Группа №10

М.А. Зайнуллин[1], А.А. Ратахина[2], М.В. Максимов[3], П.В. Терещенко[4]

[1] – студент Санкт-Петербургского государственного университета

[2] – студентка Тульского государственного университета

[3] – студент университета ИТМО

[4] – студентка Южно-Уральского государственного университета (НИУ)

ФОРМИРОВАНИЕ ПАТЕНТНОГО ЛАНДШАФТА ОБЪЕДИНЁННОГО ИНСТИТУТА ЯДЕРНЫХ ИССЛЕДОВАНИЙ (ОИЯИ)

ДУБНА 2025

Содержание

1 Аннотация

Разработана автоматизированная система анализа патентного ландшафта ОИЯИ на основе методов машинного обучения. Реализован шестиэтапный пайплайн: сбор данных из патентных баз и репозиториев → LLM-извлечение тегов → генерация эмбедингов → иерархическая кластеризация → 3D визуализация. Собран единый датасет из 1948 документов (патенты, научные публикации, программное обеспечение, базы данных), построена иерархическая структура научных направлений с 79 уровнями детализации, создан интерактивный веб-визуализатор (http://144.124.229.26:8050). Выявлены ключевые направления исследований и области с высоким патентным потенциалом.

Ключевые слова: патентный ландшафт, кластеризация документов, эмбединги, машинное обучение, LLM, визуализация данных

2 Введение

2.1 Актуальность исследования

Объединённый институт ядерных исследований (ОИЯИ) генерирует значительный объём интеллектуальной собственности: патенты, публикации, программное обеспечение, базы данных. Систематический анализ этого массива критически важен для:

- Выявления ключевых направлений научно-технологического развития
- Оценки уровня патентной активности по различным областям
- Идентификации междисциплинарных связей и перспективных направлений
- Поддержки принятия стратегических решений в области R&D
- Визуализации структуры научной деятельности института

Традиционные методы ручной экспертизы неэффективны при работе с большими объёмами данных. Применение больших языковых моделей (LLM) и машинного обучения открывает новые возможности для автоматизированного извлечения знаний и визуализации сложных взаимосвязей.

2.2 Цели и задачи

Цель: создание автоматизированной системы анализа и визуализации патентного ландшафта ОИЯИ с использованием LLM и машинного обучения.

Задачи:

1. Парсинг и сбор данных об объектах интеллектуальной собственности ОИЯИ из различных источников
2. Структурирование и унификация собранных данных в единый формат
3. Автоматическое извлечение научных областей из документов с помощью LLM

4. Генерация векторных представлений (эмбедингов) для документов и тематик
5. Реализация иерархической кластеризации документов по научным областям
6. Создание интерактивной системы визуализации патентного ландшафта
7. Проведение экспериментального анализа и интерпретация результатов

3 Методология исследования

3.1 Концептуальный подход

Центральная идея проекта - многоэтапное преобразование неструктурированных текстовых данных из различных источников в структурированный, визуализируемый патентный ландшафт:

```
              →              → LLM (   ) →
→         →           → 3D
```

Подход обеспечивает автоматизацию, объективность (температура LLM = 0.0), масштабируемость и интерпретируемость результатов.

3.2 Шестиэтапный пайплайн обработки данных

3.2.1 ЭТАП 0: Сбор и унификация данных

Цель: формирование единого датасета из различных источников интеллектуальной собственности ОИЯИ.

Источники данных:

- **Патентные базы данных:** lens.org, WIPO Patentscope, Espacenet, ФИПС
- **Научные публикации:** официальные репозитории ОИЯИ, библиографические базы
- **Программное обеспечение:** Git-репозитории, системы регистрации ПО
- **Базы данных:** внутренние каталоги ОИЯИ

Процесс сбора:

1. **Парсинг данных:** автоматизированный сбор информации с использованием web-scraping и API

2. **Извлечение метаданных:** название, авторы, дата, идентификатор (DOI, номер патента)

3. **Извлечение текстового контента:** рефераты, описания, аннотации

4. **Унификация формата:** приведение к единой структуре JSONL

Структура унифицированного документа:

```
{
  "id": 0,
  "name": "           ",
  "type_of_document": "patent|article|software|databases",
  "date": "YYYY.MM.DD",
  "authors": "        ",
  "identifier": "DOI          ",
  "text_of_document": "           "
}
```

Результат: единый датасет `full_dataset.jsonl` из 1948 документов, включающий:

- Патенты на изобретения и полезные модели
- Научные публикации
- Зарегистрированное программное обеспечение
- Базы данных

Модуль: `src/id_assigner.py`
Алгоритм двухпроходного присвоения ID:

1. Первый проход по файлу: сбор всех существующих ID в множество
2. Второй проход: для документов без ID присваивается минимальный свободный идентификатор (заполнение “дыр” в нумерации)
3. Замена исходного файла обработанным

Математическая формализация:

$$\text{Пусть } D = \{d_1, d_2, \ldots, d_n\} \text{ - множество документов}$$
$$ID_{\text{существующие}} = \{id(d_i) \mid d_i \in D, id(d_i) \neq \emptyset\}$$
$$\text{Для каждого } d_j \text{ без ID: } id(d_j) = \min(\mathbb{N} \setminus ID_{\text{существующие}})$$

Результат: 1948 документов с уникальными ID от 0 до 1947.

3.2.2 ЭТАП 2: Извлечение научных тегов с помощью LLM

Модуль: `src/tag_extractor.py`

Цель: автоматическое извлечение научных областей (тегов) из текстовых описаний документов с присвоением весов важности.

