Выявление устойчивых трендов и редких тем в патентных данных на основе консенсуса трех принципиально разных методов и семантической унификации

Алексей Артамонов, Екатерина Бажанова,
Дарья Зрелова, Петр Зрелов, Владимир Кореньков

Лаборатория информационных технологий
им. М.Г. Мещерякова
Объединенный институт ядерных исследований

Дубна, 2026

Введение

Патентные исследования: виды и ограничения

Патентные исследования – «прикладные комплексные научные исследования… включающие поиск, анализ и систематизацию патентной и иной информации», направленные на выявление «тенденций развития продукции».

Виды патентных исследований:

1. На патентоспособность
2. На патентную чистоту
3. Целевые патентные исследования
4. Исследования на уровень техники и определение тенденций развития (ключевые для стратегического планирования)

Традиционный подход Преимущественно статистический анализ метаданных: количество патентов, география, заявители, IPC-классы. Семантический анализ текстов патентов не регламентирован.

Ключевые ограничения:

- Количественные методы не выявляют содержательную динамику технологий
- Низкая информативность метаданных о технологическом содержании
- Низкая прогностическая надежность выводов

Структура методики патентного анализа

Цель: получить объективную, воспроизводимую картину патентной активности. Востребован для:

- патентные отделы предприятий,
- юристы, менеджеры проектов,
- службы мониторинга (оперативная оценка)

- Динамика патентной активности
- Ключевые игроки
- Направления патентования
- Географическое распределение
- …

Статистический анализ

I уровень

Формулировка исследовательского запроса

Определение предметной области → Выбор БД и формирование запроса → Формирование и стандартизация корпуса

Выбор источника:

- Базы данных патентной информации
- Комбинирование нескольких источников

Формирование запроса: по МПК, ключевым словам, временному интервалу.

Получение первичной выборки → стандартизация данных

II уровень

Семантическое моделирование

- Выявление содержательных технологических тем
- Эволюция тем во времени (возникновение, слияние, затухание)
- Семантические связи и кросс-технологические пересечения
- …

Цель: выявить содержание технологий, их структуру, эволюцию и связи. Востребован для:

- Научных организаций и вузов
- Фондов развития, министерств и ведомств
- Стратегических подразделений корпораций
- (обоснование НИОКР, госполитики, прогнозов).

*В докладе подробно рассматривается Уровень II

Данные

Источник данных: Lens.org

Поисковый запрос: robotics AND (intelligent OR quantum)

Исходная выборка: 48 052 документа

Критерии отбора для семантического анализа:
Наличие непустого поля «abstract» (аннотация)
Язык аннотации — английский (латинская графика)

Формирование корпуса:
Отобрано: 45 170 патентов (94% выборки)
Исключено: 2 882 патента (6%)
2 595 (5,4%) — без аннотации
287 (0,6%) — на языках с нелатинской графикой

Вывод: Корпус репрезентативен и пригоден для семантического моделирования.

1 **Проблема: юридический шум и полисемия**

Патентные тексты содержат:

- **Юридические клише** и **технические шаблоны** маскирующие реальную техническую семантику.
- **Полисемию** — многозначность технических терминов

Результат: Отдельные методы тематического моделирования дают нестабильные или малоинтерпретируемые результаты.

Решение: Консенсусный подход трех независимых методов.

Предварительная обработка данных

Очистка текста и борьба с патентным шумом

1. Удаление разметки, специальных символов и тегов

2. Нормализация

Приведение к нижнему регистру, удаление пунктуации и цифр (кроме входящих в состав токенов).

3. Расширенный список стоп-слов

Стандартные списки стоп-слов недостаточны для работы с патентными данными.

Тексты патентов характеризуются наличием юридических клише (embodiment, discloses, herein, plurality, comprises a, according to, the invention) и технических шаблонов (method, system, device, apparatus), маскирующих реальную техническую семантику.

Поэтому был сформирован расширенный словарь из 752 слов, включающий стандартные английские стоп-слова, патентные юридические клише и технические шаблоны.

Предварительная обработка данных

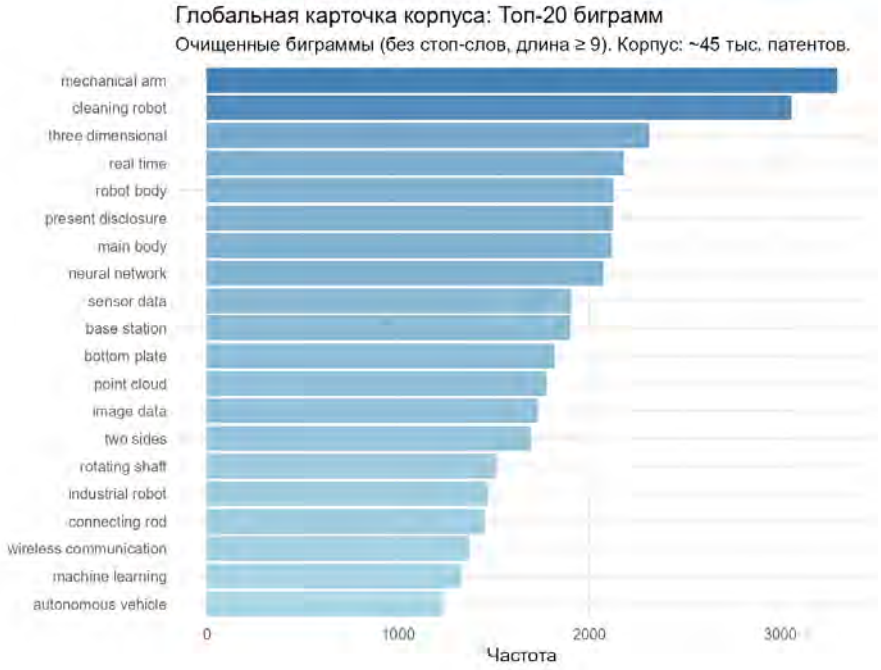
Извлечение и фильтрация биграмм

В технической документации смысл часто определяется устойчивыми словосочетаниями.

