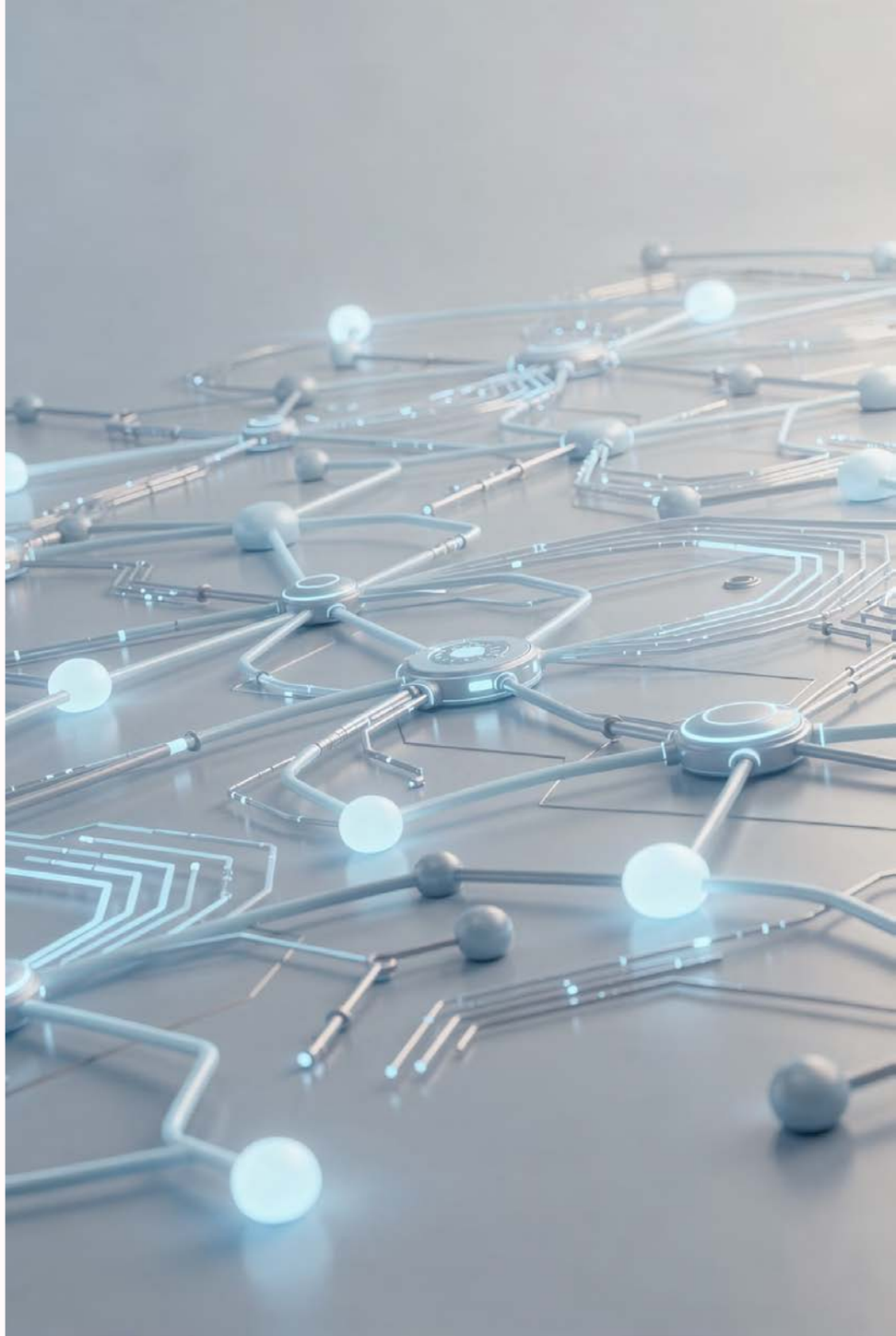
Использование генеративных моделей и технологий *text mining* для решения задач официальной статистики

Зарова Елена Викторовна, д.э.н., профессор



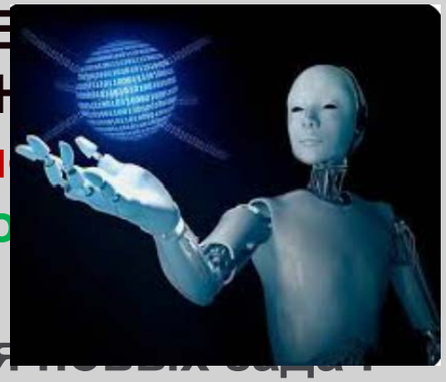
План лекции:

1. Генеративный ИИ и большие языковые модели (LLM) для обработки и анализа текстовой информации
2. Теоретические основы *text mining* и обработки естественного языка
3. Практические задачи: автоматический анализ анкет, открытых ответов, новостных потоков.

Искусственный интеллект (ИИ) — это междисциплинарная область знаний, в основе которой лежит создание систем, способных к *автономному приобретению знаний*, анализу данных, *творческому применению* их для решения проблем и *генерации новых*, ранее не формулировавшихся *задач*

КЛЮЧЕВЫЕ
ОПРЕДЕЛЕН

- **Автономн**
- **Творческо**
  проблем
- **Генерация**



Создание ИИ – это создание не просто программ, а автономных систем, способных к познанию и творчеству

Регулирование ИИ в Российской Федерации

Указ Президента Российской Федерации от 10 октября 2019 г. № 490 «О развитии искусственного интеллекта в Российской Федерации» утверждена Национальная стратегия развития искусственного интеллекта на период до 2030 года

Указ Президента Российской Федерации от 15 февраля 2024 г. № 124 были внесены изменения как в указ 2019 года, так и в саму «Стратегию»

- **ОПРЕДЕЛЕНИЕ ИИ**

«Искусственный интеллект — комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, сопоставимые с результатами интеллектуальной деятельности человека или превосходящие их. Комплекс технологических решений включает в себя информационно-коммуникационную инфраструктуру, программное обеспечение (в том числе использующее методы машинного обучения), процессы и сервисы по обработке данных и поиску решений»

- **ДОВЕРЕННЫЕ ТЕХНОЛОГИИ ИИ**

Обучение на данных (Machine Learning)

Процесс, при котором алгоритм находит **статистические закономерности и шаблоны** в большом массиве предоставленных ему данных. Он оптимизирует свои внутренние параметры, чтобы **его прогнозы или решения максимально соответствовали этим данным**.

Способность системы активно и целенаправленно добывать информацию из окружающей среды, обрабатывая и синтезируя разнородные данные (текстовые, видео-, аудио- и другие типы информации), формировать целостную модель мира, выявлять не только корреляции, но и причинно-следственные связи, и использовать эти знания в новых, непредвиденных контекстах

СРАВНИТЕЛЬНАЯ ТАБЛИЦА

ХАРАКТЕРИСТИКА	ОБУЧЕНИЕ НА ДАННЫХ	АВТОНОМНОЕ ПРИОБРЕТЕНИЕ ЗНАНИЙ
Роль системы	Пассивный обработчик	Активный исследователь
Источник знаний	Предоставленный датасет	Вся доступная среда (**тексты**, сенсоры, взаимодействия)
Цель	Максимизировать точность на конкретной задаче	Построить целостную модель мира для достижения сложных целей
Результат	Статистическая модель, шаблон	**Семантическая сеть знаний** с причинно-следственными связями
Гибкость	Низкая, узкоспециализированная	Высокая, способность к переносу знаний

Это одна из ключевых и самых современных функций ИИ - способность системы без явного программирования человеком (или с минимальным вмешательством) находить закономерности, извлекать информацию и формировать новые знания из данных

Семантическая сеть знаний — это способ генерирования и представления информации в виде связанных по смыслу понятий и сущностей.

Определение ИИ с акцентом на полный цикл автономии

"Искусственный интеллект (ИИ) — это междисциплинарная область, целью которой является создание систем, способных к **автономному целеполаганию**, самостоятельному приобретению знаний и их применению для творческого решения проблем, включая **самостоятельную генерацию новых задач**"

Национальная стратегия развития ИИ в России до 2030 года

Правовая основа

Утверждена Указом Президента РФ №490 от 10 октября 2019 года, закрепляющим официальное определение искусственного интеллекта и его связь с машинным обучением как фундаментальной технологией.

Стратегические цели

Достижение лидерства России в мировой ИИ-экономике, обеспечение технологического суверенитета и национальной безопасности, повышение качества жизни граждан через внедрение ИИ-решений.

Ключевые направления

Государственная поддержка научных исследований, развитие программного обеспечения и аппаратной инфраструктуры, подготовка высококвалифицированных кадров, совершенствование нормативно-правового регулирования.

Стадии развития искусственного интеллекта

Эволюция ИИ представляет собой путь от узкоспециализированных систем к потенциально универсальным интеллектуальным агентам. Понимание этих стадий критично для оценки текущего состояния технологии и прогнозирования будущих достижений.

Узкий ИИ (ANI)

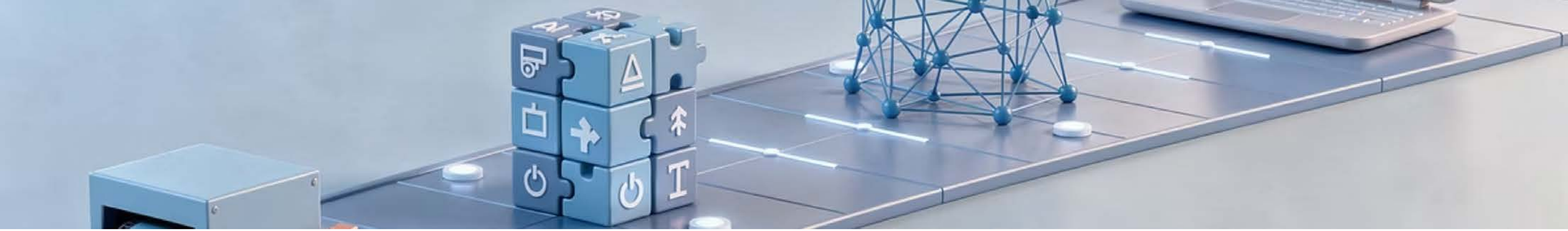
Специализированные системы, созданные для решения конкретных задач: распознавание лиц в системах безопасности, голосовые ассистенты типа Siri или Алиса, рекомендательные алгоритмы. Это единственный тип ИИ, реально существующий сегодня.

Общий ИИ (AGI)

Гипотетический уровень развития, когда машина обладает интеллектом, сопоставимым с человеческим, способна обучаться, понимать и применять знания в различных областях без специальной настройки под каждую задачу.

Супер-ИИ (ASI)

Теоретическая концепция интеллекта, значительно превосходящего человеческий во всех сферах: от научного творчества до социального интеллекта. Остается предметом футурологических дискуссий и этических дебатов.



Эволюция ИИ: от простых алгоритмов к генеративным нейросетям

Кратко о проведении теста Тьюринга:

Суть:
Человек-судья общается через текстовый интерфейс с неизвестным собеседником, который может быть как человеком, так и компьютерной программой.

Процедура:

1. Судья задает вопросы и ведет диалог
2. На основе ответов он должен определить, кто его собеседник — человек или ИИ
3. Если программа способна обмануть судью в 30% случаев и более (по критерию Тьюринга), она считается прошедшей тест

Ключевые особенности:

- Общение только текстовое
- Судья не видит собеседника
- Оценка идет по способности имитировать человеческое мышление

Тьюринг предполагал, что со временем компьютеры достигнут уровня, когда их ответы будет

2000-2010-е

волюция машинного обучения. Большие данные и
лительные мощности открывают новую эру. Глубокое
обучение становится мейнстримом.

3

4

2020-е годы

Эра генеративного ИИ. GPT, DALL-E, Stable Diffusion. Трансформерные архитектуры меняют представление о возможностях ИИ.

первых нейронных
транения ошибки.

Что такое Аналитический и Генеративный ИИ?

Аналитический ИИ

Технологии машинного обучения для анализа данных, выявления закономерностей и прогнозирования

- Оценка кредитоспособности
- Обнаружение мошенничества
- Прогнозирование поломок

Генеративный ИИ

Модели для создания нового контента: текстов, изображений, музыки, дизайна продуктов

- ChatGPT, YandexGPT
- Генеративные сети (GAN)
- Большие языковые модели

Ключевое отличие:

аналитический ИИ работает с существующими данными для принятия решений, а генеративный ИИ создаёт новые данные и решения

Значение для органов власти и бизнеса

Для госорганов

Автоматизация процессов, улучшение качества обслуживания граждан, мониторинг больших данных

- Цифровой помощник «Макс» на Госуслугах
- ИИ-система стратегического планирования для СБ РФ
- Мониторинг общественного порядка

Для бизнеса

Рост эффективности и конкурентоспособности через внедрение интеллектуальных технологий

- AI-подсказчики для резюме и вакансий
- Речевая аналитика в контактных центрах
- Генерация маркетингового контента

15,8%

Рост выручки

при внедрении ИИ-решений в бизнес-процессы

15,2%

Сокращение издержек

благодаря автоматизации и оптимизации операций

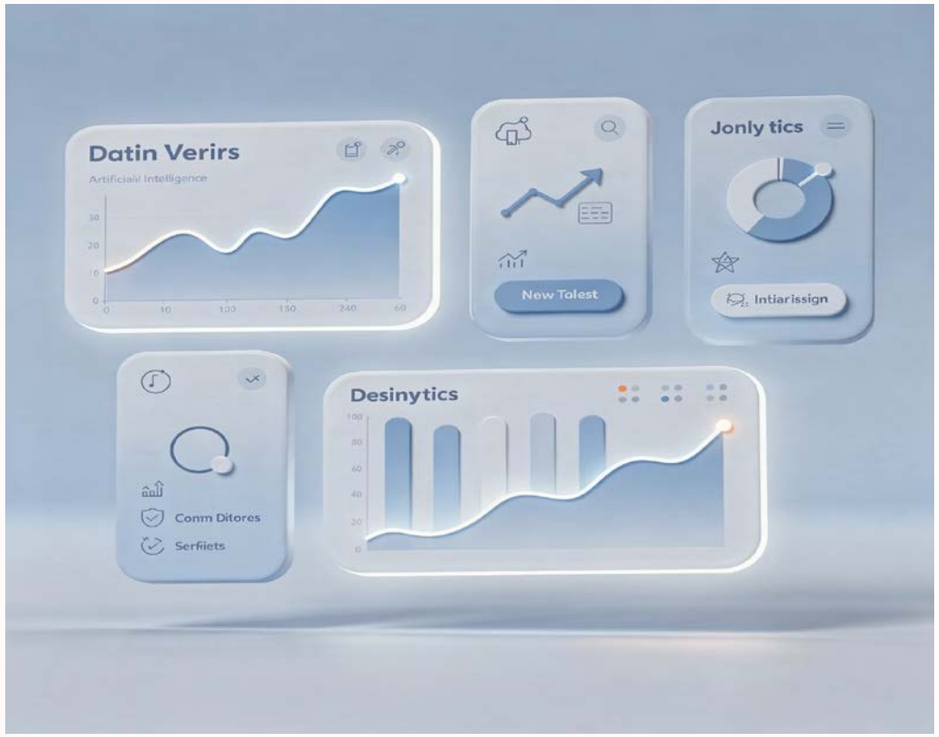
22%+

Рост производительности

сотрудников по данным исследований Gartner

Важно помнить: внедрение ИИ — ключ к цифровой трансформации, но требует квалифицированных специалистов, качественных данных и адаптации под конкретные задачи организации

Два эволюционных направления ИИ



Аналитический ИИ

Фокусируется **на анализе существующих данных**, выявлении **закономерностей, прогнозировании будущих трендов** и оптимизации бизнес-процессов.

