость впервые жизнь период 2004 выявить впервые темп прирост первичный заболеваемость подросток общий заболеваемость заболеваемость целое отчетный документация изучение динамика

# Кластеризация текстовых документов на основе матрицы «терм-документ» (TF-IDF)



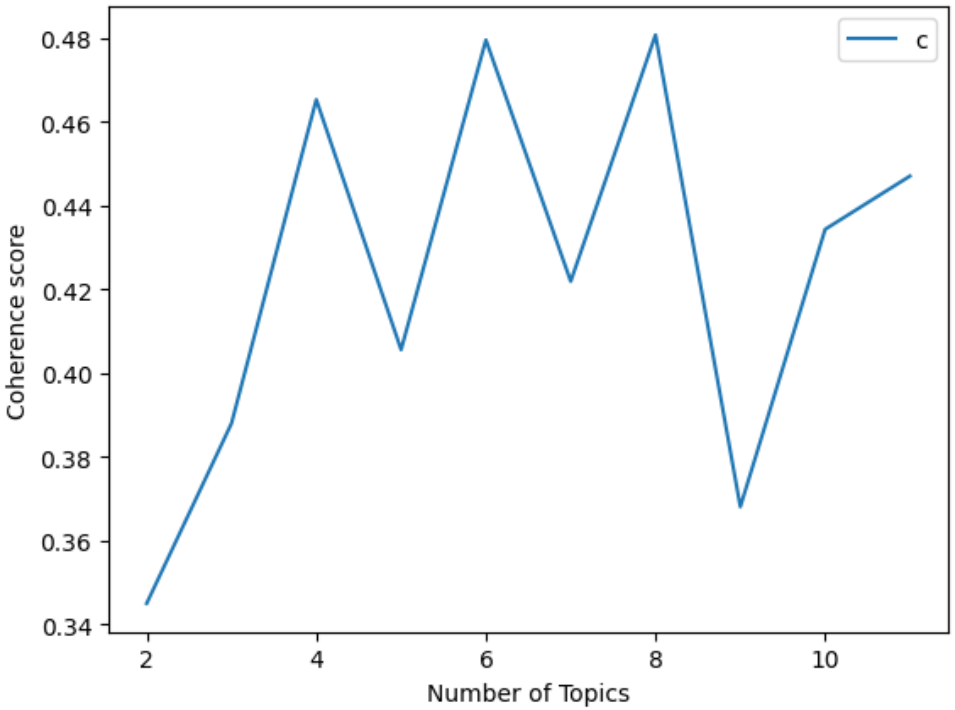

Визуализация кластеров документов с помощью понижения размерности признакового пространства методом главных компонент. Анализ показывает приемлемое качество выделенной структуры документов