Например, униграммы **neural** и **network** по отдельности могут относиться к разным темам, тогда как биграмма **neural network** однозначно относится к искусственному интеллекту.
Использование биграмм повышает точность тематического моделирования. Именно биграммы, а не униграммы, позволяют модели различать полисемичные технические термины.

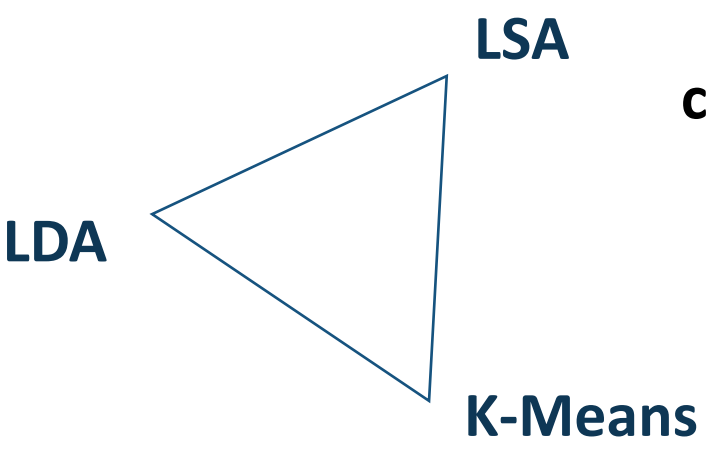
Для дальнейшего анализа использовались биграммы.

1. **Токенизация** (пример: neural network → neural_network)
2. **Фильтрация по длине**
   Убираем биграммы < **6** символов ИЛИ слова < **3** символов.
3. **Фильтрация по частоте**
   Остаются только статистически значимые биграммы $\mathbf{n \geq 10}$.
4. **Удаление стоп-слов внутри пар**
   Пары, содержащие хотя бы одно стоп-слово из расширенного списка, исключаются.



После фильтрации из общего количества **38 743** осталось **14 826** уникальных биграмм.

Методологический базис



K-Means на основе векторизации с использованием TF-IDF + усеченное сингулярное разложение (TruncatedSVD)

TF-IDF → DTM → TrancatedSVD → 50D vectors → K-Means → Clusters

$$X_{n\times m} \approx \boxed{U_{n\times 50} \times \Sigma_{50\times 50}} \times V^T_{50\times m}$$

Документы x Биграммы | Документы в латентном пространстве | Сингулярные значения | Биграммы в латентном пространстве

Суть метода. Геометрический подход. Кластеризация документов в векторном пространстве на основе минимизации внутрикластерного расстояния.

Входные данные. Эмбеддинги документов (**50D**), полученные с помощью TF-IDF + TruncatedSVD*.

Кластеризация производится в пространстве **50D** методом **K-Means.**

Роль метода в консенсусе: Учитывает **локальную плотность** и геометрию. Проверяет, образуют ли документы компактные группы в пространстве, независимые от вероятностных допущений.

*Рассматривались также альтернативные подходы с Word2Vec/Doc2Vec и GloVe.

Методологический базис

LDA (Latent Dirichlet Allocation)

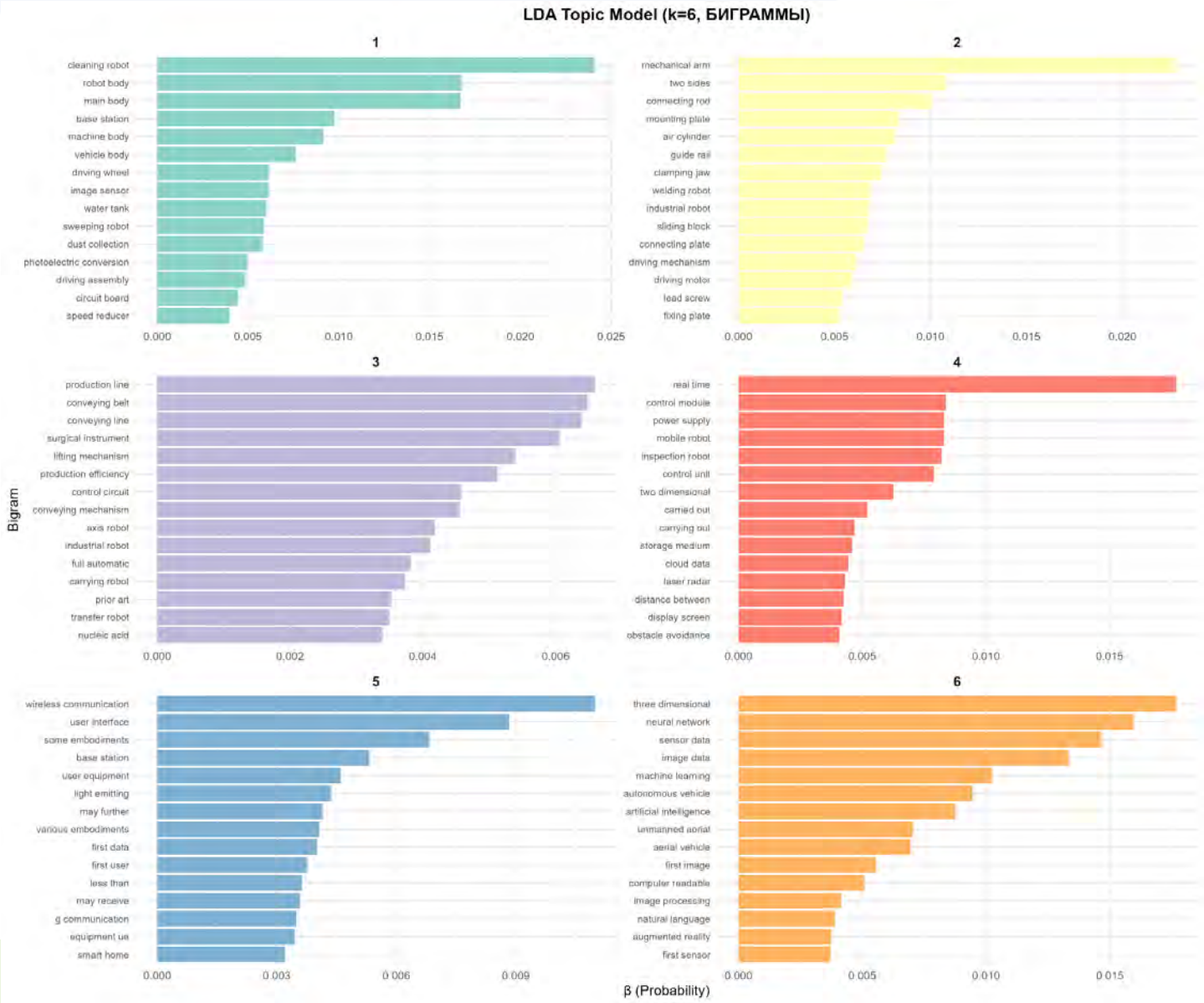
Суть метода: Вероятностная модель, представляющая каждый документ смесью скрытых тем, а каждую тему — распределением слов.
Выявляет **латентные семантические структуры** через статистические закономерности совместной встречаемости терминов.