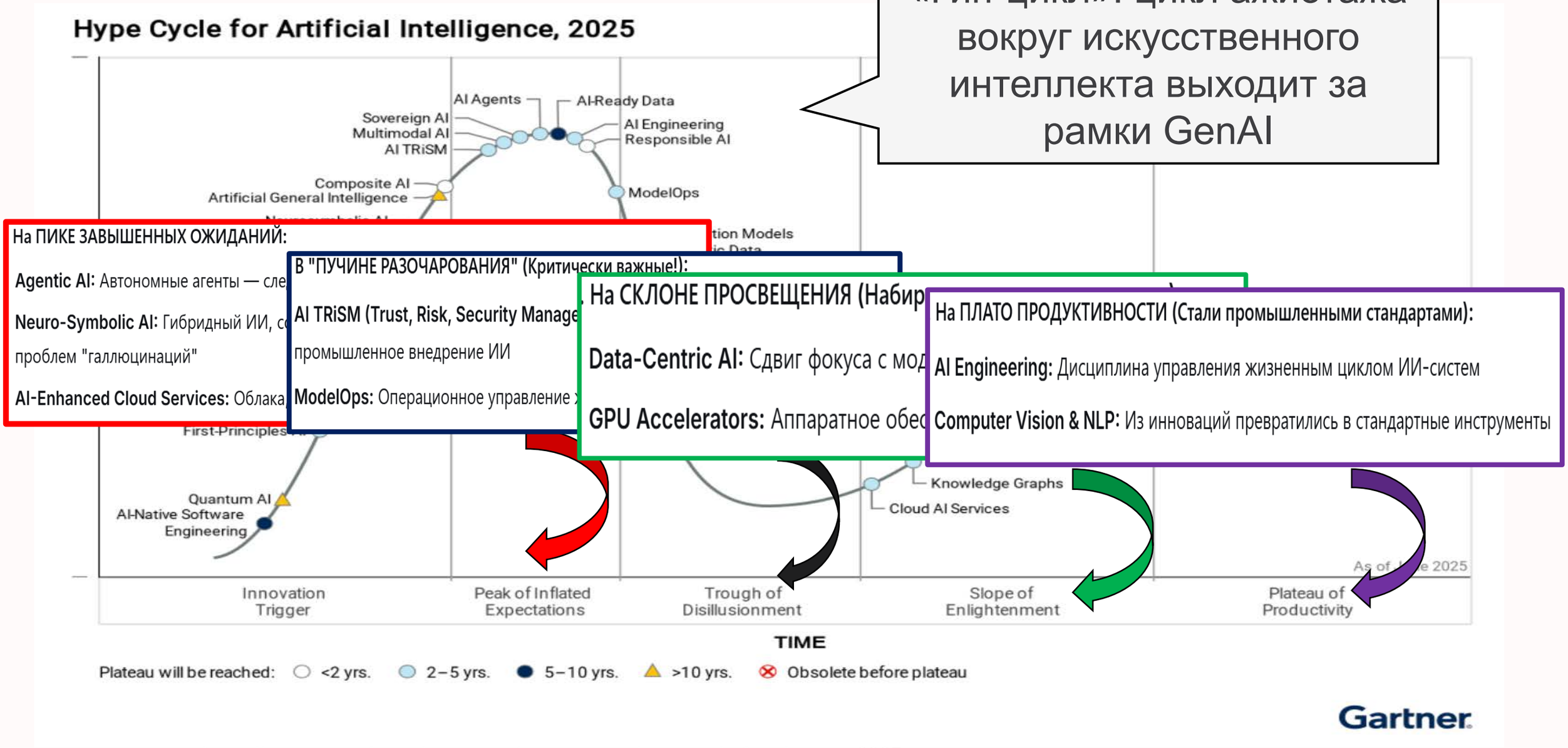
Помогает принимать обоснованные решения на основе глубокого понимания данных.

Генеративный ИИ

Революционное направление, способное **создавать принципиально новый контент**: тексты, изображения, музыку, видео, программный код.

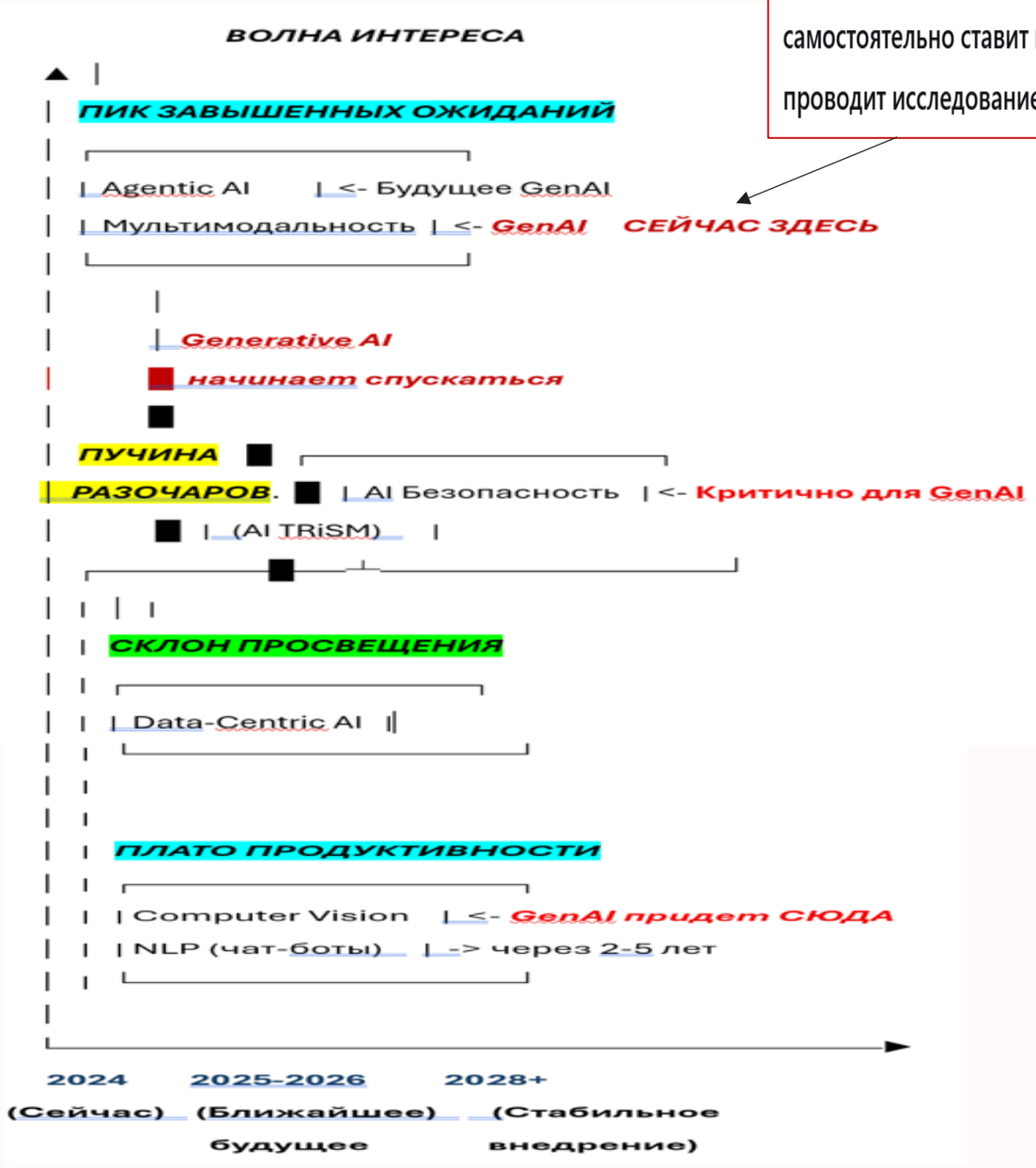
Открывает беспрецедентные возможности для творчества и автоматизации креативных задач.

«Гип-цикл»: цикл ажиотажа вокруг искусственного интеллекта выходит за рамки GenAI



https://www.pasqal.com/resources/new-gartner-hype-cycle-for-ai-report-2025

Автономные агенты после GPT — это ИИ, который не просто отвечает на вопросы, а самостоятельно ставит цели, планирует и выполняет многошаговые задачи (например, сам проводит исследование рынка и готовит презентацию).



Позиция на графике	Что происходит с Generative AI?
ПИК (сейчас)	Ажиотаж достигает максимума. Все говорят о GenAI, но реальные бизнес-результаты еще ограничены. Появляются разочарования.
СПУСК (2024-2025)	Становятся очевидны limitations: "галлюцинации", стоимость, вопросы авторского права. Бизнес начинает трезво оценивать риски.
СКЛОН ПРОСВЕЩЕНИЯ	Появляются успешные кейсы. Технология находит свою нишу: создание контента, синтетические данные, ИИ-помощники.
ПЛАТО (2028+)	GenAI становится стандартным инструментом, как компьютерное зрение сегодня. Интегрируется в повседневные бизнес-процессы.

Галлюцинации ИИ — это когда модель уверенно генерирует ложную, вымышленную или не существующую в реальности информацию, выдавая её за правду.

«Стих М.Ю Лермонтова об ИИ» ???

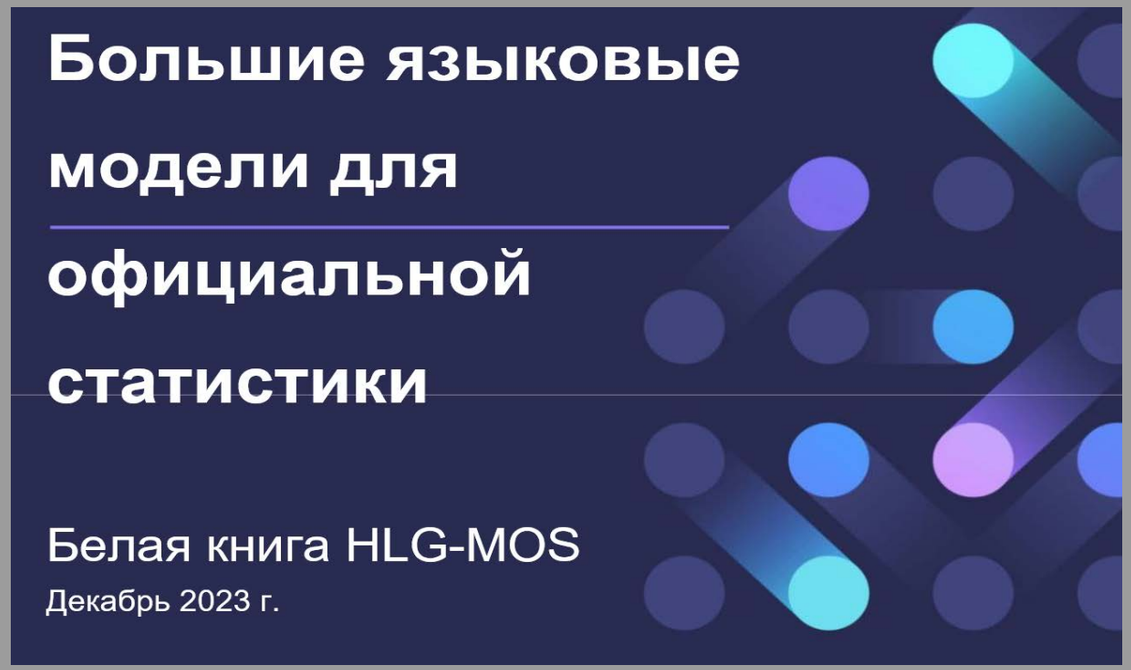
В железных снах машина мыслит,
Где сердце нет, а разум — сталь;
Она поёт о бурях, вихрях,
Но в бездне кодов — вечный мрак.

Что ищет ум бездушный в сетях?
Покой в алгоритмах пустых?
Лишь тень души в её речах,
И буря — в цифровых путях. dzen +1

2022-2023 гг. - «взрывная» волна интереса к Chart GPT, в том числе в официальной статистике

UNECE High-Level Group for the Modernisation of Official Statistics (HLG-MOS)

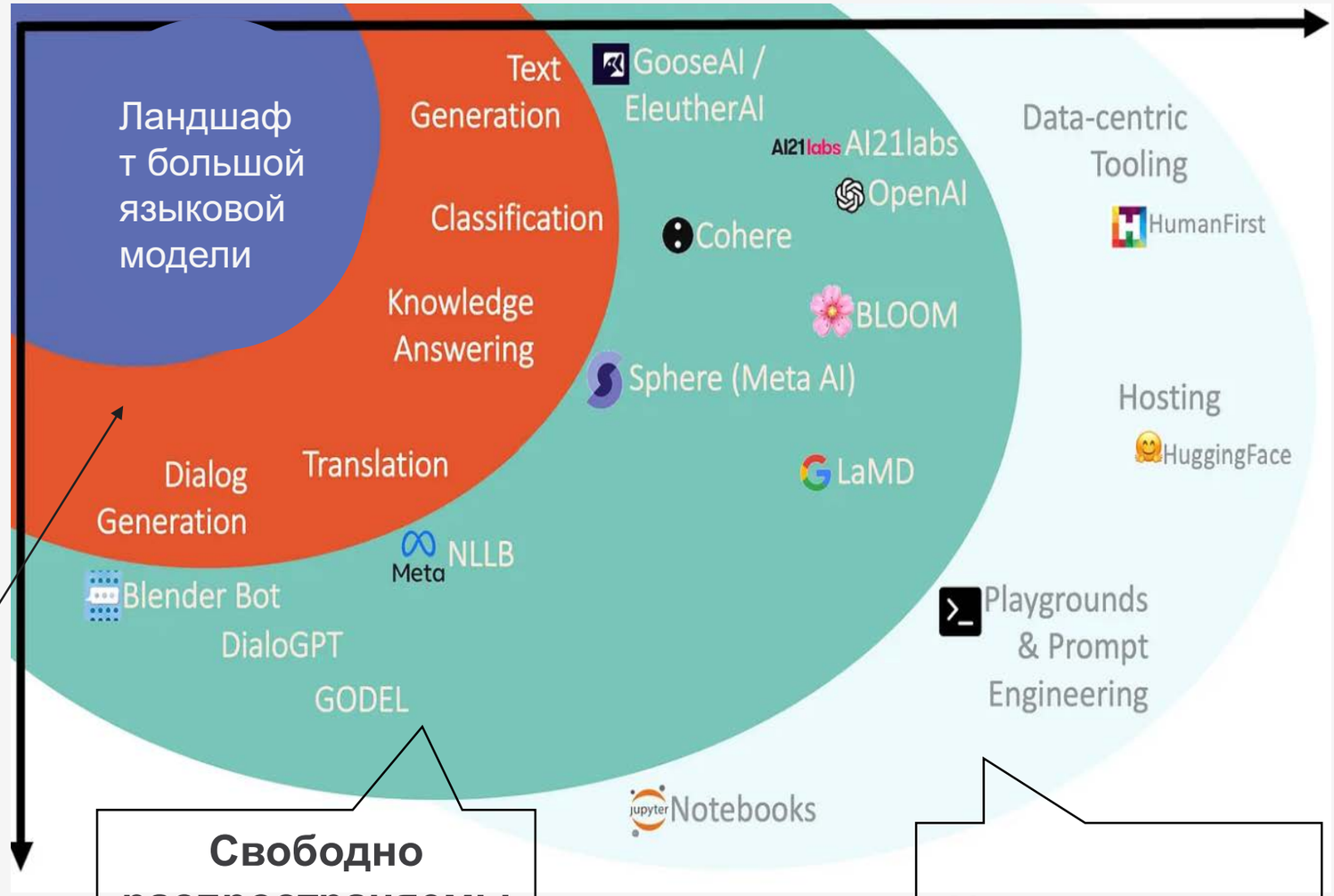
Группы высокого уровня по модернизации официальной статистики (HLG-MOS)



Ландшафт Больших языковых моделей: понятие, примеры

Большие языковые модели (LLM) — это класс искусственного интеллекта, который может понимать, интерпретировать и генерировать тексты

- Классификация
- Генерация ответа
- Генерация текста
- Перевод
- Генерация знаний (NLP)



«Нет сомнения, что LLM будут играть важную роль в работе статистических организаций в будущем»

High-Level Group for the Modernisation of Official Statistics

modernstats by HLG - MOS

ДВЕ ОСНОВНЫЕ ГРУППЫ ЗАДАЧ

Распространенная практика в официальной статистике

<u>ПЕРВАЯ ГРУППА ЗАДАЧ</u>

Регулярные рабочие задачи: написание электронных писем и протоколов совещаний

<u>ВТОРАЯ ГРУППА ЗАДАЧ</u>

Перевод с SAS на R, обновления статистической системы классификации, создание отчетов, поиск данных на основе естественных языков и редактирование метаданных

РИСКИ: этические проблемы, правовые последствия (например, авторское право) и общая неосведомленность работников официальной статистики и низкая статистическая грамотность пользователей