# LDA

(0, '0.015*"рт" + 0.009*"ад" + 0.008*"ага" + 0.007*"хсн" + 0.007*"мм" + ' '0.007*"артериальный" + 0.006*"ст" + 0.005*"индекс" + 0.005*"давление" + ' '0.005*"суточный"'),
(1, '0.034*"медицинский" + 0.019*"помощь" + 0.010*"здравоохранение" + ' '0.009*"население" + 0.008*"заболевание" + 0.008*"организация" + ' '0.008*"врач" + 0.007*"год" + 0.007*"г" + 0.007*"больной"'),
(2, '0.027*"заболеваемость" + 0.021*"болезнь" + 0.019*"ребенок" + 0.017*"орган" ' '+ 0.016*"здоровье" + 0.015*"дыхание" + 0.013*"заболевание" + ' '0.011*"уровень" + 0.011*"система" + 0.011*"показатель"'),
(3, '0.015*"больной" + 0.012*"пациент" + 0.010*"хобл" + 0.009*"заболевание" + ' '0.009*"группа" + 0.007*"легкое" + 0.007*"исследование" + ' '0.007*"бронхиальный" + 0.007*"хронический" + 0.006*"ба"'),
(4, '0.012*"инфекция" + 0.012*"туберкулез" + 0.010*"препарат" + ' '0.009*"пневмония" + 0.009*"респираторный" + 0.008*"лечение" + '0.007*"острый" + 0.007*"заболевание" + 0.006*"болезнь" + 0.006*"ребенок"'), (5, '0.017*"пациент" + 0.014*"группа" + 0.013*"больной" + 0.012*"заболевание" + ' '0.012*"случай" + 0.010*"мужчина" + 0.010*"женщина" + 0.010*"смерть" + ' '0.009*"сердце" + 0.008*"причина"'),
(6, '0.021*"работник" + 0.019*"профессиональный" + 0.016*"труд" + 0.014*"группа" ' '+ 0.013*"риск" + 0.010*"фактор" + 0.010*"условие" + 0.009*"рабочий" + ' '0.008*"заболевание" + 0.008*"р"'),
(7, '0.011*"вута" + 0.009*"оренбургский" + 0.008*"медногорск" + ' '0.007*"вскармливание" + 0.004*"обл" + 0.004*"комбинат" + '0.004*"свердловский" + 0.003*"асбест" + 0.003*"газотранспортный" + ' '0.003*"буровой"'),
(8, '0.042*"болезнь" + 0.033*"год" + 0.025*"население" + 0.021*"г" + ' '0.016*"система" + 0.016*"смертность" + 0.014*"орган" + 0.013*"показатель" + ' '0.012*"возраст" + 0.011*"причина"'),
(9, '0.015*"курение" + 0.013*"курить" + 0.009*"респондент" + 0.007*"вологда" + ' '0.006*"табачный" + 0.006*"табак" + 0.004*"курильщик" + 0.004*"дым" + ' '0.004*"сигарета" + 0.004*"табакокурение"'),
(10, '0.024*"воздух" + 0.018*"население" + 0.016*"загрязнение" + 0.014*"район" + ' '0.014*"атмосферный" + 0.013*"среда" + 0.010*"вещество" + 0.010*"зона" + ' '0.010*"влияние" + 0.010*"фактор"')

С помощью латентного размещения Дирихле выполнено тематическое моделирование с оценкой каждой из выделенных тем