Входные данные: Та же, что и в предыдущем методе, разреженная матрица «документ × биграмма» (DTM), построенная на очищенном словаре (**14 826** биграмм после удаления **752** стоп-слов и фильтрации).

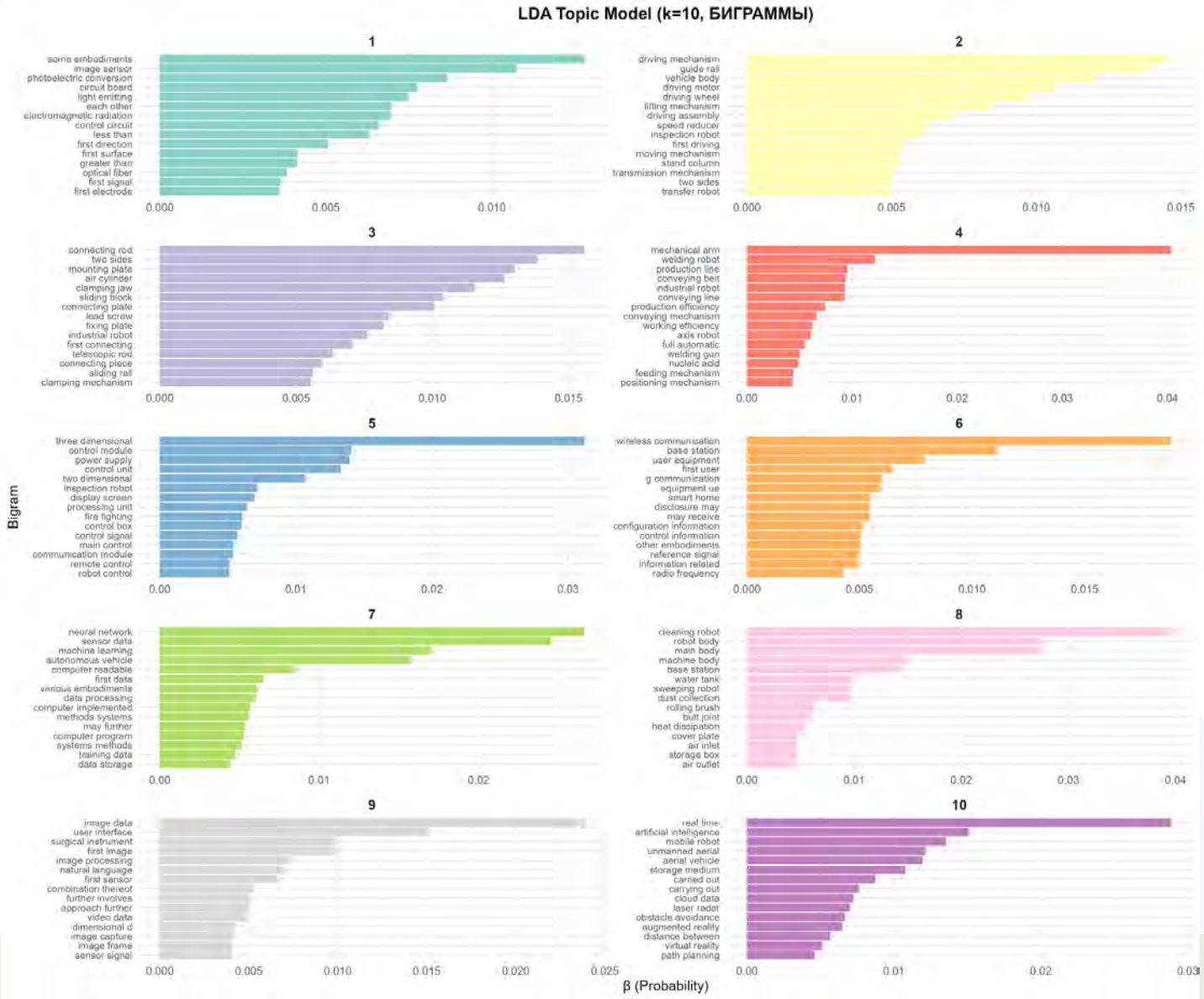
Параметры: **k** ∈ {4, 5, 6, 8, 10}. Выбор диапазона **k** обусловлен поиском баланса между гранулярностью и устойчивостью консенсуса.