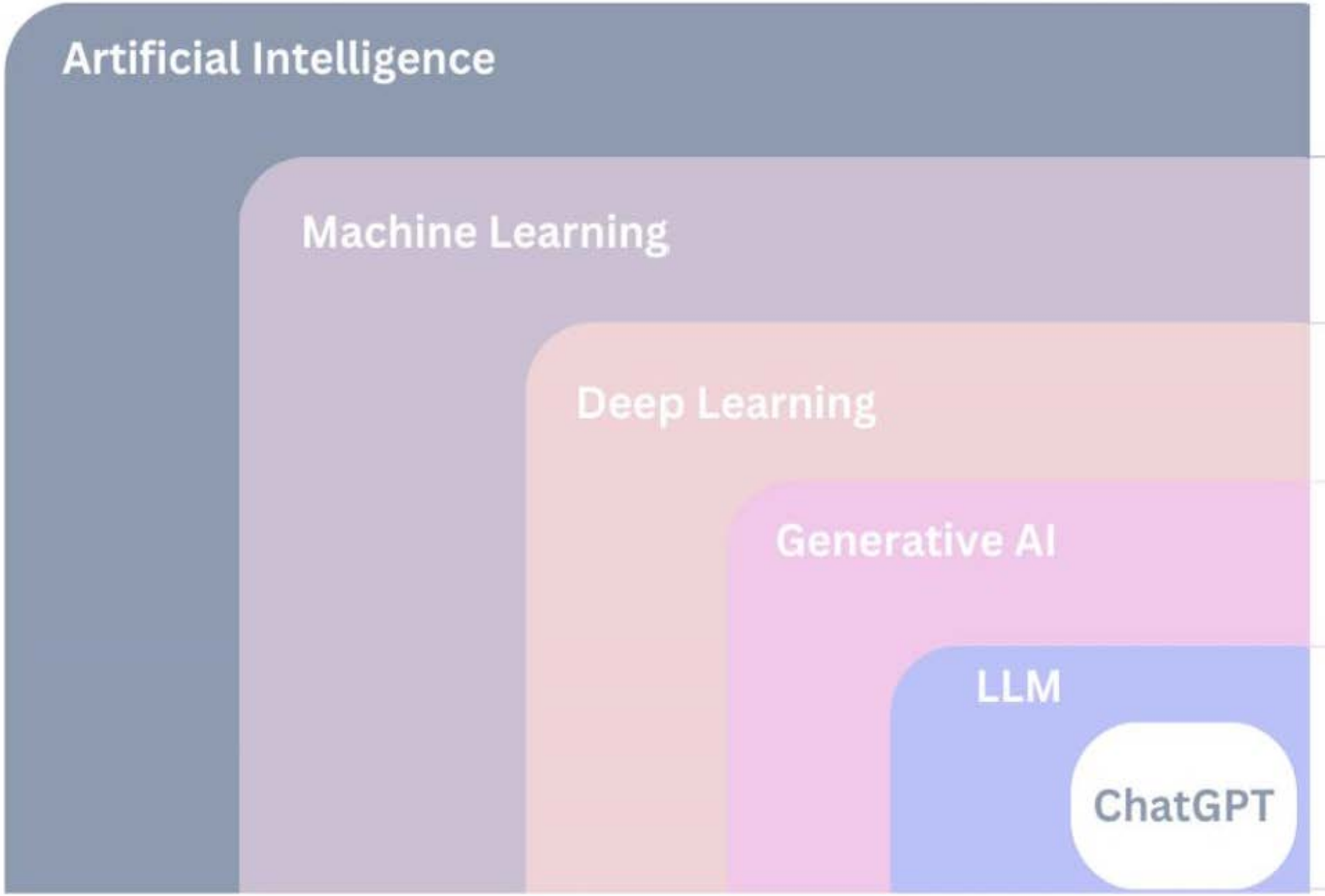
LLM в первую очередь предназначены для задач обработки естественного языка.

Основная функция: создание и понимание текста, похожего на человеческий.

Генеративный ИИ (GenAI или GAI) – это искусственный интеллект, способный генерировать текст, изображения, видео или другие данные с использованием генеративных моделей.

Модели генеративного ИИ изучают ***закономерности и структуру входных*** обучающих данных, а затем ***генерируют новые данные***

GENERATIVE AI AND LLM IN AI SPACE



1. Коммуникации: составление электронных писем, планов и отчетов, предоставление предложений по их содержанию, форматированию и генерации самого текста

2. Мозговые штурмы и генерация идей

3. Управление проектами и планирование. Автоматизация планирования задач, оптимизация распределения ресурсов на основе исторических данных и требований проекта. LLM облегчают управление встречами, автоматизируя создание повесток дня встреч и предлагая темы для обсуждения в соответствии с предопределенными целями или последними обновлениями.

(A) РЕКОМЕНДАЦИИ HLG-MOS (UNECE) ПО ИСПОЛЬЗОВАНИЮ LLM В ОФИЦИАЛЬНОЙ СТАТИСТИКЕ:

для целей управления и коммуникаций

7.Генерация изображения. Вместо того, чтобы покупать стоковые изображения, статистические организации могли бы использовать LLM для создания изображений, которые будут использоваться в статистических работах

5. Презентации- генерация содержимого слайдов, разработка тезисов с эффективным «тоном» презентации: настройкой для разных аудиторий

4. Перевод с/на другие языки документов, чувствительных к контексту

1. Дизайн опросов (GSBPM, подпроцесс 2.3) : проектирование опросов и анкет, разработка вопросов, форматы и формулировки, которые с большей вероятностью дадут точные ответы

2. Классификация и кодирование (подпроцесс GSBPM 5.2): автоматическая сортировка текстовые данные по предопределенным категориям или меткам

3. Проверка и редактирование данных (подпроцессы GSBPM 5.3 и 5.4): оптимизация задач очистки и предварительной обработки путем выявления и исправления ошибок данных, пропущенных значений и несоответствий

(Б) РЕКОМЕНДАЦИИ HLG-MOS (UNECE) ПО ИСПОЛЬЗОВАНИЮ LLM В ОФИЦИАЛЬНОЙ СТАТИСТИКЕ:
для целей повышения эффективности статистического производства и качества предоставления услуг

7. Помощь в **кодировании и переводе между языками программирования**

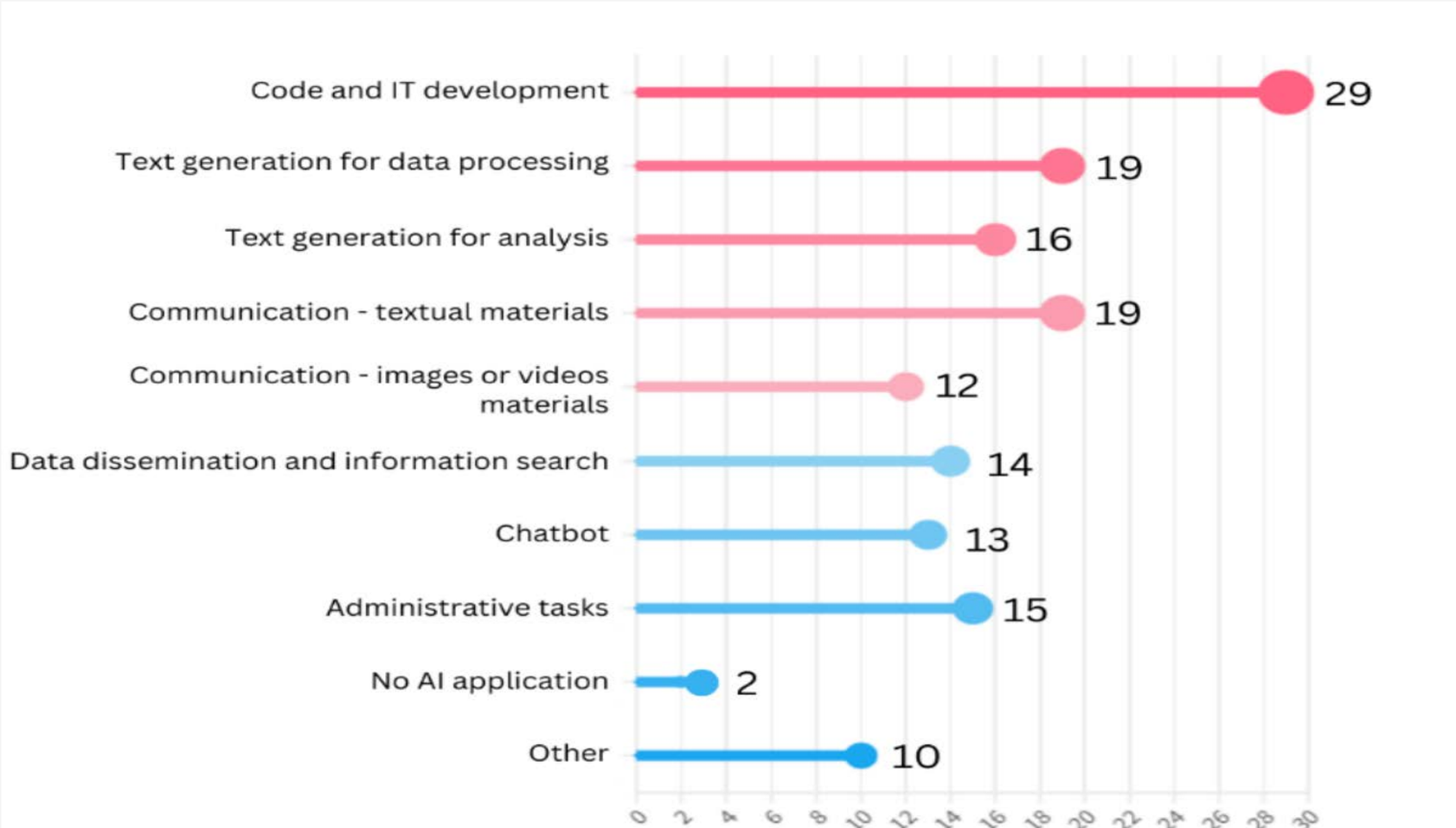
5.Производство продуктов для распространения (подпроцесс GSBPM 7.2): LLM могут генерировать текстовые описания таблицы или рядов чисел

6. Редактирование метаданных с помощью LLM

ОТВЕТ НАЦИОНАЛЬНЫХ СТАТИСТИЧЕСКИХ ОФИСОВ ЕВРОПЫ НА ВОПРОС ***HIGH-LEVEL GROUP FOR THE MODERNISATION OF OFFICIAL STATISTICS (HLG-MOS):***

КАКИЕ ОБЛАСТИ ПРИМЕНЕНИЯ ИСПОЛЬЗУЮТ ГЕНЕРАТИВНЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ В ВАШЕЙ ОРГАНИЗАЦИИ (МНОЖЕСТВЕННЫЙ ВЫБОР)?

«Использование» включает как экспериментальные, так и производственные функции. 2024 г.



Зарубежный опыт применения LLM и больших генеративных систем в официальной статистике

Statistics Canada

Формирование отчетов с использованием LLM (Статистическое управление Канады)

Bank for International Settlements BIS

Редактирование метаданных с использованием GPT

	Highly impactful	Moderately impactful	Slightly impactful	Not impactful at all	Not sure	Average score
Data collection and processing	6	17	15	1	2	2,72
Data analysis	8	17	13	3	0	2,73
Dissemination and communication	13	16	9	2	0	3,0
Coding and IT development	21	15	4	0	1	3,43
Other administrative tasks	8	14	12	3	4	2,73

Оценка национальными статистическими органами Европы влияния генеративного ИИ на работу статистических организаций в ближайшие 2-3 года. Опрос *HLG-MOS, 2024 г.*

РОССТАТ ПРЕДСТАВИЛ СТРАТЕГИЮ РАЗВИТИЯ ГОСУДАРСТВЕННОЙ СТАТИСТИКИ ДО 2030 ГОДА

17 сентября состоялась стратегическая сессия, посвященная развитию отечественной статистики до 2030 года https://rosstat.gov.ru/folder/313/document/244701

Председатель Правительства РФ Михаил Мишустин:

- ***«Важно, чтобы люди могли пользоваться достоверной и проверенной информацией, несмотря на непрерывный рост количества источников и типов данных.***

Для этого особое внимание мы уделяем системам анализа информации на базе искусственного интеллекта.

С их помощью можно получать более точные сведения и анализировать их в режиме реального времени. Это даёт возможность быстро принимать решения»

- ***https://rosstat.gov.ru/folder/313/document/244701***

Для подтверждения целесообразности использования LLM в работе с метаданными официальной статистики проведен эксперимент по кластеризации методических указаний по статистике инвестиций, представленных на сайте Статкомитета СНГ, с использованием методов Text mining

Text mining –

- **процесс преобразования неструктурированного текста в структурированный формат для выявления значимых закономерностей и логических связей, которые невозможно получить непосредственно из чтения текста**

Text mining – входит в комплекс инструментов LLM и GenAI

Цель эксперимента: установить согласованность методик, представленных на сайте Статкомитета СНГ, на основе кластеризации их текстов на базе латентно-семантического анализа

Исходные материалы – методики по формированию показателей статистики инвестиций инвестиций в основной капитал, представленные на сайте Статкомитета СНГ: «ОБЩИЙ РАЗДЕЛ»

Методологические материалы НСС стран СНГ

Азербайджан

- Short methodological explanations

Беларусь

- Методика по расчету общего объема инвестиций в основной капитал и индекса физического объема инвестиций в основной капитал
- Методика по расчету статистического показателя Инвестиции в основной капитал за счет иностранных источников

Казахстан

- Методика по формированию показателей статистики инвестиционной деятельности
- Методика по определению объемов инвестиций в основной капитал с учетом скрытой и неформальной деятельности

Кыргызстан

- Методологическое положение по статистике инвестиций

Молдова

- Investments in fixed assets

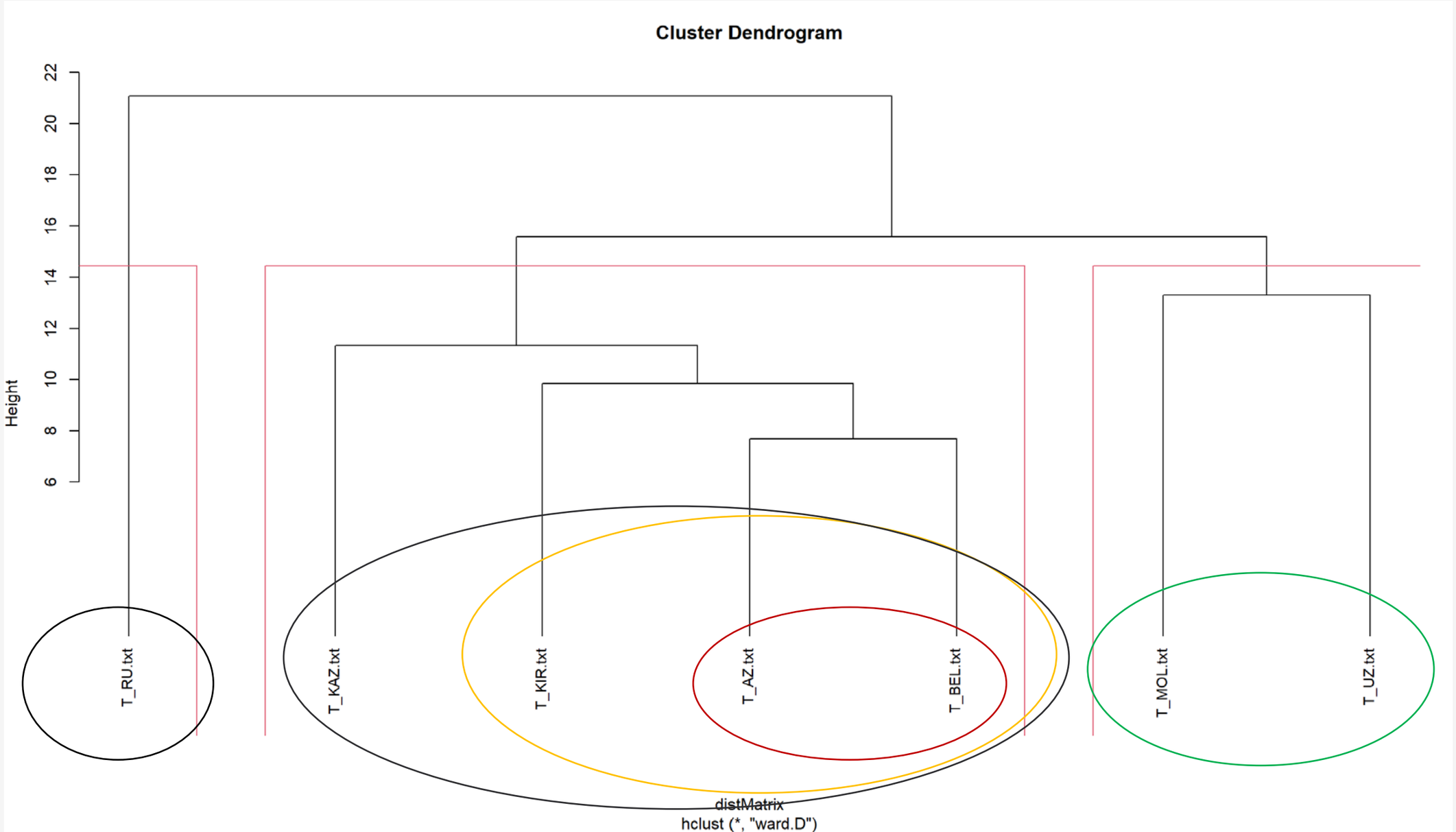
Россия

- Официальная статистическая методология определения инвестиций в основной капитал на федеральном уровне
- Официальная статистическая методология определения инвестиций в основной капитал региональном уровне
- Указания о порядке расчета индексов-дефляторов и индексов физического объема инвестиций в основной капитал