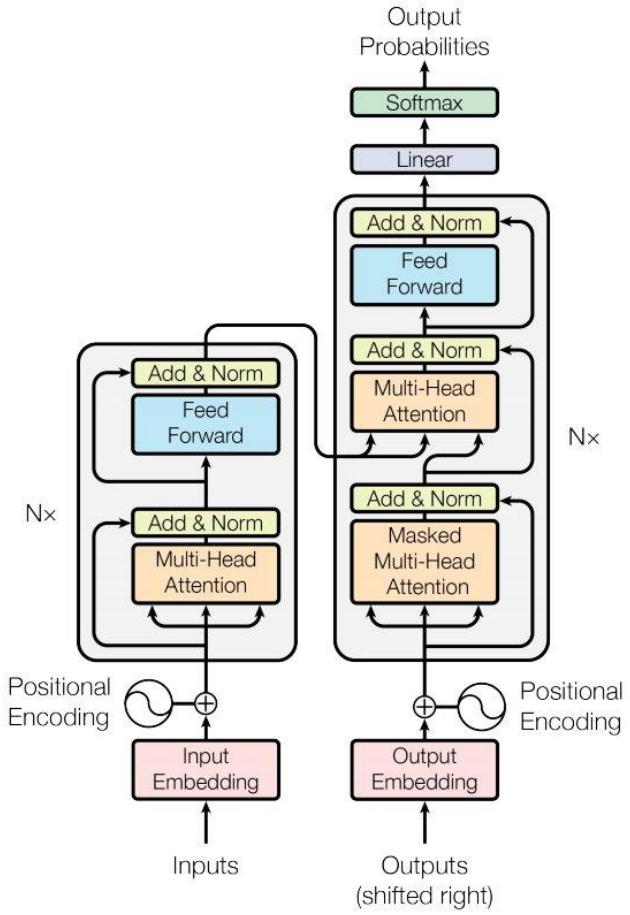
# LSA

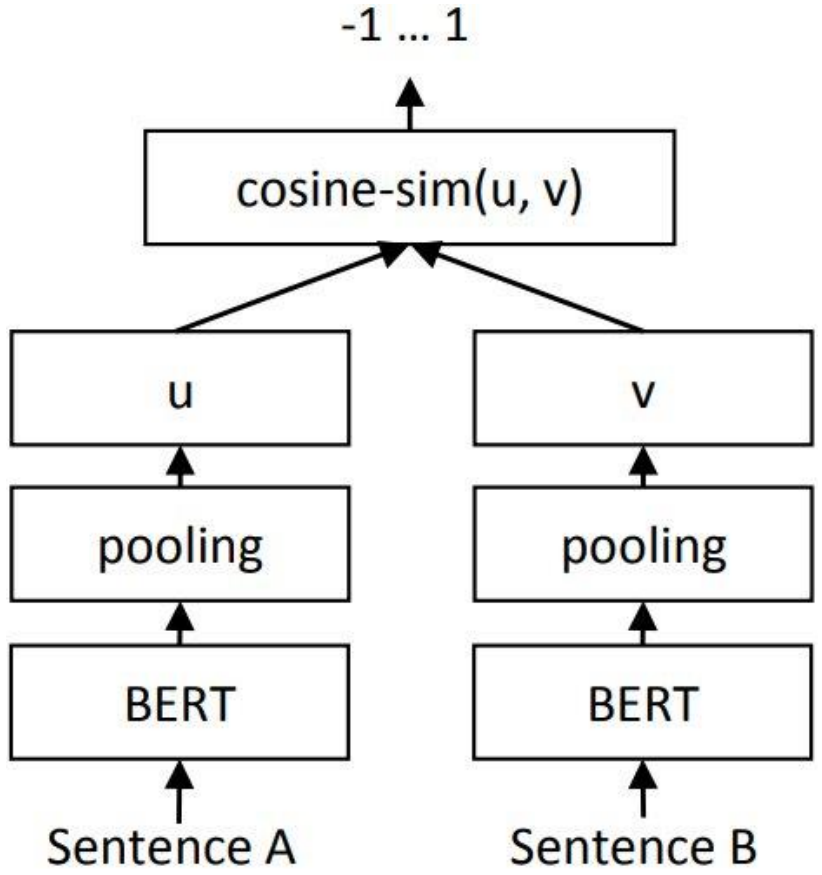



(0, '0.423*"болезнь" + 0.336*"год" + 0.224*"население" + 0.216*"г" + 0.188*"система" + 0.183*"орган" + 0.176*"заболеваемость" + 0.171*"показатель" + 0.146*"заболевание" + 0.141*"смертность"'),
(1, '-0.433*"болезнь" + -0.337*"количество" + -0.177*"система" + 0.152*"ребенок" + 0.143*"группа" + 0.142*"смертность" + -0.139*"рассматривать" + 0.133*"больной" + 0.132*"заболевание" + -0.124*"целое"'), (2, '-0.500*"смертность" + -0.244*"причина" + -0.224*"смерть" + -0.218*"год" + -0.165*"население" + 0.162*"заболевание" + -0.159*"мужчина" + 0.157*"больной" + 0.152*"ребенок" + -0.135*"женщина"'), (3, '0.364*"заболеваемость" + 0.323*"население" + -0.241*"количество" + 0.239*"г" + -0.191*"больной" + -0.153*"год" + -0.152*"пациент" + -0.150*"смертность" + -0.142*"причина" + -0.133*"смерть"'), (4, '-0.699*"ребенок" + -0.220*"год" + 0.188*"смерть" + 0.187*"больной" + 0.154*"медицинский" + 0.144*"причина" + 0.141*"население" + -0.112*"группа" + 0.108*"хобл" + -0.106*"заболеваемость"'), (5, '0.340*"болезнь" + -0.324*"г" + -0.268*"показатель" + -0.252*"количество" + 0.243*"смерть" + 0.186*"ребенок" + 0.174*"причина" + 0.144*"возраст" + 0.137*"заболеваемость" + 0.134*"случай"'), (6, '-0.468*"г" + -0.214*"больной" + 0.202*"воздух" + 0.198*"риск" + 0.186*"фактор" + 0.138*"смертность" + 0.138*"среда" + -0.135*"пациент" + -0.132*"инвалидность" + 0.131*"здоровье"')

Зависимость согласованности от количества тематик, полученные с помощью модели LSA