Роль в треугольном консенсусе: Обеспечивает семантическую интерпретируемость. LDA лучше других методов объясняет содержательную часть патента через распределения слов, но чувствительна к шуму — поэтому требует подтверждения алгебраическим (LSA) и геометрическим (K-Means) методами..

LDA Topic Model K=6 Top 15 Bigrams



LDA Topic Model K=10 Top 15 Bigrams



Методологический базис

LSA (Latent Semantic Analysis / SVD)

TDM → SVD → Latent Semantic Factors

$$X_{m\times n} \approx \boxed{U_{m\times 50}} \times \Sigma_{50\times 50} \times V^T_{50\times n}$$

Биграммы x Документы | Биграммы в латентном пространстве | Сингулярные значения | Документы в латентном пространстве

Суть метода: Алгебраический подход, основанный на сингулярном разложении (SVD) матрицы «Терм × Документ» (TDM). Метод факторизует исходное пространство в набор ортогональных латентных факторов, упорядоченных по убыванию объясняемой дисперсии.

В отличие от LDA, LSA не делает вероятностных предположений и выявляет глобальные линейные зависимости между терминами на основе их совместной встречаемости во всём корпусе.

Входные данные: TF-IDF взвешенная матрица биграмм (**14 826 × 42 660**).

Обозначение **LSA_Bi** подчёркивает, что SVD применяется именно к **биграммной** матрице.

Параметры: **k = 50** (сохраняется ~70% общей дисперсии матрицы). Выбор размерности согласован с этапом векторизации для K-Means, обеспечивая единую семантическую базу для геометрической кластеризации.

Роль в треугольном консенсусе: Выступает стабилизирующим мостом между вероятностным (LDA) и геометрическим (K-Means) подходами. Устойчив к синонимии и вариативности патентных формулировок. Компенсирует локальную нестабильность LDA.

Треугольный консенсус: верифицированные результаты

Метрика согласованности

Для оценки устойчивости используется интегральный показатель:

$$\text{Avg_Consensus}(K) = \frac{C_{\text{KM}-\text{LDA}} + C_{\text{KM}-\text{LSA}} + C_{\text{LDA}-\text{LSA}}}{3} \quad (1)$$

Где K – количество кластеров/тем/топиков, C – консенсус для двух методов, вычисляемый как:

$$C_{A,B}(K_A, K_B) = \frac{\sum_{i=1}^{K_A} M_{ii}}{\sum_{i=1}^{K_A} \sum_{j=1}^{K_B} M_{ij}} \quad (2)$$

M_{ij} – число документов, отнесённых к теме i методом A и к кластеру j методом B

K_A – число тем/кластеров метода A, K_B – число тем/кластеров метода B

- **Core Topics**: Avg_Consensus ≥ 0.55. Порог установлен эмпирически.
- **Rare Topics**: Size ≤ 50 в LSA_{Bi}. Выявление микро-ниш, игнорируемых глобальными метриками.

Треугольный консенсус: верифицированные результаты

Таблица верифицированного консенсуса

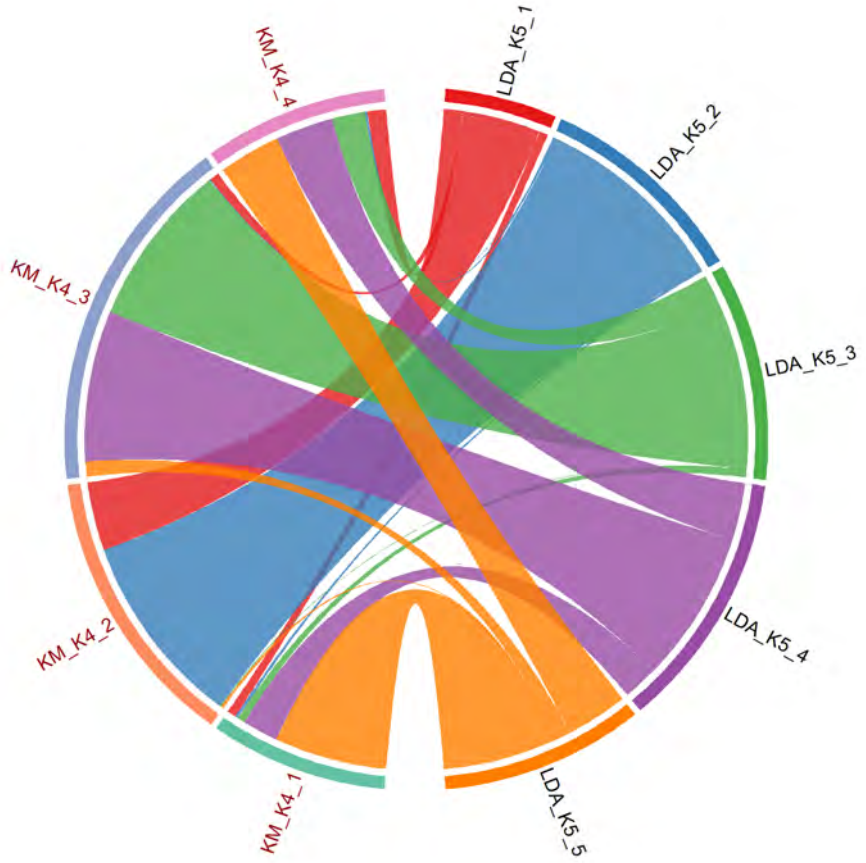
K	KM-LDA	KM-LSA	LDA-LSA	Avg_Consensus	Rank	Core	Примечание
4	0.5403	0.7076	0.7082	0.6520	4	Да	Основа макро-структуры
5	0.2696	0.9624	0.9457	0.7259	2	Да	
6	0.4615	0.7677	0.7638	0.6643	3	Да	
8	**0.4390**	**0.8807**	**0.8861**	**0.7353**	**1**	**Да**	Наиболее устойчивый уровень
10	0.4106	0.7128	0.6808	0.6014	5	Погран.	Фрагментация