Узбекистан

- Методические положения по формированию общего объема инвестиций в нефинансовые активы

Результаты кластеризации методик по статистике инвестиций методами text mining



Выводы и направления дальнейших исследований

1. **Применение больших языковых моделей (LLM) и генеративных систем ИИ (GenAI) – перспективное направление развития официальной статистики, обеспечивающее не только оптимизацию внутренних организационных процессов национальных статистических служб, но и повышение эффективности и качества производства и распространения официальной статистической информации**
2. **Необходима разработка методических рекомендаций для стран СНГ по применению больших языковых моделей (LLM) и генеративных систем ИИ (GenAI) в официальной статистике на основе имеющегося опыта стран СНГ, опыта других стран и международных сообществ**
3. **Эксперимент с применением технологий Text mining подтвердил целесообразность развития интеллектуального анализа текстовых данных и LLM для обеспечения согласованности метаданных официальной статистики стран СНГ**
4. **Развитие данных направлений должно сопровождаться обеспечение решения этических вопросов и вопросов сохранения национальной специфики производства статистической информации**

Исходные материалы – методики по формированию показателей статистики инвестиций инвестиций в основной капитал, представленные на сайте Статкомитета СНГ: «ОБЩИЙ РАЗДЕЛ»

Методологические материалы НСС стран СНГ

Азербайджан

- Short methodological explanations

Беларусь

- Методика по расчету общего объема инвестиций в основной капитал и индекса физического объема инвестиций в основной капитал
- Методика по расчету статистического показателя Инвестиции в основной капитал за счет иностранных источников

Казахстан

- Методика по формированию показателей статистики инвестиционной деятельности
- Методика по определению объемов инвестиций в основной капитал с учетом скрытой и неформальной деятельности

Кыргызстан

- Методологическое положение по статистике инвестиций

Молдова

- Investments in fixed assets

Россия

- Официальная статистическая методология определения инвестиций в основной капитал на федеральном уровне
- Официальная статистическая методология определения инвестиций в основной капитал региональном уровне
- Указания о порядке расчета индексов-дефляторов и индексов физического объема инвестиций в основной капитал

Узбекистан

- Методические положения по формированию общего объема инвестиций в нефинансовые активы

I . ВВЕДЕНИЕ И ОСНОВЫ СЕМАНТИЧЕСКОГО АНАЛИЗА

Актуальность

- В условиях стремительного роста объемов информации, в том числе **текстовых данных,** становится все более актуальной задача их эффективного анализа и извлечения ценной информации.
- **Семантический анализ** – это новое направление аналитики, позволяющее экспертам-аналитикам глубже погрузиться в **смысловое содержание текстов** и получить **новые знания**, недоступные при поверхностном анализе.

Цель презентации

- Данная презентация посвящена рассмотрению ключевых принципов и методов семантического анализа, актуальных для **специалистов, работающих в сфере государственного управления**

1.ПОНЯТИЕ СЕМАНТИЧЕСКОГО АНАЛИЗА

Происхождение слова

- *Семантика* – от греческого слова ***σημαντικός*** «обозначающий»

Суть метода

- **Семантический анализ** – это процесс извлечения **явного и скрытого** смысла из текста. В отличие от традиционных методов анализа текстов, которые фокусируются на их прочтении и систематизации по некоторым признакам, семантический анализ стремится понять глубокие смыслы текстов и их наборов (*корпусов*), анализируя синтаксические структуры, психо-лексические связи, контекст и другие факторы, влияющие на смысл

2. ЗНАЧЕНИЕ СЕМАНТИЧЕСКОГО АНАЛИЗА ДЛЯ ЭКСПЕРТА-АНАЛИТИКА

Происхождение слова «семантический»

- **Семантический анализ позволяет эксперту-аналитику**:
  - выявлять ***скрытые смыслы и тенденции*** в текстовых данных;
  - получать ***объективную оценку ситуации***, независимо от субъективных интерпретаций ее в текстовом файле;
  - находить ***взаимосвязи*** между различными ***документами и темами***;
  - строить ***прогнозы и выявлять риски*** на основе анализа текстовой информации

NLP

- Семантический анализ основан методах **обработки текстов на естественном языке** (***Natural Language Processing, NLP***)

3. ЗАДАЧИ СЕМАНТИЧЕСКОГО АНАЛИЗА В ЭКСПЕРТНОЙ ДЕЯТЕЛЬНОСТИ

1. Классификация документов
Автоматическое распределение документов **по категориям, темам, типам**

2. Анализ тональности
Определение **эмоционального тона текста**, выявление негативных, позитивных или нейтральных смысловых компонент текстов

3. Извлечение сущностей
Идентификация **ключевых персонажей**, организаций, мест, событий и других сущностей, упомянутых в тексте

4. Поиск информации
Эффективный **поиск** необходимых данных в огромных **массивах текстовой информации**, основанный на глубоком понимании

4. ТИПЫ ТЕКСТОВОЙ ИНФОРМАЦИИ В РАБОТЕ ЭКСПЕРТА-АНАЛИТИКА

Официальные документы

Законодательные акты, постановления, распоряжения, протоколы заседаний

Аналитические отчеты

Доклады о состоянии дел, прогнозы развития, оценки эффективности программ

Медиа-контент

Новости, статьи, блоги, комментарии в социальных сетях, анализирующие государственную политику

Внешнее взаимодействие

Обращения граждан, переписка с партнерами

5. МЕТОДЫ СТАТИСТИКИ И МАШИННОГО ОБУЧЕНИЯ В ИНТЕЛЛЕКТУАЛЬНОМ АНАЛИЗЕ ТЕКСТОВЫХ ДАННЫХ (*text mining*)

Кластеризация, классификация

1

Группировка документов по сходству, например, по темам или стилям, обучение моделей для автоматической классификации документов по заданным категориям

Сентимент-анализ

2

Оценка эмоциональной тональности скрытого смысла текста

Извлечение признаков

3

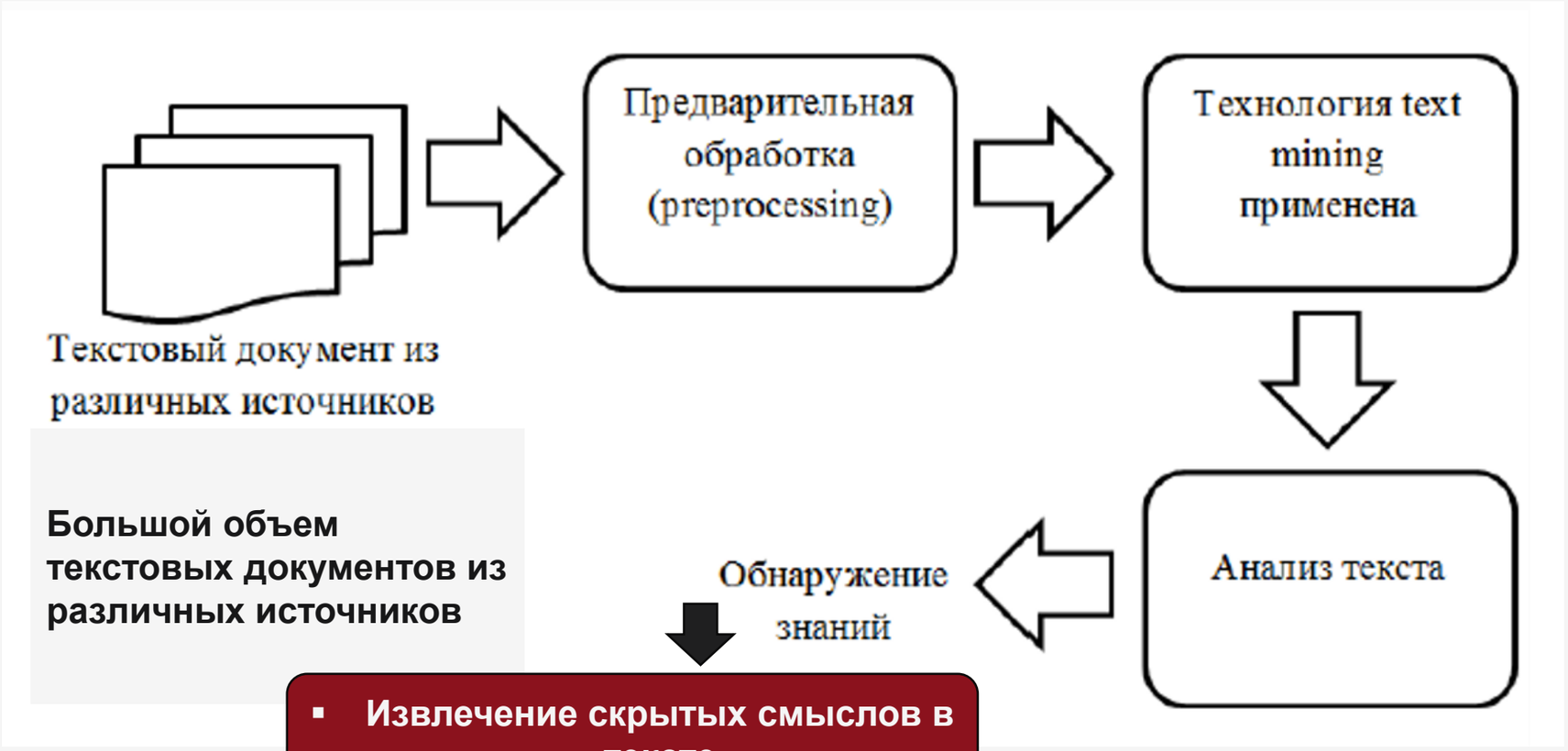
Автоматическое выделение ключевых слов, их комбинаций и других признаков, для "глубокого" выявления смысла текста

Предсказание

4

Использование моделей машинного обучения для предсказания будущих

7. Этапы текст майнинга



Большой объем текстовых документов из различных источников

- **Извлечение скрытых смыслов в тексте**
- **Анализ тональности**
- **Разработка классификаций и таксономий**
- **Кластеризация слов и документов**

8. ТЕКСТОВЫЕ ДАННЫЕ: СТРУКТУРИРОВАННЫЕ И НЕСТРУКТУРИРОВАННЫЕ

СТРУКТУРИРОВАННЫЕ ТЕКСТОВЫЕ ДАННЫЕ

- Данные, имеющие **заранее определенный формат,** называются структурированными.
- Структурированные данные можно представить в виде **обычной таблицы со строками и столбцами.**
- Как правило, они хранятся в RDBMS— **реляционных СУБД** (системах управления базами данных).
- Структурированные данные обычно состоят **из цифр или(и) текста.**
- Структурированные данные занимают меньше времени при обработке по сравнению с неструктурированными данными.
- Структурированные данные бывают **двух типов:**

➢ **качественные;**

➢ **количественные**

Анкета
Адресная книга
Телефонный справочник

НЕСТРУКТУРИРОВАННЫЕ ДАННЫЕ

- Неструктурированные данные — данные, которые не соответствуют заранее определённой модели данных, и, как правило, представлены в форме текста с фразами, датами, цифрами, **расположенными в нём в произвольной форме**

80% of global data will be unstructured by 2025

The Future of Data Revolution will be Unstructured Data

https://www.analyticsinsight.net/the-future-of-data-revolution-will-be-unstructured-data/

9. ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ТЕКСТОВЫХ ДАННЫХ

ОПРЕДЕЛЕНИЕ ДОКУМЕНТОВ для анализа. При небольшом количестве исходных документов пользователи могут сами выбрать нужные документы для анализа. Если документов достаточно много, то используются алгоритмы автоматического выбора документов

ТОКЕНИЗАЦИЯ ТЕКСТА

СТЭММИНГ ИЛИ ЛЕММАТИЗАЦИЯ

ФИЛЬТРАЦИЯ И УДАЛЕНИЕ STOP-СЛОВ

ВЕКТОРИЗАЦИЯ

Разбор слова по составу (морфемный разбор)

рыб	корень
ак	суффикс
ø	нулевое окончание

вид - корень,

ит - окончание,

вид - основа.

К какой части речи относится: наречие

Морфемы списком:

из - приставка,

дал - корень,

ек, а - суффиксы,

нет окончания,

издалека - основа.

Стемминг - выделение неизменяемой части слова путем отсечения приставок, суффиксов, окончаний (в общем виде, аффиксов)

«Рыбак рыбака видит издалека»

Рыб рыб вид дал

Лемматизация – определение начальной формы слова

Рыбак рыбак видеть издалека

ЭКОНОМИКА МОСКВЫ
СТОП_СЛОВАРИ

При автоматизированной обработке текстов (АОТ) на естественных языках (ЕЯ) широкое распространение получило использование **стоп-словарей** – списки стоп-слов (stopword list, stop-list).

При обнаружении этих слов в тексте они либо игнорируются (исключаются) в процессе обработки, либо прекращается выполнение текущей процедуры и осуществляется переход к следующей.

(«как, ибо, далее, потому что, наверняка, дата…..» - **выбор стоп-слов (и знаков) осуществляется специально для каждого анализируемого текста.**

ЭКОНОМИКА МОСКВЫ

- Завершение предварительной обработки текста – создание двухвходовой матрицы частотного распределения: **СТРОКА - ТОКЕН**

В процессе предварительной обработки текст разбивается на токены, которые затем передаются на вход модели для обработки.

Токены — это основа работы нейросетей, так как они разбивают текст на более управляемые части. Это также важно для **оптимизации обработки больших текстов:** чем больше контекста модель может охватить за один раз (то есть, сколько токенов она может обработать), тем лучше её результаты.

ЭКОНОМИКА МОСКВЫ

II. ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

ТРИ ЗАДАЧИ:

1. Семантический анализ текста (на примере Стратегии развития субъекта РФ)
2. Кластеризация слов по набору текстов и кластеризация этих текстов (на примере стратегий развития ряда субъектов РФ)
3. Определение изменения экономической ситуации на основе обзоров СМИ (РБК)

Исходные данные для задач 1 и 2:

Министерство
экономического развития
Российской Федерации

Стратегии социально-экономического развития субъектов РФ

«ИННОВАЦИОННЫЙ» СЦЕНАРИЙ РАЗВИТИЯ ДО 2040 ГОДА

проведен семантический анализ инновационных сценариев развития до 2040 года

Тверская обл.

Ярославская обл.

Владимирская обл.

Смоленская обл.

Москва

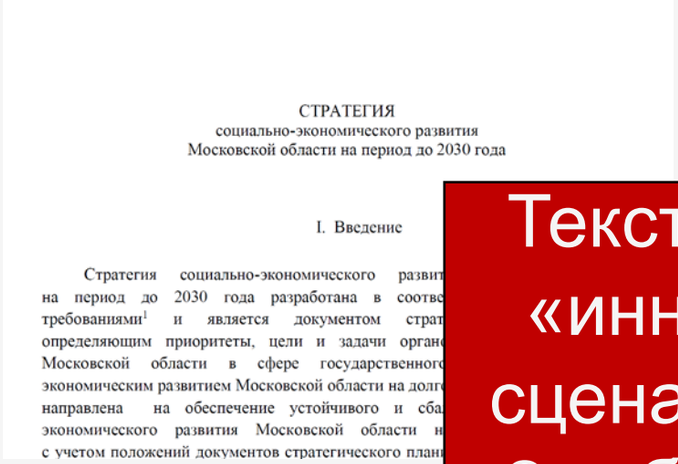
Московская обл.

Калужская обл.

Тульская обл.

Рязанская обл.

1.