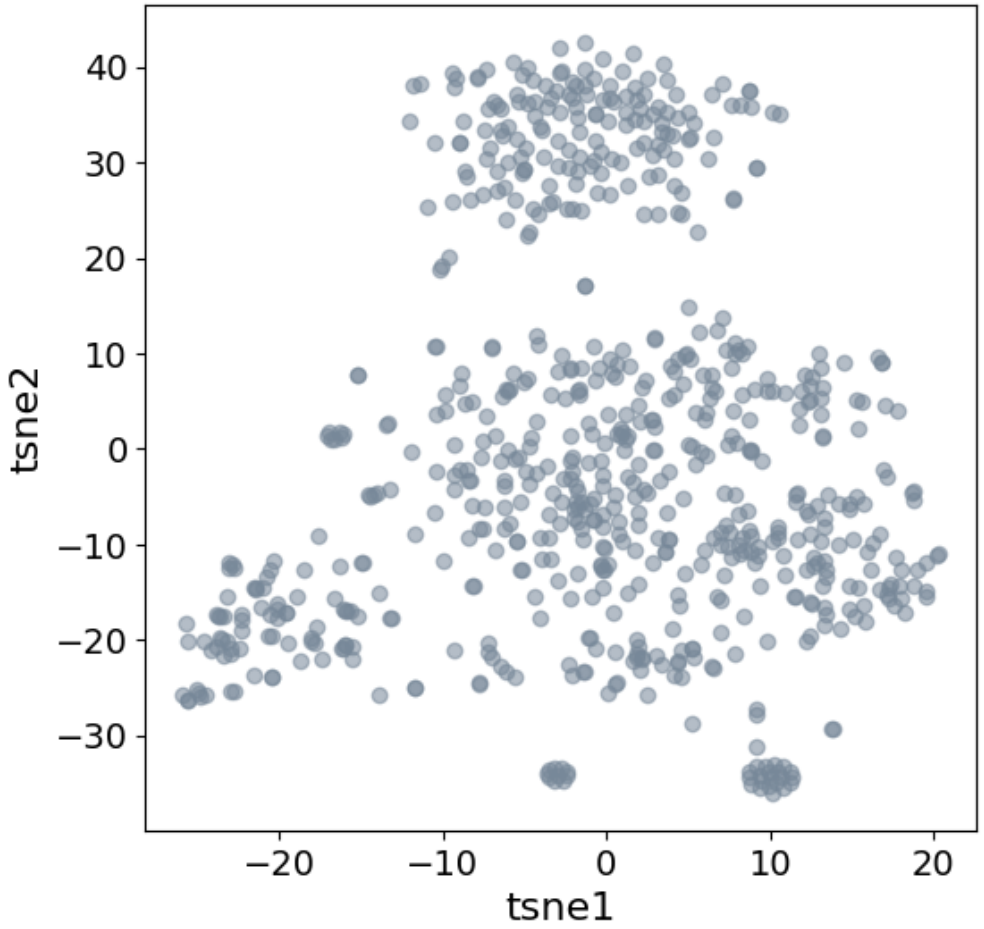
# BERT модель для построения векторов вложений на уровне предложений и текстов