Уровень K=8 является оптимальным для стратегического анализа (Avg = 0.7353).
LSA выступает стабилизирующим мостом: попарные метрики с участием LSA систематически превышают KM-LDA.

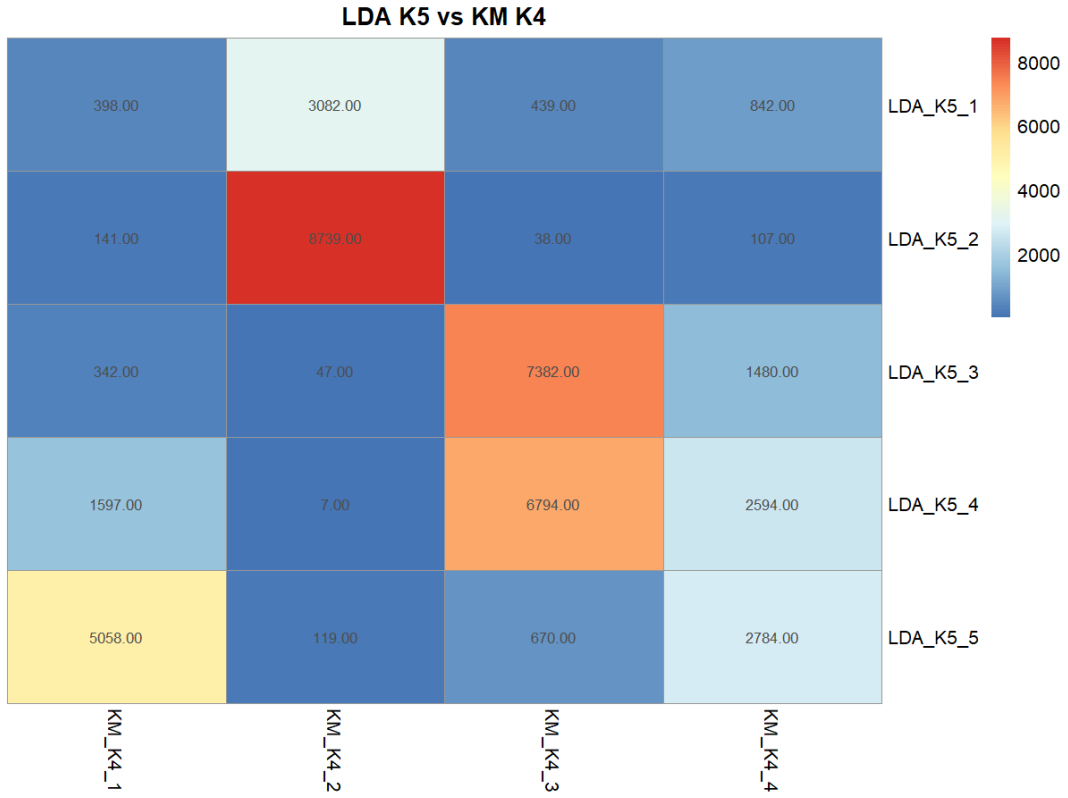
Кросс-валидация методов (Chord Diagrams)

Визуальное подтверждение связей между методами для выбранных уровней гранулярности

LDA (K=5) vs K-Means (K=4)