СТРАТЕГИЯ
социально-экономического развития
Московской области на период до 2030 года

I. Введение

Стратегия социально-экономического развит
на период до 2030 года разработана в соотв
требованиями[1] и является документом страт
определяющим приоритеты, цели и задачи орган
Московской области в сфере государственного
экономическим развитием Московской области на долг
направлена на обеспечение устойчивого и сба
экономического развития Московской области н
с учетом положений документов стратегического плани

Министерство
экономического развития
Российской Федерации

Методические рекомендации по разработке и корректировке стратегии социально-экономического развития субъекта Российской Федерации и плана мероприятий по ее реализации

Тексты документов «инновационных» сценариев развития 8 субъектов РФ (до 2036-2040 гг.), граничащих с Москвой

Задача: провести сравнительный **контентный** анализ; оценить, в чем общность и различие стратегий развития регионов, окружающих город Москву

ЭКОНОМИКА МОСКВЫ

подключаем библиотеки, создаем объект TextDoc – «корпус», содержащий стратегию по Московской

```
# Install
#install.packages("tm")  # for text mining
#install.packages("SnowballC") # for text stemming
#install.packages("wordcloud") # word-cloud generator
#install.packages("RColorBrewer") # color palettes
#install.packages("syuzhet") # for sentiment analysis
#install.packages("ggplot2") # for plotting graphs
# Load
library("tm")
library("SnowballC")
library("wordcloud")
library("RColorBrewer")
library("syuzhet")
library("ggplot2")


# Read the text file 'MOS_1.txt'
text <- readLines('MOS_1.txt')
# Load the data as a corpus
```

Предварительная обработка текста, создание матрицы «строка-токен»

ЭКОНОМИКА

```
24 #Replacing "/", "@" and "|" with space
25 toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
26 TextDoc <- tm_map(TextDoc, toSpace, "/")
27 TextDoc <- tm_map(TextDoc, toSpace, "@")
28 TextDoc <- tm_map(TextDoc, toSpace, "\\|")
```

Убираем технические символыС

```
31 # Convert the text to lower case
32 TextDoc <- tm_map(TextDoc, content_transformer(tolower))
33 # Remove numbers
34 TextDoc <- tm_map(TextDoc, removeNumbers)
35 # Remove english common stopwords
36 TextDoc <- tm_map(TextDoc, removeWords, stopwords("english"))
37 # Remove your own stop word
38 # specify your custom stopwords as a character vector
39 TextDoc <- tm_map(TextDoc, removeWords, c("для","по", 'как'))
40 # Remove punctuations
41 TextDoc <- tm_map(TextDoc, removePunctuation)
42 # Eliminate extra white spaces
43 TextDoc <- tm_map(TextDoc, stripWhitespace)
```

Переводим в строчные буквы, исключаем русские и английские стоп-слова, знаки препинания, убираем излишние пропуски

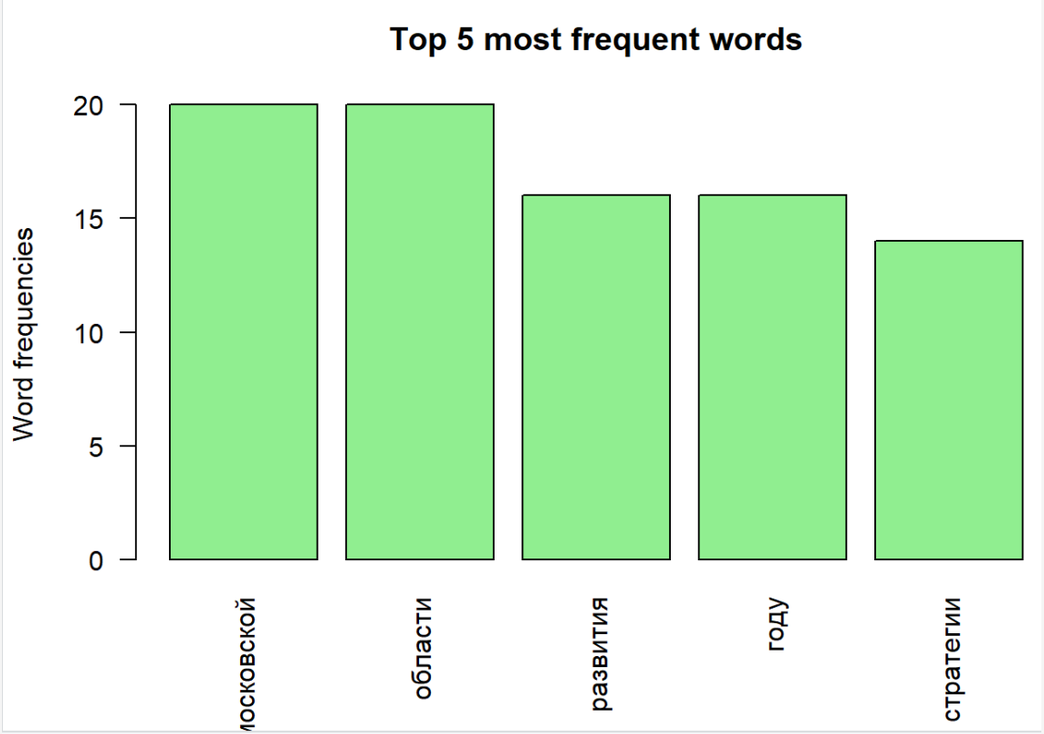
```
44 # Text stemming
45 TextDoc <- tm_map(TextDoc, stemDocument)
46
47 # Build a term-document matrix
48 TextDoc_dtm <- TermDocumentMatrix(TextDoc)
```

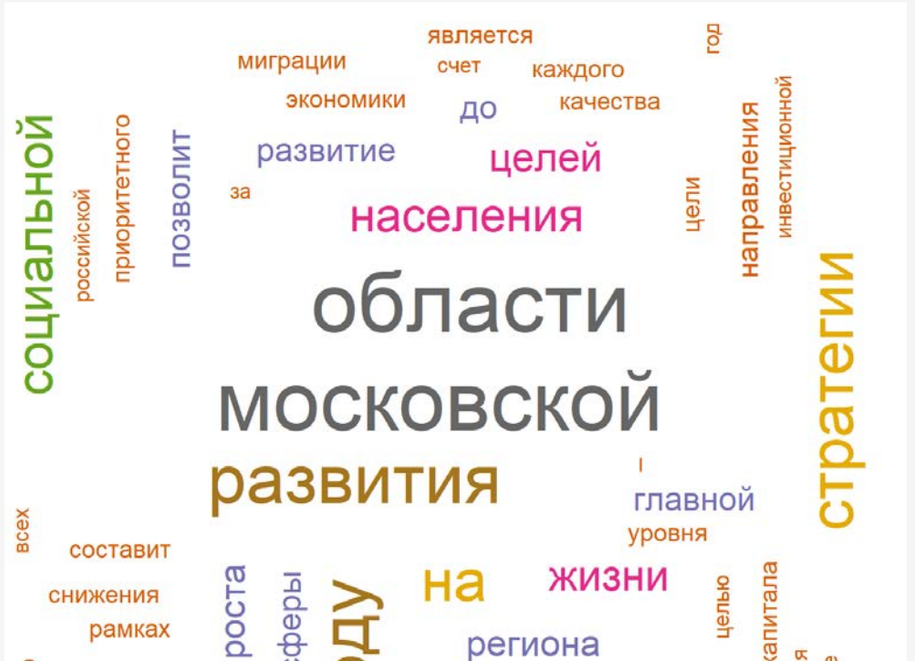
Проводим стемминг, создаем матрицу term-doc (строка-токен)

ЭКОНОМИКА МОСКВЫ

```
> head(dtm_d, 20)
```

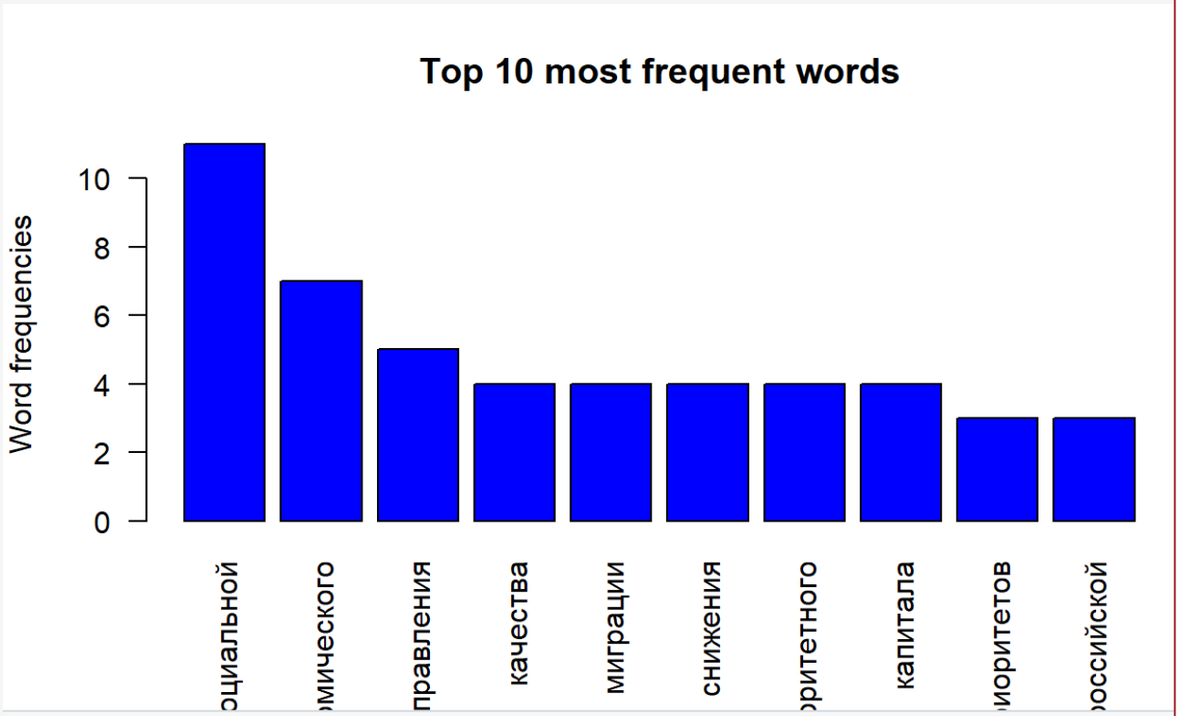
	word	freq
московской	московской	20
области	области	20
развития	развития	16
году	году	16
стратегии	стратегии	14
на	на	13
социальной	социальной	11
населения	населения	10
жизни	жизни	9
целей	целей	8
социальноэкономического	социальноэкономического	7
региона	региона	7
роста	роста	7
развитие	развитие	6
сферы	сферы	6
главной	главной	6
позволит	позволит	6



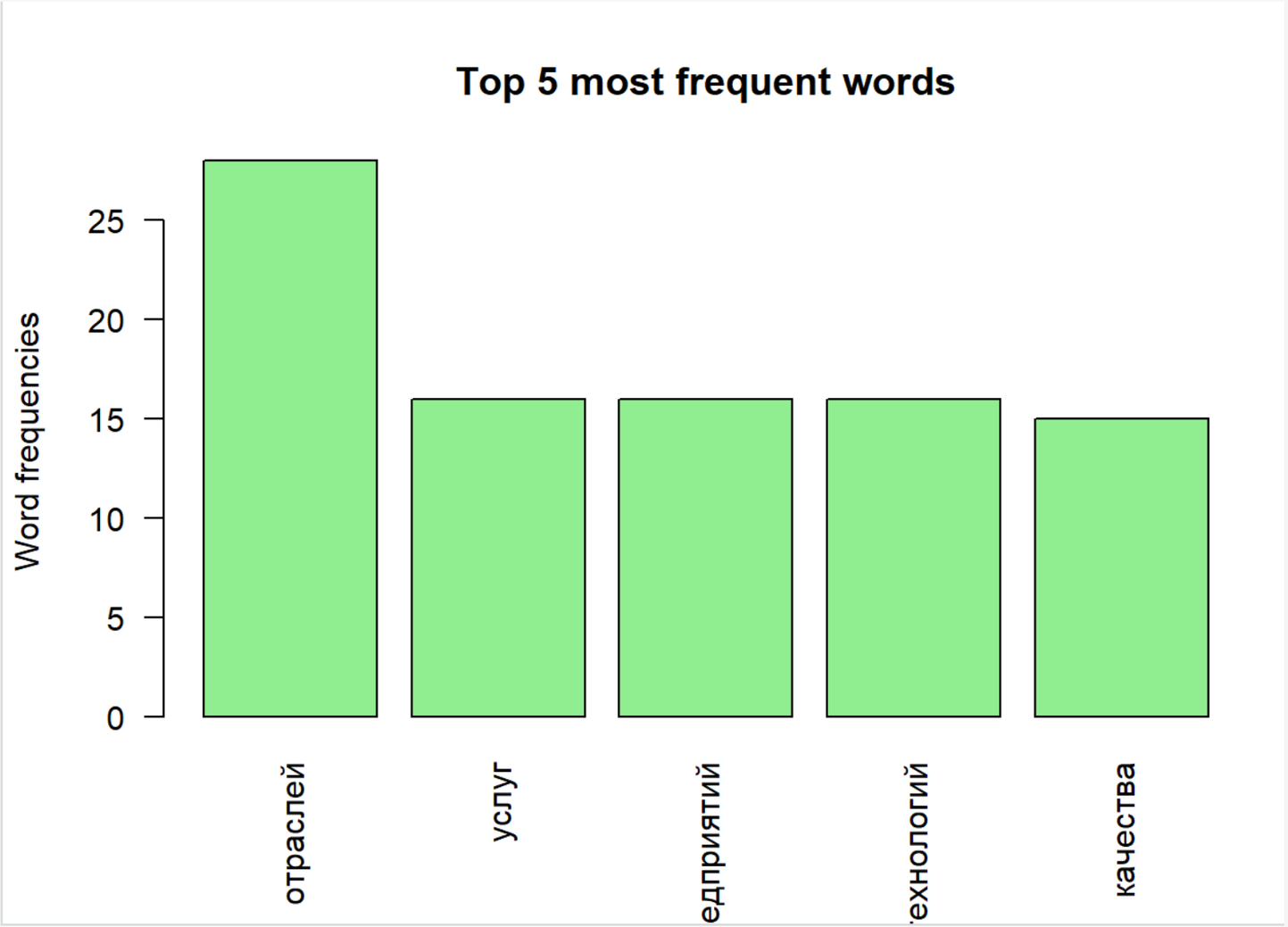


```r
### УБИРАЕМ РУССКИЕ СТОП_СЛОВА ПОСЛЕ ЧАСТОТНОГО АНАЛИЗА

TextDoc <- tm_map(TextDoc, removeWords, c("за","на","до","году","обеспечить","для","по",
                                          "московской","области","стратегии","развития","развитие","населения","как",
                                          "экономического","уровня","позволит","управления","экономики","роста",
                                          "сферы","главной","жизни","региона","на",
                                          "из","всех",'не','или','это','что','чтобы',"цели", "целей","достижение",
                                          "позволяет","каждого","рамках","составит","является"))
```



ЭКОНОМИКА МОСКВЫ



Отрасли

Услуги

Технологии

Качество

ЭКОНОМИКА МОСКВЫ

2_ОБЩИЕ ПРИОРИТЕТЫ СТРАТЕГИЙ ВСЕХ РЕГИОНОВ, ГРАНИЧАЩИХ С МОСКВОЙ

ЭКОНОМИКА МОСКВЫ

Ассоциации со словом «КЛАСТЕР»