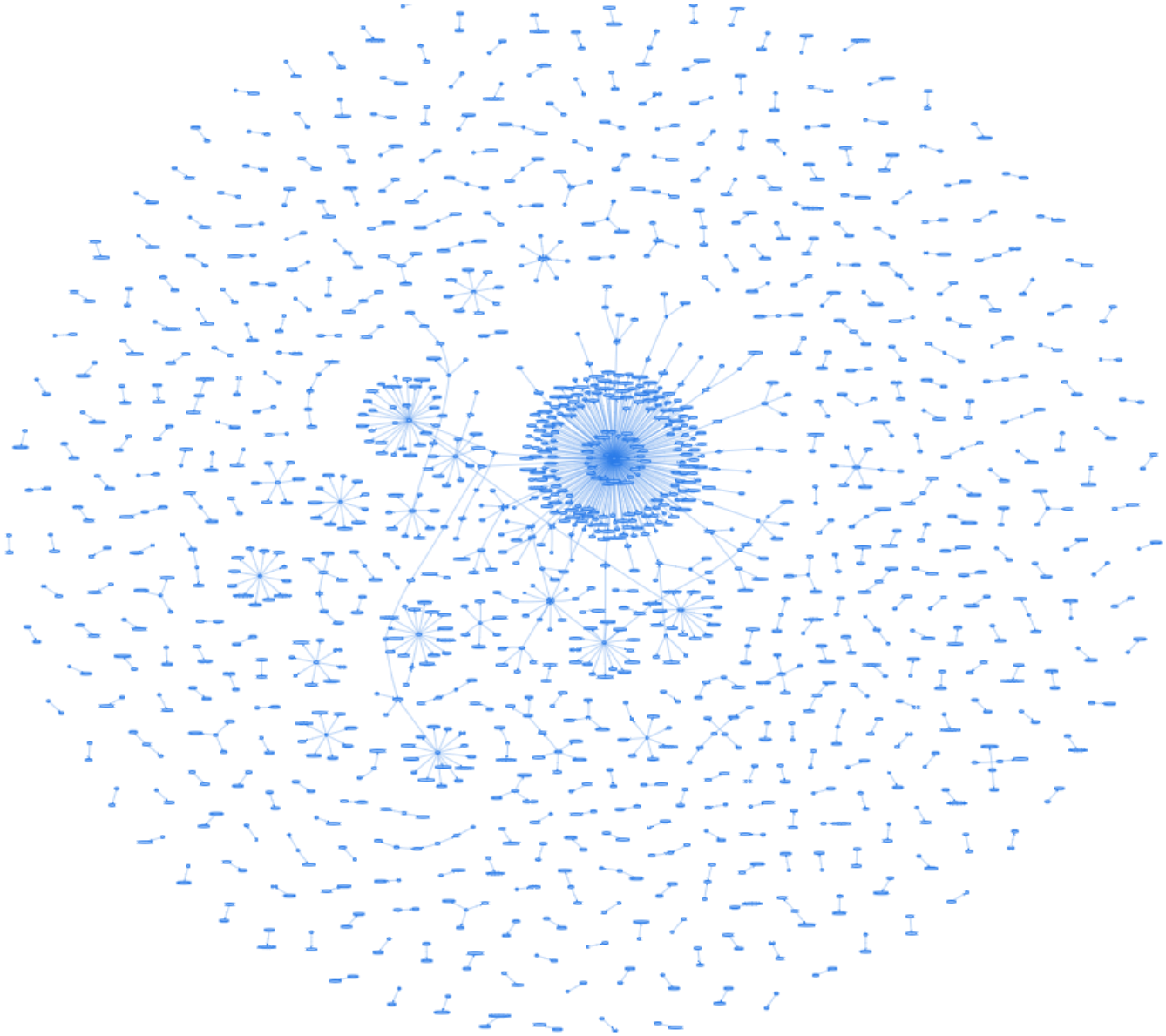

Модель трансформер использована для получения векторного представления текстовых документов на уровне предложений с последующим усреднением. Формализация документов с помощью предобученной модели, предоставленной Сбертех, позволяет сократить размерность признакового пространства (использован слой с размерностью 1024)