HITMAP

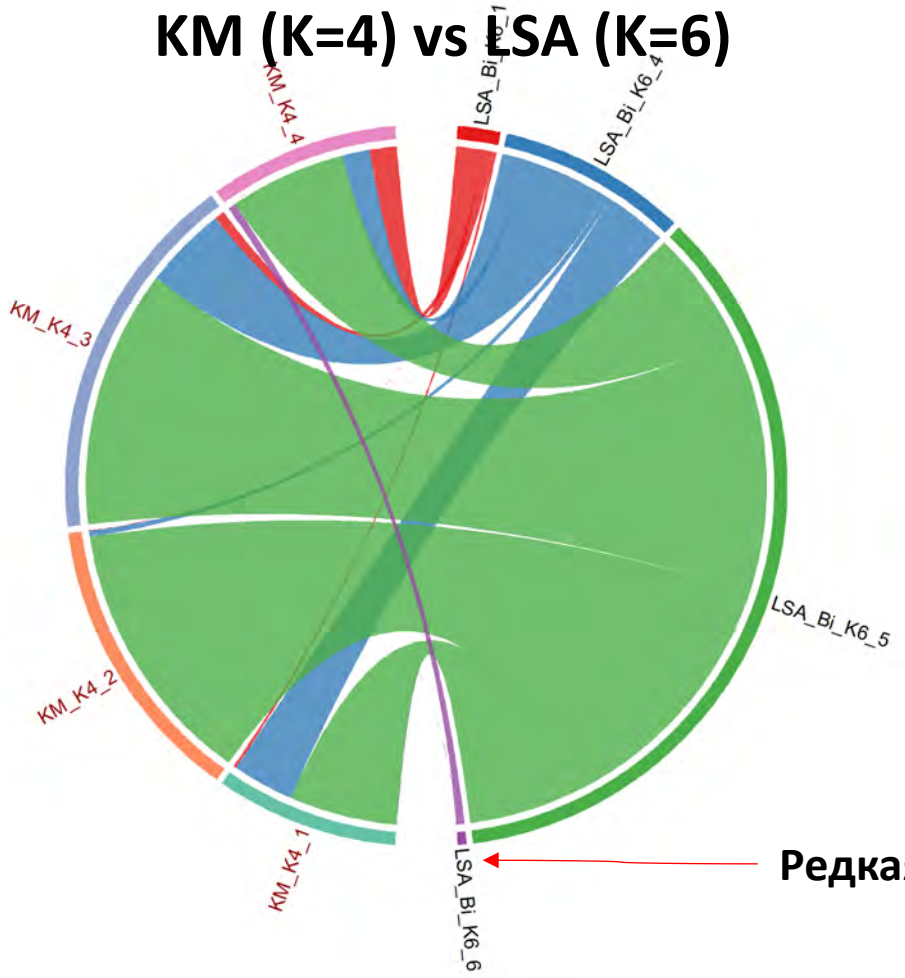


Связи между **вероятностными темами** (LDA) и **геометрическими кластерами** (K-Means). Толстые линии показывают сильное соответствие, тонкие — расщепление тем LDA на несколько кластеров K-Means (и наоборот).

Кросс-валидация методов (Chord Diagrams)

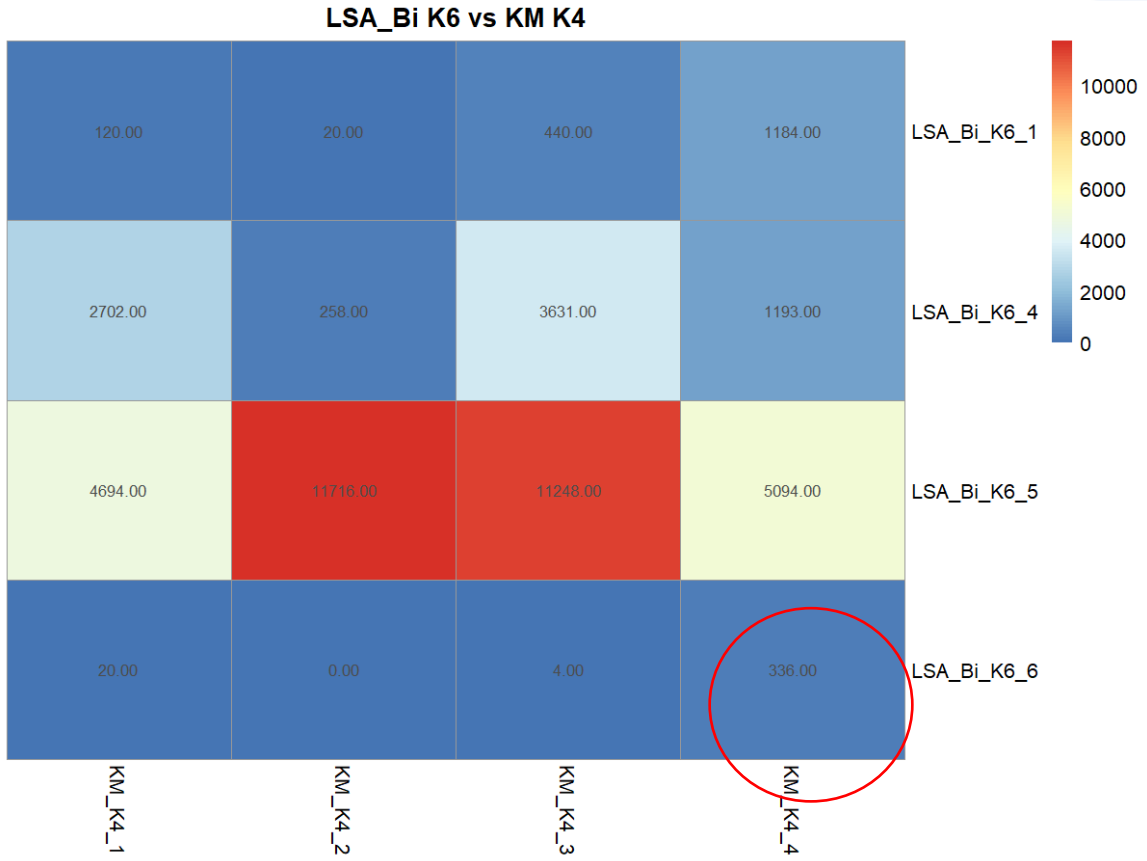
Хордовая кросс-валидация может служить дополнительным методом выявления достаточно редких тем

KM (K=4) vs LSA (K=6)



Редкая тема

HITMAP



Редкая нишевая тема LSA_K6_6, которая идентифицирована как «Биотехнологии и фармацевтические методы» по результатам анализа топ-биграмм . Тема включает термины pharmaceutical_compositions, genetic_variations, methods_treating .

Иерархическая структура и Супер-темы

«Треугольный консенсус» помог выявить, что разные уровни **K** оптимальны для разных частей семантического пространства. Корпус содержит патенты из разных предметных областей. Тематическая структура каждой из этих областей проявляется на разных уровнях гранулярности.

В отличие от традиционных подходов, фиксирующих единый уровень гранулярности (плоская структура), предлагаемая схема объединяет темы с трех уровней (K=4, K=6, K=8).