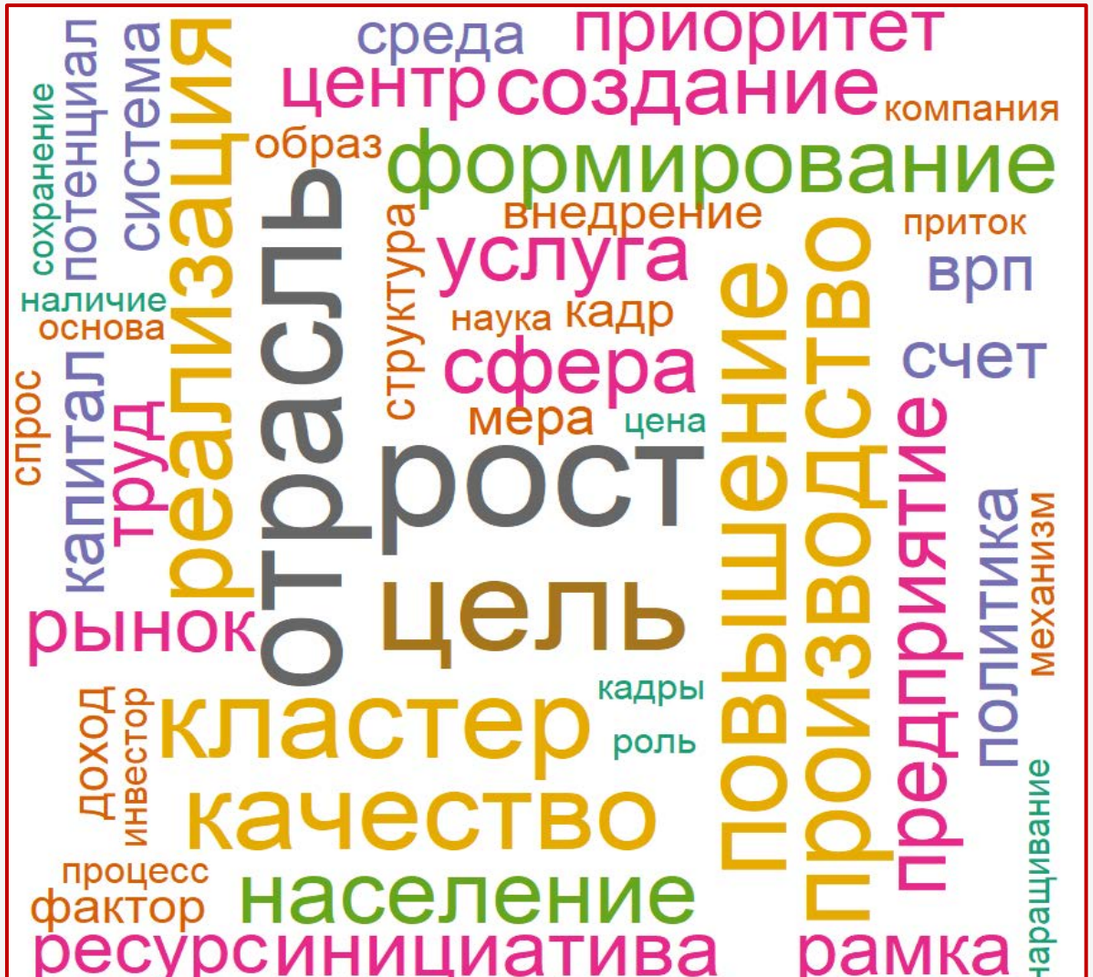
$кластер

туристскорекреационный	агропромышленный	газотурбостроения
0.37	0.37	0.26
энергетического	фармацевтической	автокомпонентов
0.26	0.26	0.26
ит	бизнесаутсорсинга	железнодорожный
0.26	0.26	0.26
транспортнологистический	биоэнергетический	фармацевтический
0.26	0.26	0.26

Особенности стратегических приоритетов: семантическая оценка

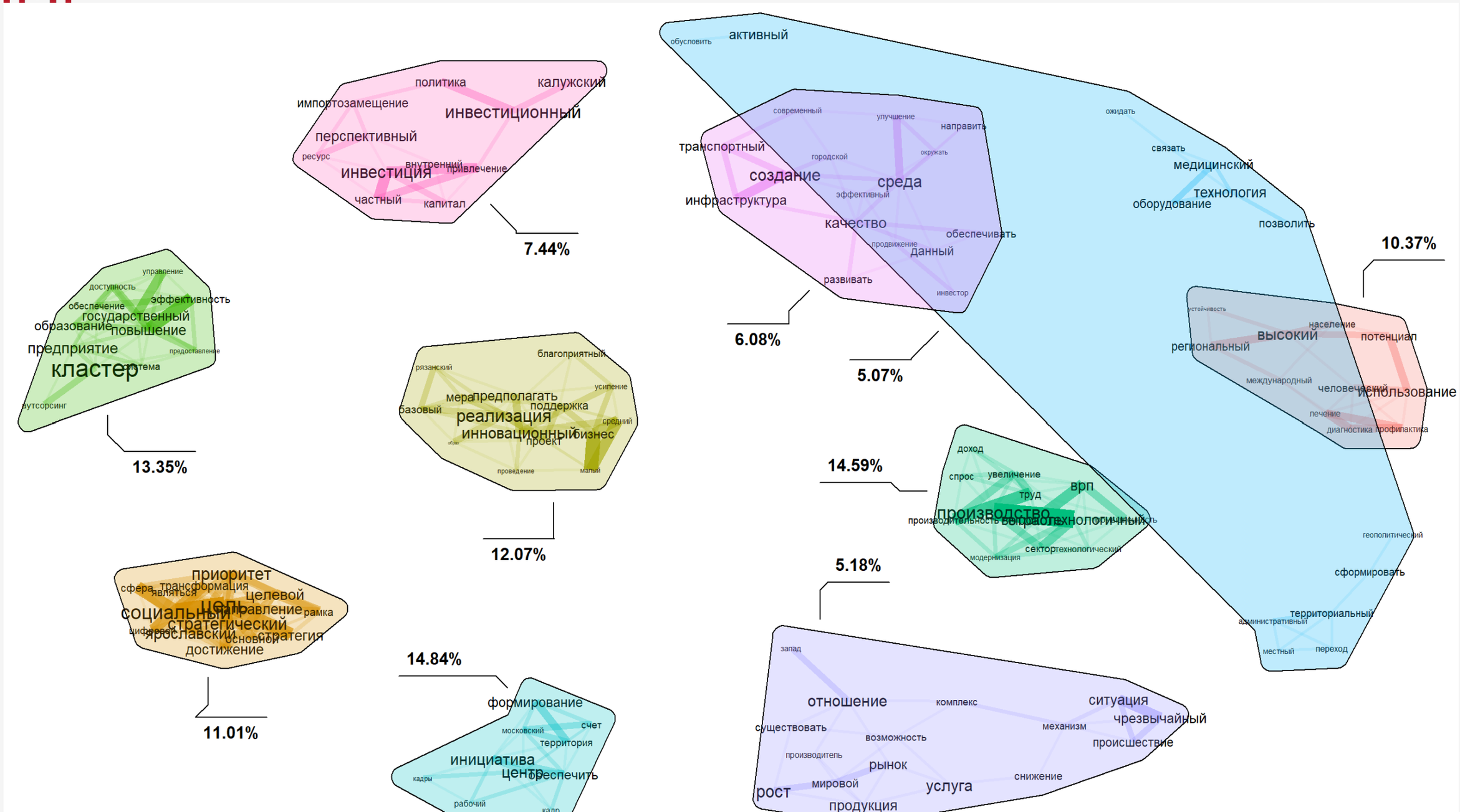
Московская область

Все регионы, граничащие с Москвой

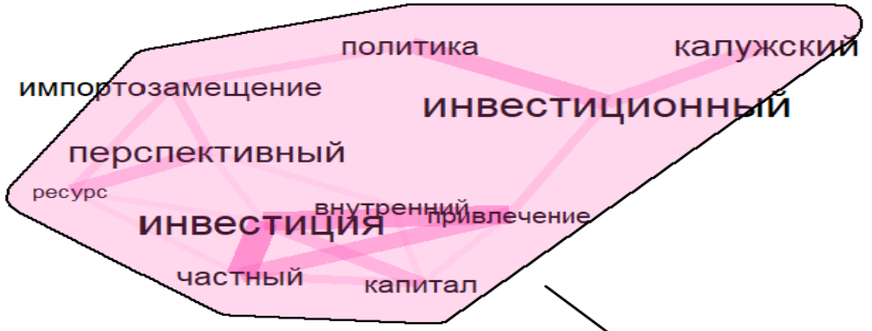


быстродействия: авторский словарь стоп-слов (фрагмент)

	A
1	**lemma**
2	август
3	автор
4	агент
5	акт
6	актив
7	акцент
8	апрель
9	апробация
10	архивный
11	аудитория
12	аый
13	балл
14	баррель
15	беспрецедентный
16	библиотека
17	блик
18	блок
19	будущее
20	будущий
21	быстрый
22	ввод
23	век
24	величина
25	вероятность
26	вероятный
27	весовой
28	внимание
29	возникновение
32	г.
33	газ
34	глава
35	год
36	голикова
37	голиковый
38	горячий
39	гост
40	грядущий
41	гчп
42	дело
43	деловой
44	демографический
45	денежный
46	день
47	детализация
48	детальный
49	деятельность
50	диалог
51	добавить
52	документ
53	документооборот
54	документооборота
55	доля
56	дополнение
57	дополнительный
58	доступ
59	доступа
61	е
62	едакция
63	единый
64	еисжс
65	епг
66	есиа
67	жк
68	задача
69	зал
70	запись
71	зачисление
72	ижс
73	изучение
74	имя
75	иной
76	ит
77	итог
78	ить
79	иэп
80	июль
81	июнь
82	ия
83	й
84	кабинет
85	карта
86	картина
87	категория
88	квазь
90	китать
91	клуб
92	ключевой
93	ко
94	конец
95	конечный
96	конкретизация
97	конкретный
98	контекст
99	контент
100	коэффициент
101	край
102	крайний
103	крый
104	лаборатория
105	лабораторный
106	легкий
107	линейка
108	линия
109	лицо
110	льший
111	май
112	март
113	меж

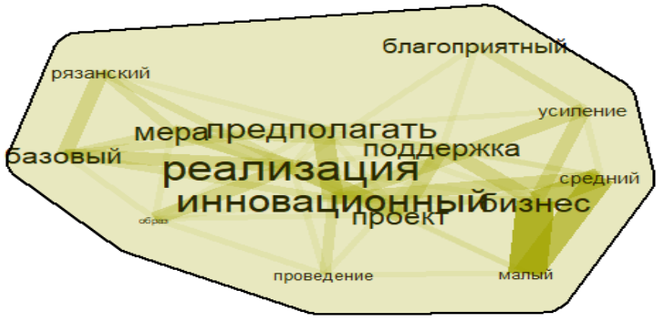

политика
калужский
импортозамещение
инвестиционный
перспективный
ресурс
внутренний
привлечение
инвестиция
частный
капитал
7.44%
управление
доступность
эффективность
обеспечение
государственный
образование
повышение
предприятие
предоставление
кластер
система
аутсорсинг
13.35%
благоприятный
рязанский
усиление
мера
предполагать
поддержка
базовый
реализация
средний
инновационный
бизнес
проект
проведение
малый
12.07%
приоритет
сфера
трансформация
являться
цель
целевой
социальный
управление
рамка
стратегический
ярославский
стратегия
основной
достижение
11.01%
14.84%
формирование
счет
московский
территория
инициатива
центр
обеспечить
кадры
рабочий
кадр
обусловить
активный
современный
улучшение
направить
ожидать
транспортный
городской
окружать
связать
медицинский
создание
среда
технология
инфраструктура
эффективный
оборудование
качество
позволить
обеспечивать
продвижение
данный
6.08%
развивать
инвестор
5.07%
10.37%
устойчивость
население
высокий
потенциал
региональный
использование
международный
человеческий
печение
диагностика
профилактика
14.59%
доход
спрос
увеличение
труд
врп
производство
производительность
технологичный
сектортехнологический
модернизация
геополитический
сформировать
территориальный
административный
местный
переход
5.18%
запад
отношение
комплекс
ситуация
чрезвычайный
существовать
механизм
возможность
происшествие
производитель
рынок
снижение
рост
мировой
услуга
продукция