# T-SNE проекция формализованного с помощью предобученной модели-трансформера BERT корпуса текстов

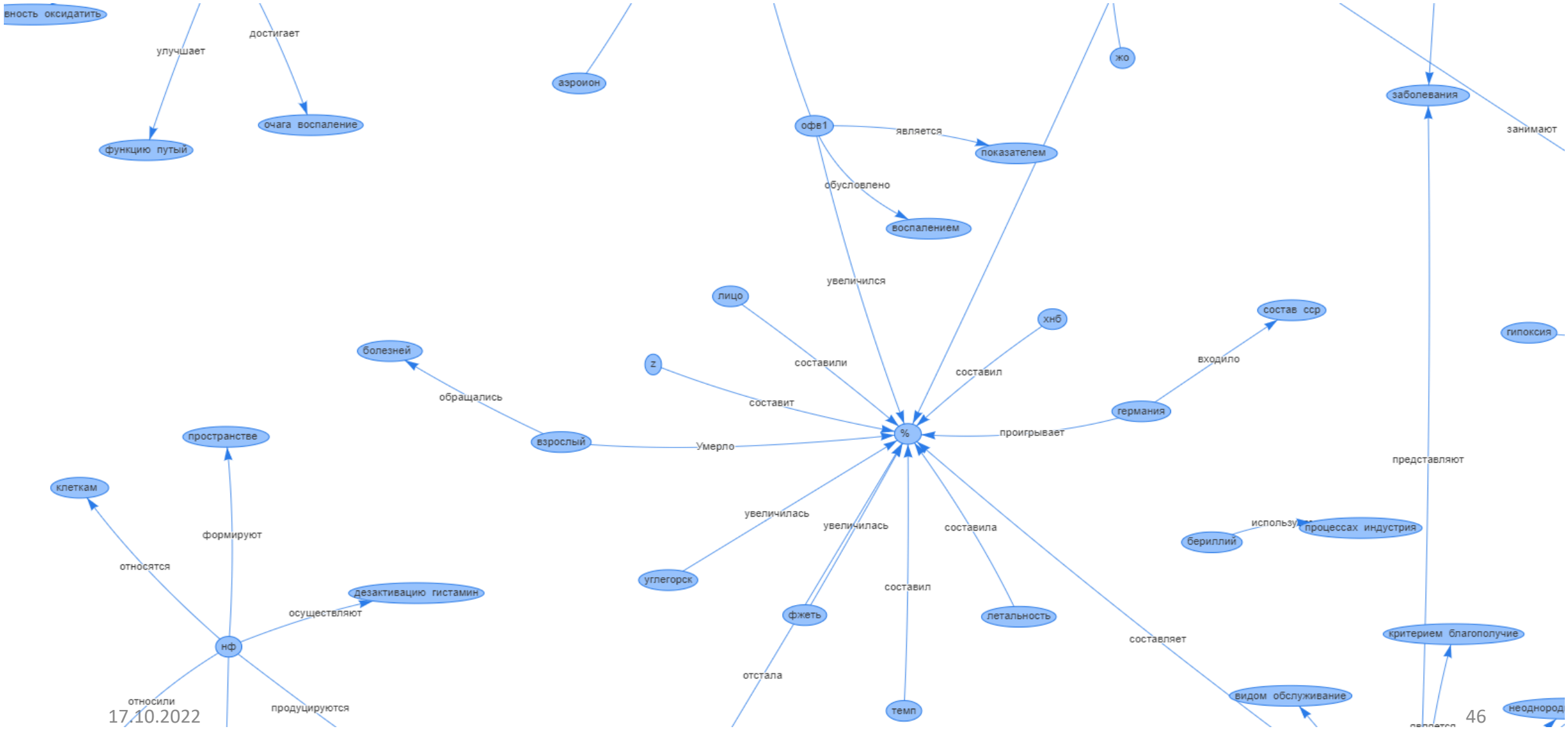



Формализованное представление документов, полученное с помощью модели трансформера и спроецированное на двухмерную плоскость с помощью Стохастического вложения соседей с t-распределением, представлено на слайде. Структура документов позволяет выделить группы документов на основе оценки их семантической близости. Необходимо применение метода кластеризации для оценки границ кластеров

# Фрагмент структуры построенного графа знаний



# Фрагмент структуры построенного графа знаний



# Анализ результатов

- применение моделей машинного обучения и интеллектуального анализа на основе тонкой настройки существующих нейросетевых лингвистических моделей, построенных на обобщенных корпусах текстов, позволит существенно повысить качество анализа исходных, «сырых» слабоструктурированных данных и снизить требования к предварительно построенным глоссариям и кодификаторам;
- реализация классического метода Байеса для многоклассовой классификации корпуса текстовых документов продемонстрировал наилучший результат: лучшая оценка точности 0,92 и сравнительно хорошая полнота 0,86;
- модели общего назначения позволяют предварительно разметить корпус для дальнейшей верификации и уточнения разметки, а также построения специализированных моделей (оценка F1-меры для модели общего назначения – 20,4% по сравнению с вариантом использования словаря – 16,7%).
- применение языковых моделей трансформеров для выделения триплетов позволяет существенно расширить возможности по формализации знаний, построению графов знаний и решению задач построения систем поддержки принятия решений в клинической практике. Необходимым этапом является детализированная разметка корпуса текстовых документов для учета специфической лексики и обилия сокращений.
- для русскоязычного сегмента представлены лишь отдельные достаточно скромные по размеру корпусы размеченных медицинских текстов, что объясняется высокой трудоемкостью их сбора и отсутствием общепринятого протокола разметки (по сравнению с англоязычными решениями), учитывающего структуру и номенклатуру отечественной документации;
- отдельной проблемой является формирование открытых кодификаторов и глоссариев именованных сущностей для построения моделей NER и дальнейшей автоматизации конструирования графов знаний (для русскоязычного сегмента);
- недостаточное количество полилингвальных языковых предметно-ориентированных моделей семейства BERT, T5 и т.д;
- необходимость интеграции и адаптации методов построения графовых моделей, языковых моделей и традиционных моделей баз знаний из других предметных областей.