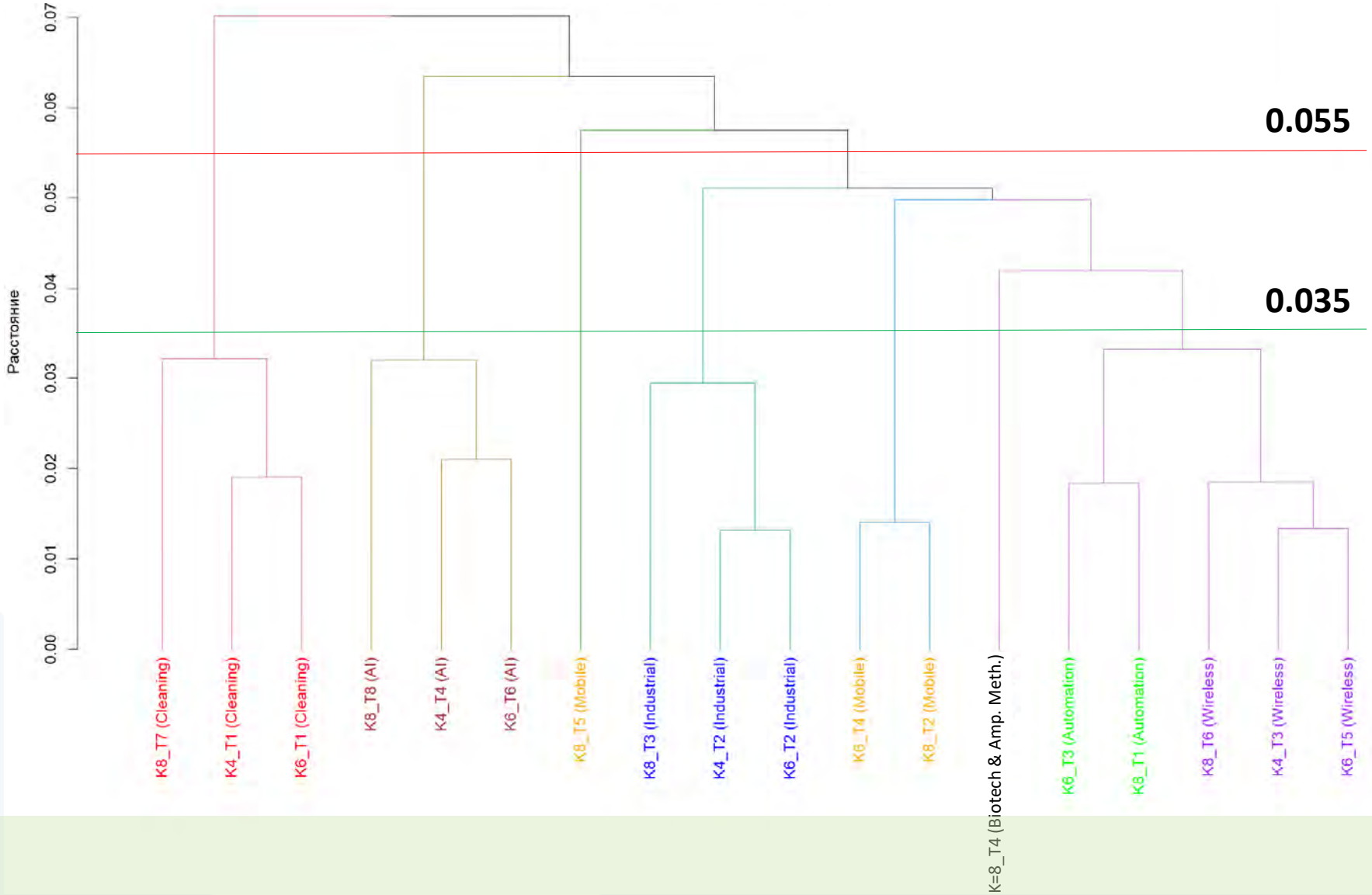
Уровень K=8 оптимален для детальных технических тем.
Уровень K=5/6 лучше выделяет нишевые инновации.
Уровень K=4 формирует устойчивые макро-темы.

Объединение этих уровней в единую дендрограмму позволяет сохранить естественную гранулярность каждой предметной области, избегая как излишнего укрупнения, так и фрагментации.

Смешанная дендрограмма отражает эту неоднородность структуры данных.

Иерархическая структура и Супер-темы

Дендрограмма Core-тем, объединяющих уровни К=4, К=6 и К=8



Метрика:
Cosine

Метод:
Complete Linkage

- Cleaning
- Industrial
- Automation
- Mobile
- Wireless
- AI
- Biotech & Amp. Methods

Иерархическая структура и Супер-темы

Экспертные названия 18 Core-тем (Mixed K=4/6/8) 1/2

Таблица **18** верифицированных **тем**, представленных на **дендрограмме** на предыдущем слайде.

Каждая тема имеет экспертное название, отражающее его техническую суть.
Названия тем сформированы на основе анализа топ-15 биграмм и репрезентативных аннотаций с последующей валидацией на соответствие поисковому запросу.

ID Темы	Уровень	Индекс темы	Экспертное название	Русский аналог
1	K=8	K8_T7	Biotech & Amplification Methods	Биотехнологии и методы амплификации
2	K=4	K4_T1	Mechanical Assembly System	Системы механической сборки
3	K=6	K6_T1	Cleaning Robot Mechanisms	Механизмы роботов-уборщиков
4	K=8	K8_T8	Servo Drive Control	Управление сервоприводами
5	K=4	K4_T4	Intelligent Data Processing	Интеллектуальная обработка данных
6	K=8	K8_T6	Wireless Communication Modules	Модули беспроводной связи
7	K=8	K8_T5	Optical Display Components	Компоненты оптических дисплеев
8	K=8	K8_T3	Fixed Connection Structures	Структуры фиксированных соединений
9	K=4	K4_T2	Electronics & Autonomy	Электроника и автономность

Иерархическая структура и Супер-темы

Экспертные названия 18 Core-тем (Mixed K=4/6/8) 2/2

Таблица **18** верифицированных **тем**, представленных на дендрограмме (**продолжение**).