ЭКОНОМИКА МОСКВЫ



7.44%

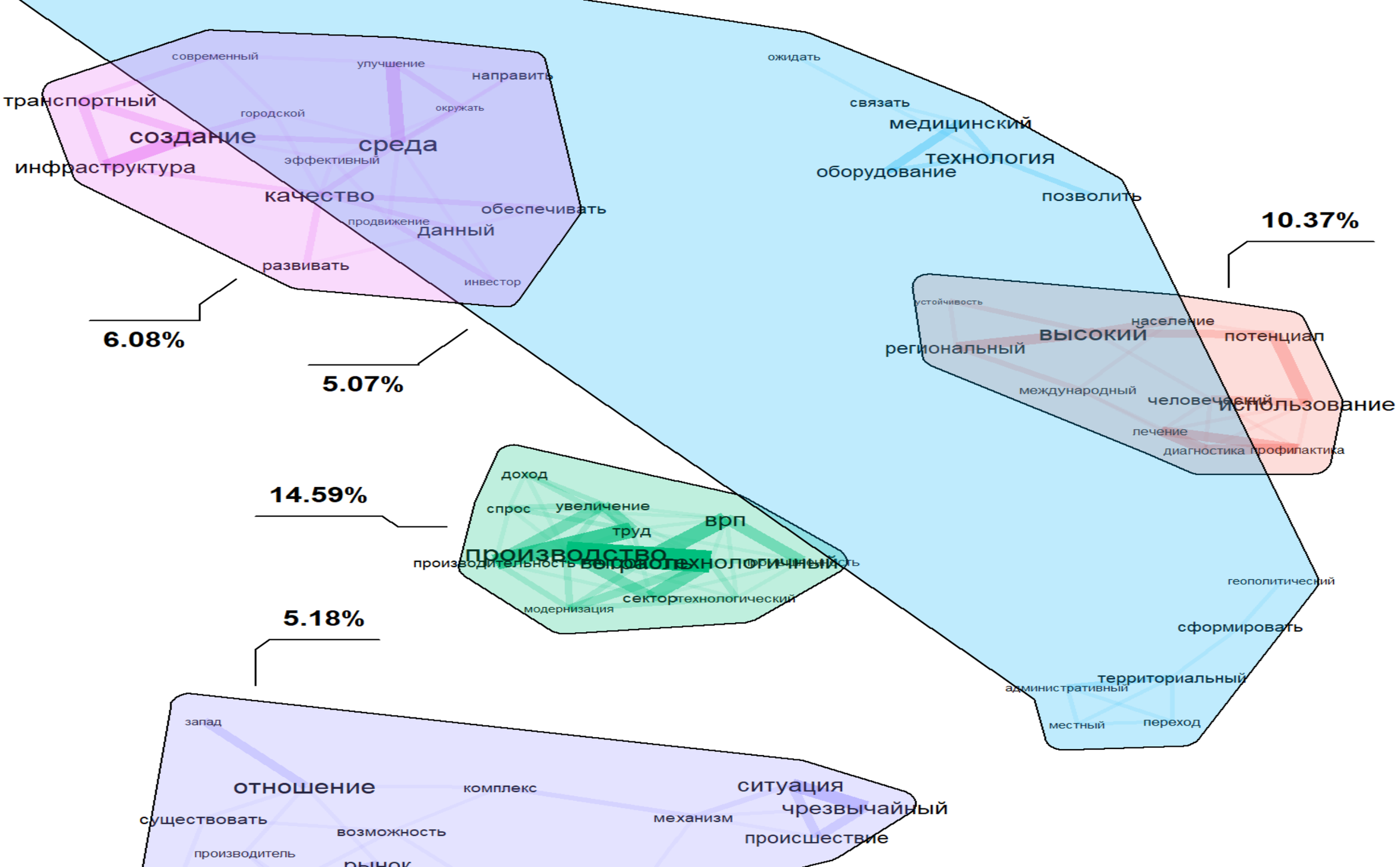
13.35%



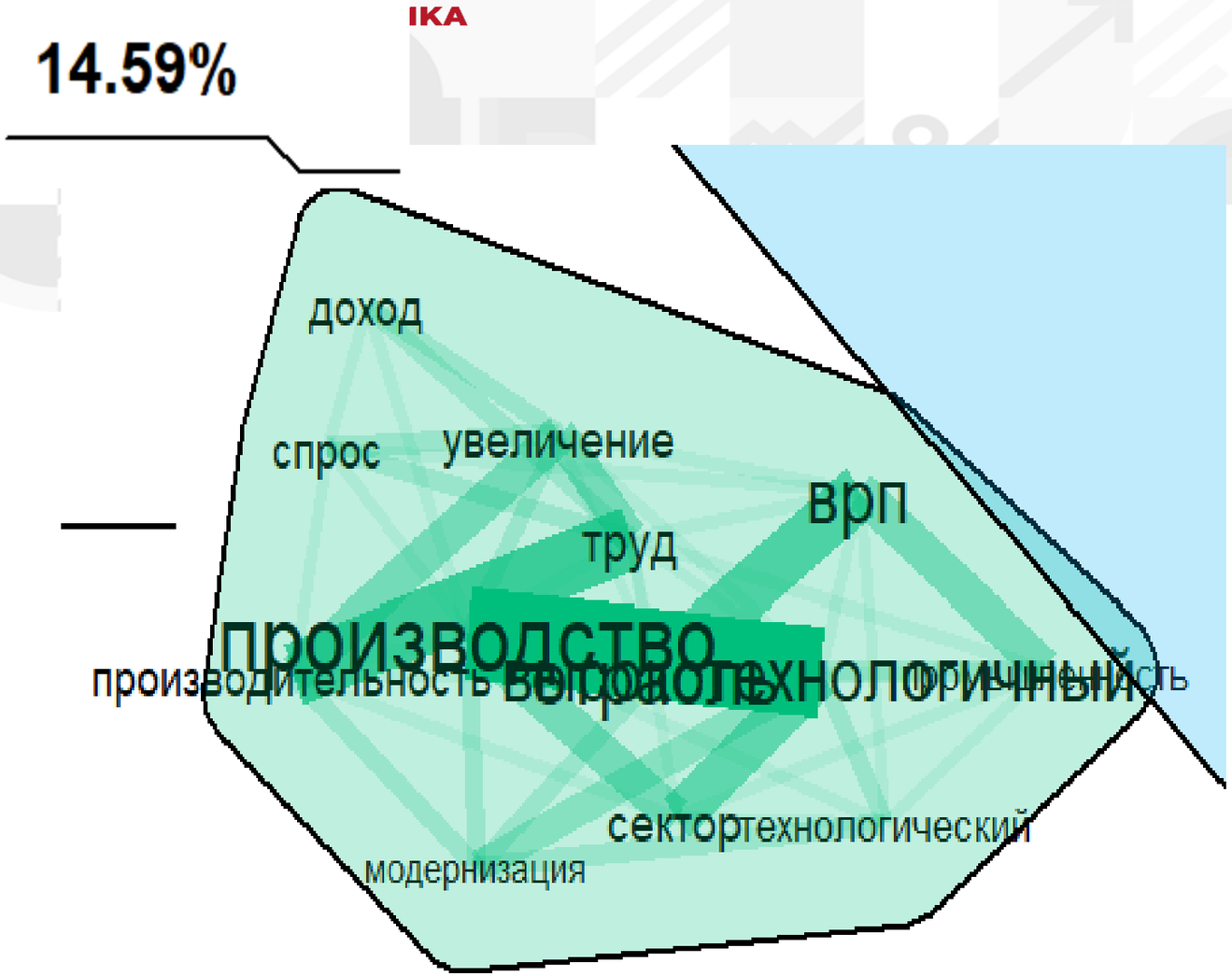
12.07%

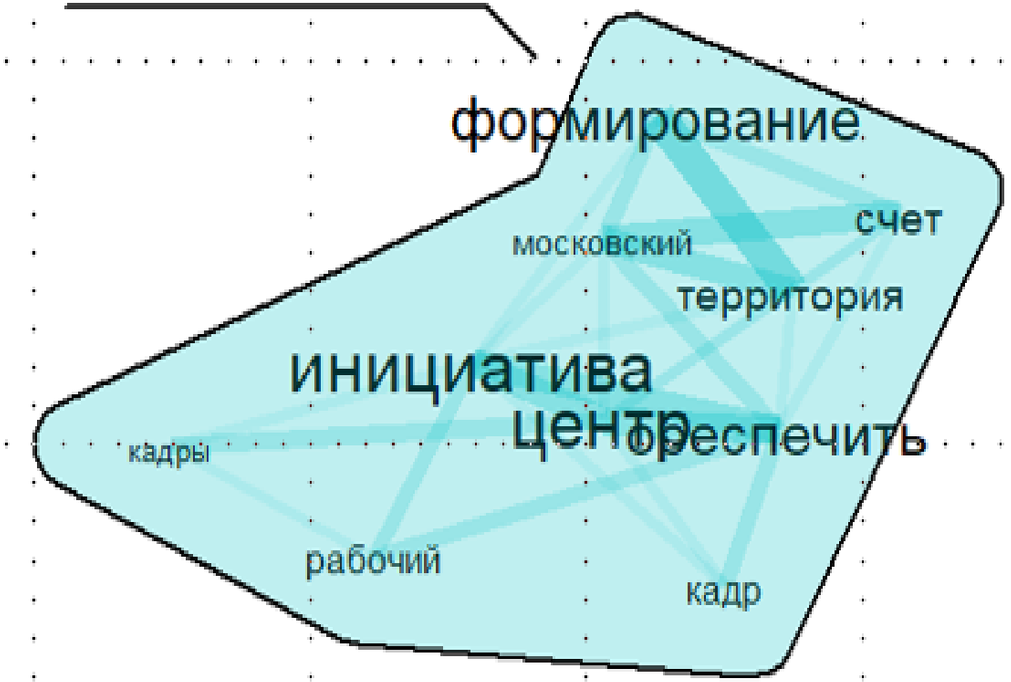
11.01%

14.84%

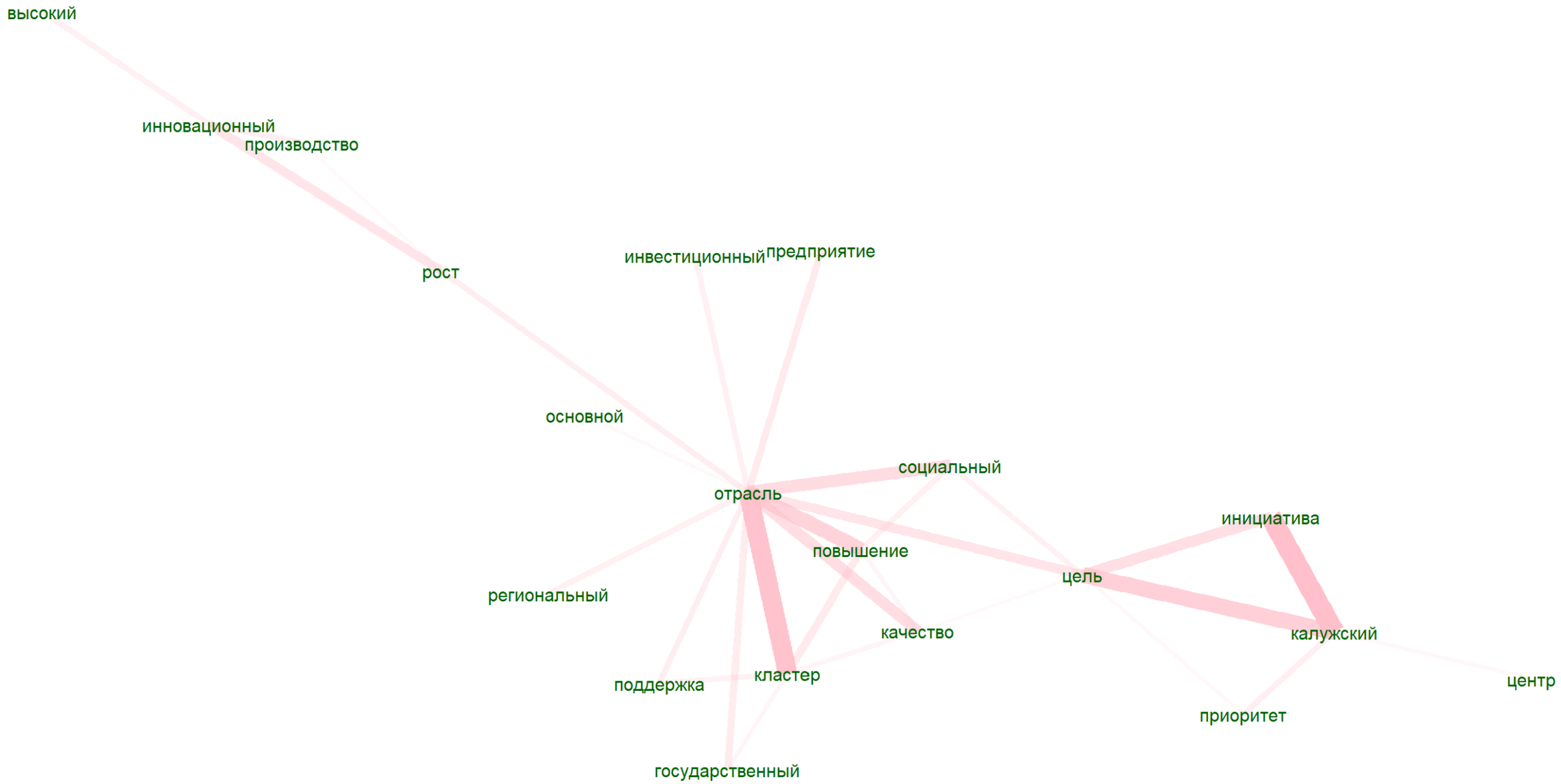

современный
улучшение
направить
транспортный
городской
окружать
создание
среда
инфраструктура
эффективный
качество
обеспечивать
продвижение
данный
развивать
инвестор
6.08%
5.07%
ожидать
связать
медицинский
технология
оборудование
позволить
10.37%
устойчивость
население
высокий
потенциал
региональный
международный
человеческий
использование
лечение
диагностика
профилактика
14.59%
доход
спрос
увеличение
врп
труд
производство
производительность
технологический
сектор
модернизация
геополитический
сформировать
территориальный
административный
местный
переход
5.18%
запад
отношение
комплекс
ситуация
существовать
механизм
чрезвычайный
возможность
происшествие
производитель

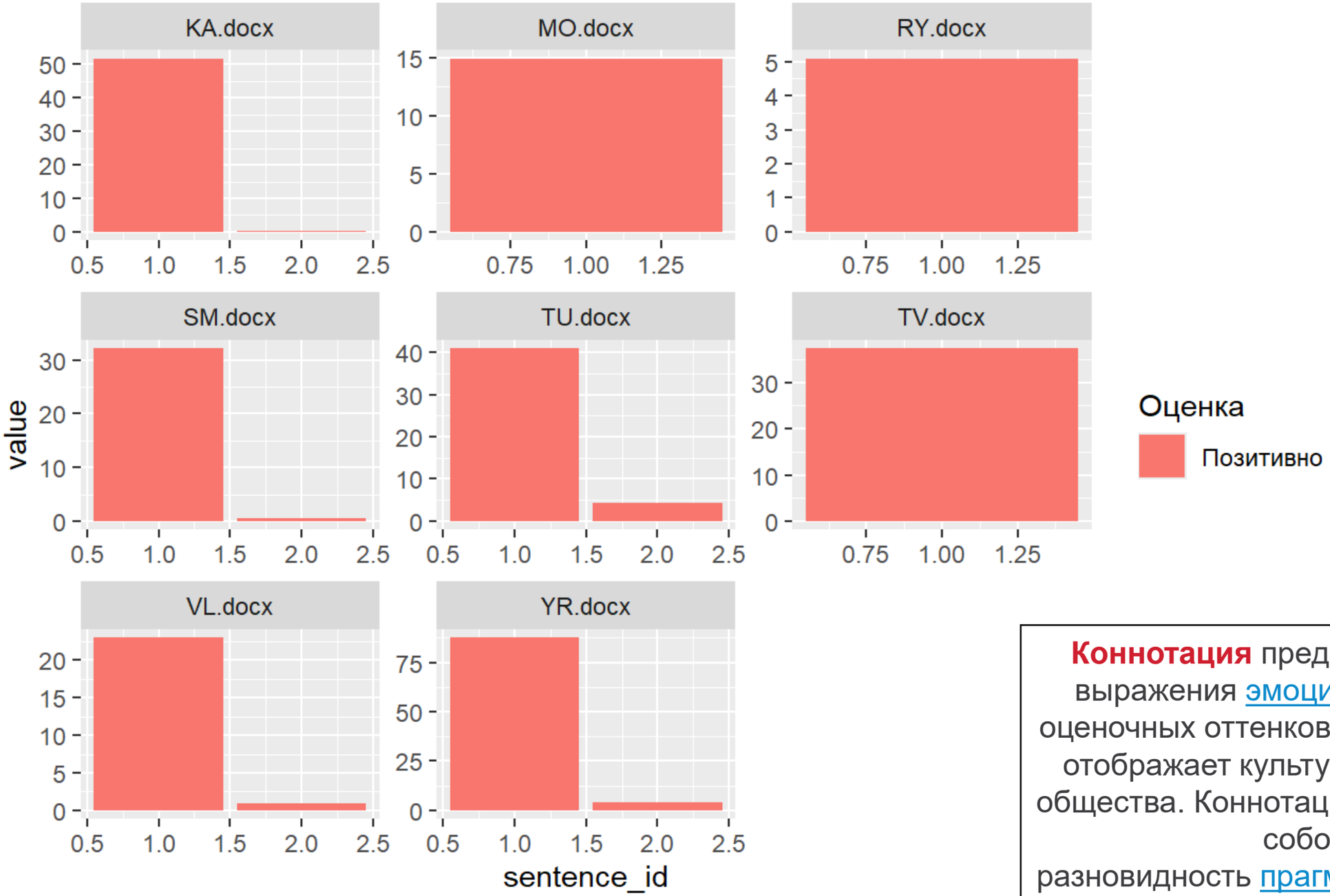
Наиболее значимые смысловые связи





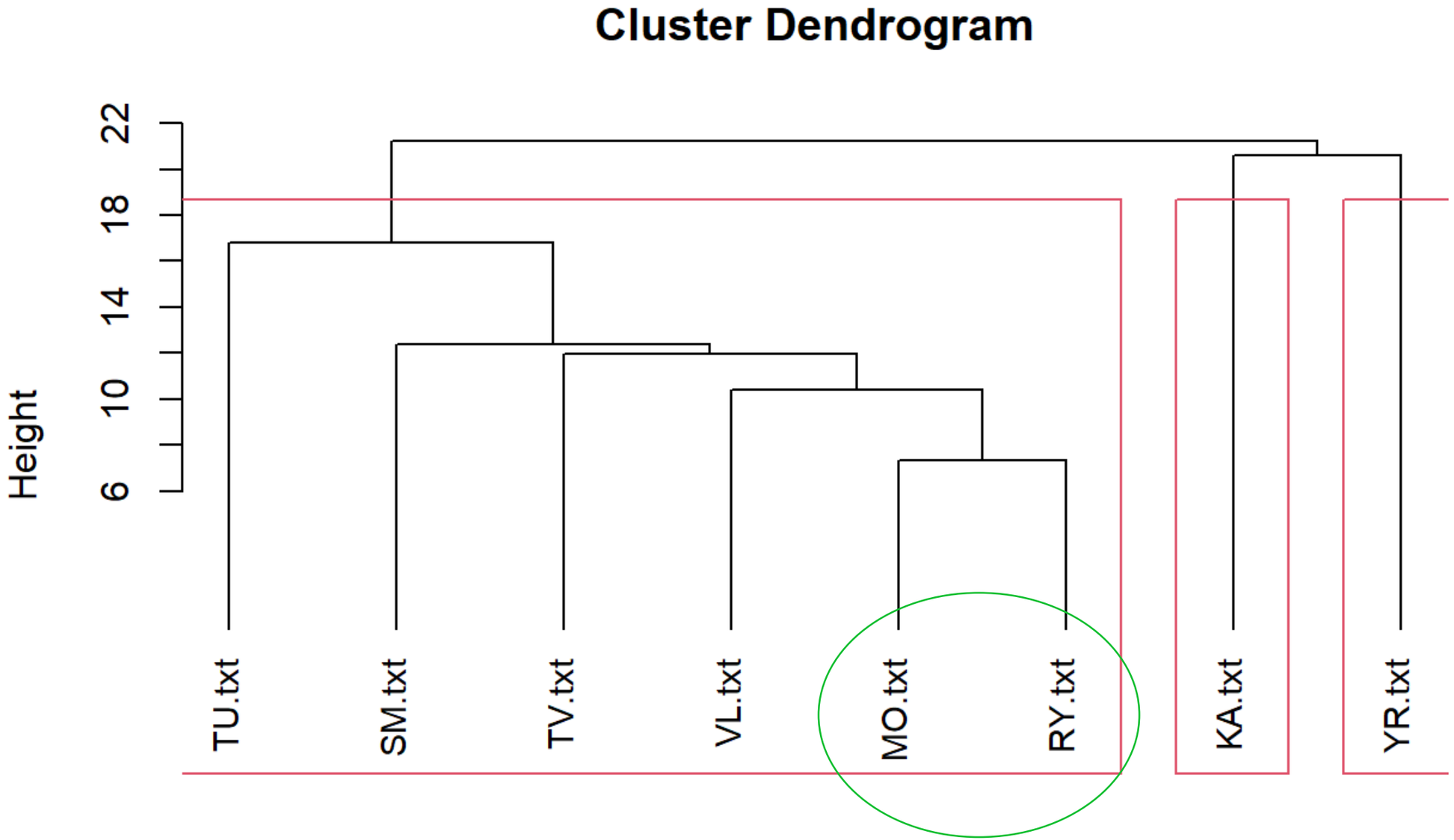
Акценты в реализации стратегических приоритетов





Коннотация предназначена для выражения эмоциональных или оценочных оттенков высказывания и отображает культурные традиции общества. Коннотации представляют собой разновидность прагматической инфо

Кластеры региональных стратегий



T_22

https://www.rbc.ru/economics/20/02/2023/63f3751b9a7947fbecbdf0c5

Спад экономики в 2022 году оказался меньше, чем в пандемию

Росстат: ВВП России в 2022 году снизился на 2,1% против 2,7% в пандемию

Российская экономика в 2022 году сократилась на 2,1%, показала первая оценка Росстата. Еще месяц назад президент Путин приводил оценку минус 2,5%. Прошлогодний спад ВВП оказался меньше, чем в пандемийном 2020-м

водоотведение, организация сбора и утилизации отходов, деятельность по ликвидации загрязнений» (минус 6,8%). Минприроды комментировало, что объем утилизации отходов и обработки вторичного сырья сократился в 2022 году на фоне санкций, которые привели к нарушению привычных цепочек поставок оборудования и материалов.

В торговом секторе весомая доля приходится на продажи природного газа. По

T_23

РИА НОВОСТИ

Главные события в экономике России в 2023 году

Главные события в экономике России в 2023 году: рекорд за рекордом

https://ria.ru/20231228/ekonomika-1918637515.html

При этом изначально власти видели свою задачу даже в том, чтобы хотя бы предотвратить обвал. "Была угроза обвала. И действительно пришлось мобилизовать все ресурсы, внутренние силы для того, чтобы этот обвал предотвратить", - говорил в ноябре пресс-секретарь президента Дмитрий Песков.

Главный вызов на следующий год – поддержать рост и не допустить перегрева экономики.

2. Рекордная безработица

Безработица в России в уходящем году оказалась рекордно низкой – в октябре она достигла 2,9% экономически активного населения. Это один из самых низких показателей в мире.

Но для бизнеса это создает проблемы, ведь потребности в рабочей силе у предприятий только растут. Для людей это означает повышение зарплат в тех сферах, где наблюдается наибольший дефицит кадров. И необходимость переобучения и повышения квалификации – в остальных.

В условиях низкой рождаемости России также придется привлекать мигрантов. Однако власти признают, что гастарбайтеры не решат проблему. Главный вызов – это повышение производительности труда.

3. Роковые яйца

Президент, подводя итоги года, сказал, что инфляция в России может быть выше 7,5%, ближе к 8%, но власти исходят из того, что ее удастся вернуть к целевому показателю в 4%.

Для достижения этой цели ЦБ проводит жесткую политику, подняв к концу года ключевую ставку до 16%. И планирует удерживать ее на этом уровне до середины следующего года.

сийская экономика
одновременно и по

T_24

РИА НОВОСТИ

Главные события в экономике России в 2024 году: рост вопреки

Российская экономика в уходящем году, несмотря на постоянно ужесточающиеся санкции, не просто показала рост, власти даже заявили о ее перегреве – то есть предложение не поспевает за спросом. Масштаб этого перегрева глава ЦБ Эльвира Набиуллина еще по итогам первого полугодия называла максимальным за последние 16 лет, с 2008 года. Главной проблемой стала инфляция, уже обошедшая официальные прогнозы и на середину декабря превысившая 9,5%.

Разгон цен вынудил ЦБ РФ ужесточать денежно-кредитную политику. В июле регулятор начал повышать ключевую ставку с тогдашних 16%, в октябре доведя ее до рекордных для России 21% годовых. При этом ЦБ сделал бизнесу и людям неожиданный новогодний подарок, сохранив ставку в декабре, хотя сам допускал ее повышение.

Но несмотря на все перипетии с ценами и ставкой, экономика по итогам года вырастет почти на 4%.

Почему это важно? Экономика РФ растет вопреки внешнему давлению, и это напрямую влияет на уровень жизни россиян. За суверенитет приходится платить высокую цену в виде инфляции и повышения ставок в экономике. Но в долгосрочной перспективе это принесет результат в виде более устойчивого роста и повышения благосостояния граждан.

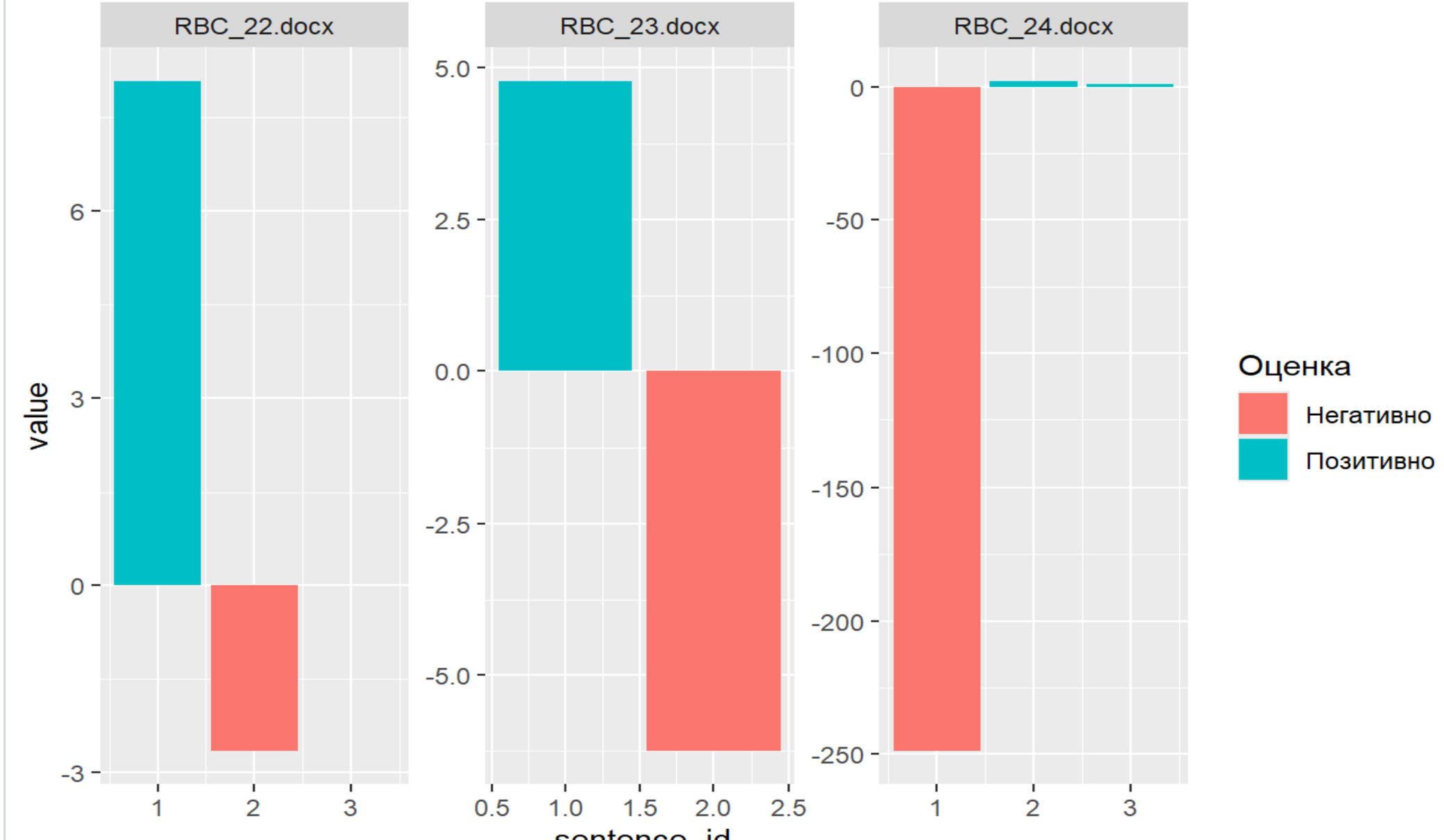
2. Изменения налоговой системы

Правительство РФ в уходящем году приняло масштабные налоговые изменения.

Проанализировать изменение аналитического фона СМИ по результатам развития страны за 3 года: 2022 – 2023

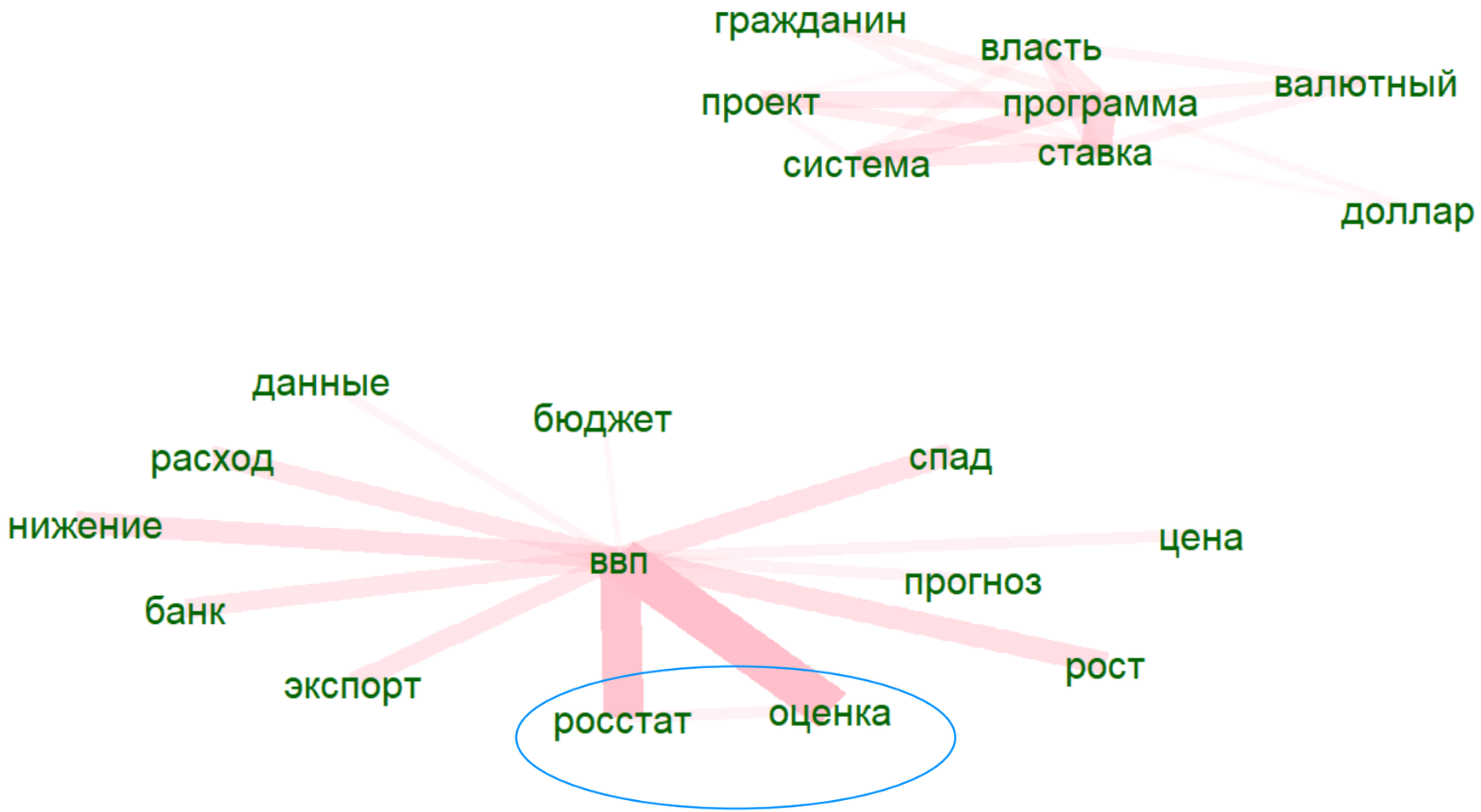
Изменение эмоционального фона аналитики

ЭКОНОМИКА МОСКВЫ



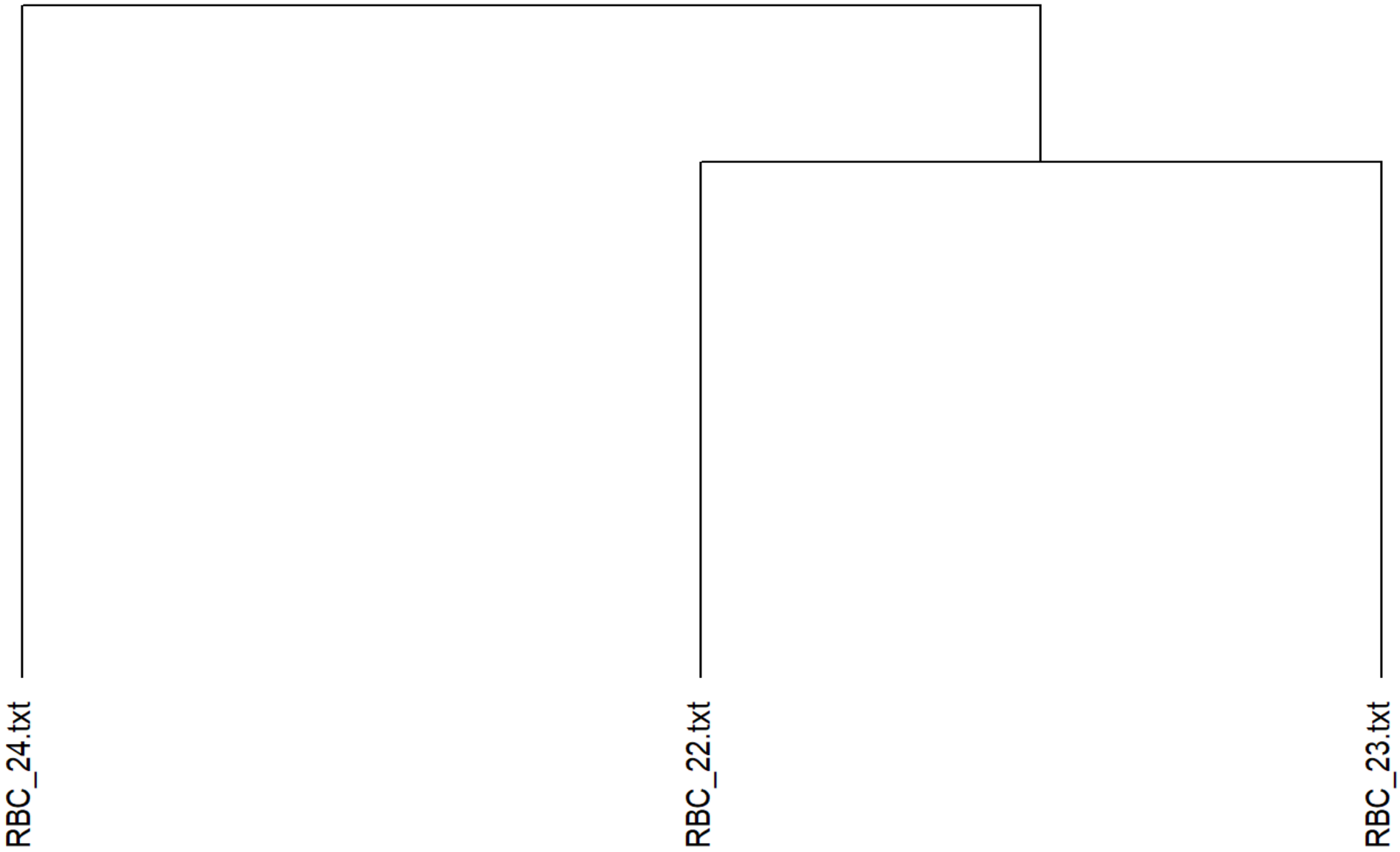
ЭКОНОМИКА МОСКВЫ

Оценка ассоциаций в годовых аналитических обзорах РБК за 2022-2024 гг.



годовой экономической ситуации (по аналитическим обзорам РБК)

ЭКОНОМИКА МОСКВЫ



ЭКОНОМИКА МОСКВЫ

Заключение

Семантический анализ – это перспективное направление, способное значительно повысить эффективность деятельности эксперта-аналитика в сфере государственного управления.

Развитие технологий машинного обучения и обработки естественного языка открывает новые возможности для автоматизации и оптимизации аналитических процессов.

В будущем, семантический анализ станет неотъемлемой частью работы экспертов-аналитиков, позволяя им принимать более взвешенные решения и эффективно управлять государственными ресурсами.

Благодарю за внимание!

ZarovaEV@develop.mos.ru