Названия сформированы на основе анализа топ-15 биграмм и репрезентативных аннотаций с последующей валидацией на соответствие поисковому запросу.

ID Темы	Уровень	Индекс темы	Экспертное название	Русский аналог
10	K=6	k6_T2	Autonomous Vehicle Navigation	Навигация автономных транспортных средств
11	K=6	k6_T4	Sensor Data Fusion	Объединение данных датчиков
12	K=8	k8_T2	System Architecture Patterns	Паттерны системной архитектуры
13	K=8	k8_T4	Feeding & Supply Mechanisms	Механизмы подачи и снабжения
14	K=8	k8_T6	Image Processing Algorithms	Алгоритмы обработки изображений
15	K=4	k4_T3	System Architecture & Computing	Системная архитектура и вычисления
16	K=6	k6_T5	Neural Network Applications	Применения нейронных сетей
17	K=6	k6_T3	Electronic Control Units	Электронные блоки управления
18	K=8	k8_T1	Patent Disclosure Methods	Способы раскрытия информации о патентах

Иерархическая структура и Супер-темы

Выделенные Супер-темы (1)

На основе разреза дендрограммы на уровне h=**0.035** выделено **7** супер-тем

№	Экспертное название	Состав супер-тем (смешанные уровни)	Комментарий
ST-1	Core Systems & Control Technologies	K4-T1, K8-T1, K8-T7	макро-домен систем + управление
ST-2	Advanced Sensing & Materials	K4-T4, K6-T6, K8-T8	материалы и сенсоры
ST-3	Electromagnetic Radiation & Optical Sensing	K8-T5	изолированная ниша оптических сенсоров
ST-4	Industrial Processing & Mechanical Methods	K4-T2, K6-T2, K8-T3	промышленные/механические методы
ST-5	Data Analysis & Computational Methods	k6_T4, k8_T2	алгоритмы и анализ данных
ST-6	Biotech & Amplification Methods	K8-T4	Редкая ниша стала самостоятельным супер-кластером
ST-7	Pharmaceuticals, Bioinformatics & Integrated Biotech	k4_T3, k6_T3, k6_T5, k8_T1, k8_T6	Объединяет фармацевтику, биоинформатику и смежные биотех-темы

Иерархическая структура и Супер-темы

Выделенные Супер-темы (2)

На основе разреза дендрограммы на уровне h=**0.055** выделено 4 супер-темы

№	Экспертное название	Состав супер-тем (смешанные уровни)	Обоснование включения разных K
ST-1	Core Systems & Control Technologies	k4_T1, k8_T1, k8_T7	Макро-домен систем + детализации по управлению и специфическим системам
ST-2	Advanced Sensing & Materials	k4_T4, k6_T6, k8_T8	Макро-домен материалов/сенсоров + ниши K=6/K=8
ST-3	Electromagnetic Radiation & Optical Sensing	k8_T5	Изолированная тема; оптические сенсоры, излучатели, пиксельные матрицы
ST-4	Biotech, Pharma & Data Processing Methods	k4_T2, k4_T3, k6_T2, k6_T3, k6_T4, k6_T5, k8_T2, k8_T3, k8_T4, k8_T6	Крупнейший кластер; биотехнологии, фармацевтика, биоинформатика, методы обработки

Заключение

Применение «треугольного» консенсуса трех принципиально различных методов позволило:

1. Выявить устойчивую тематическую структуру, не зависящую от артефактов отдельного алгоритма.
2. Определить оптимальную гранулярность анализа (K=8) на основе верифицированных метрик.
3. Обеспечить методологическую целостность и воспроизводимость результатов анализа патентных ландшафтов.
4. Выявлять как тренды, так и микро-ниши.
5. Сформировать рекомендации:
   - Использовать **K=8** как оптимальную гранулярность для стратегического анализа.
   - Применять **супер-темы** для структурирования доменов и устранения дублирования.
   - Исследовать **ниши редких тем** на предмет уникальных технических решений.

Новизна подхода – на уровне объединения различных методов, а использование смешанной дендрограммы для тем с разных уровней грануляции также обладает методологической новизной.

Thank you for your attention!

Contact: *zrelova@jinr.ru*

Иерархическая структура и Супер-темы

Подход	Описание	Ограничение
Плоская структура (например,только K=8)	Все темы одного уровня детализации	**Теряется контекст**: непонятно, какие темы являются частными случаями других
Смешанная структура (Например, K=4+6+8)	Объединение наиболее устойчивых и содержательных тем с разных уровней	**Сохраняет естественную гранулярность**: некоторые инновации видны только на K=8, а общие тренды — только на K=4

Редкие инновационные ниши

K	Topic	Size	Название	Описание
5	2	2	Роботизированные механизмы подачи и зажима	Автоматическая подача и фиксация деталей.
5	4	34	Методы амплификации нуклеиновых кислот	Изотермические методы, полимеразы
8	4	30	Биотехнологии и амплификация	Амплификация, хирургические инструменты

- **Биотехнологии** — устойчивая микро-ниша, воспроизводящаяся на k=5 и k=8
- **Механизмы подачи** — узкий аспект, интегрированный в Супер-тему автоматизации.