Технологический стек:

- **LLM модель:** OpenAI GPT-4o-mini (через OpenRouter API)
- **Температура:** 0.0 (обеспечение детерминированности)
- **Промпт-инжиниринг:** специализированный системный промпт

Требования к тегам:

- Только английский язык (валидация через регулярные выражения)
- Длина: 1-6 слов
- Формат: первое слово с заглавной буквы
- Веса: $w \in (0, 1]$, $\sum w_i = 1.0$

Математическая модель:

$$\text{Для документа } d_j: \quad Tags(d_j) = \{(t_1, w_1), (t_2, w_2), \ldots, (t_k, w_k)\}$$
$$\text{где } \sum_{i=1}^{k} w_i = 1, \quad w_i > 0$$

Примеры извлечённых тегов:

- Machine learning: 0.6, Neural networks: 0.3, Optimization methods: 0.1
- High energy physics: 0.7, Particle physics: 0.2, Experimental data: 0.1
- Nuclear reactions: 0.6, Monte Carlo methods: 0.3, Radiation safety: 0.1

Механизм надёжности:

- Exponential backoff для retry (1-60 секунд)
- До 10 попыток при сетевых ошибках
- Обработка пустых ответов

Результат: файл с тегами и весами для каждого документа.

3.2.3 ЭТАП 3: Генерация эмбедингов тегов

Модуль: `src/tag_embeddings_generator.py`
Модель эмбедингов: BAAI/bge-large-en-v1.5

- Размерность: 1024D
- Оптимизация: MPS (Mac M1) / CUDA / CPU
- Нормализация: включена (для корректного косинусного сходства)

Алгоритм:

1. Сбор всех уникальных тегов: $T = \{t_1, t_2, \ldots, t_m\}$
2. Сортировка (обеспечение детерминированности)
3. Добавление контекстного префикса: “Scientific theme: {tag}”
4. Batch-генерация эмбедингов (batch_size = 16)
5. Сохранение в формате NumPy compressed (.npz)

Математическая формализация:

$$\text{Для каждого тега } t_i: \quad embedding(t_i) = BGE(\text{“Scientific theme: ”} + t_i) \in \mathbb{R}^{1024}$$
$$\|embedding(t_i)\| = 1 \text{ (L2-нормализация)}$$

Оптимизация памяти: float32, compressed .npz, batch processing

Результат: файл `unique_tag_embeddings.npz` с эмбедингами всех уникальных тегов.

3.2.4 ЭТАП 4: Генерация эмбедингов документов

Модуль: `src/document_embeddings_generator.py`

Алгоритм взвешенной агрегации:

Для документа d_j с тегами $Tags(d_j) = \{(t_1, w_1), \ldots, (t_k, w_k)\}$:

$$embedding(d_j) = \sum_{i=1}^{k} w_i \cdot embedding(t_i)$$

Нормализация: $embedding(d_j) \leftarrow \dfrac{embedding(d_j)}{\|embedding(d_j)\|}$

Обоснование: взвешенная сумма сохраняет семантические отношения, нормализация обеспечивает корректность косинусного сходства, линейная комбинация учитывает многоаспектность документов.

Пример вычисления:

```
doc_tags = {
    "Machine learning": 0.6,
    "Neural networks": 0.3,
    "Optimization": 0.1
}
doc_embedding = (
    0.6 * emb["Machine learning"] +
    0.3 * emb["Neural networks"] +
    0.1 * emb["Optimization"]
)
doc_embedding /= np.linalg.norm(doc_embedding)
```

Результат: файл `document_embeddings.npz` (1948 × 1024).

3.2.5 ЭТАП 5: Иерархическая кластеризация

Модуль: `src/hierarchical_clustering.py`

Метод кластеризации:

- **Алгоритм:** Hierarchical Agglomerative Clustering
- **Метрика:** Cosine distance: $d(x, y) = 1 - \cos(x, y)$
- **Linkage:** Average (UPGMA - Unweighted Pair Group Method with Arithmetic mean)

Математическая формализация:

Дано: $X = \{x_1, \ldots, x_n\}$ - эмбединги документов, $x_i \in \mathbb{R}^{1024}$

Косинусное расстояние: $d(x_i, x_j) = 1 - \dfrac{x_i \cdot x_j}{\|x_i\| \cdot \|x_j\|} = 1 - x_i \cdot x_j$ (т.к. $\|x_i\| = 1$)

Average linkage: $d(C_i, C_j) = \dfrac{1}{|C_i||C_j|} \displaystyle\sum_{x \in C_i} \sum_{y \in C_j} d(x, y)$

Алгоритм Weighted Tag Frequency для названий кластеров:

Для кластера C_k:

1. Собрать все теги документов: $T_k = \bigcup_{d \in C_k} Tags(d)$
2. Вычислить суммарные веса: $score(t) = \sum_{d \in C_k : t \in Tags(d)} weight(d, t)$
3. Сортировать по убыванию $score(t)$
4. Название = топ-3 тега (адаптивно в зависимости от количества кластеров)

Метрики качества кластеризации:

1. **Silhouette Score:**

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

где $a(i)$ - среднее расстояние до точек своего кластера

$b(i)$ - среднее расстояние до точек ближайшего кластера

Диапазон: $[-1, 1]$, чем выше - тем лучше

2. **Davies-Bouldin Index:** чем меньше - тем лучше компактность кластеров

3. **Calinski-Harabasz Score:** чем больше - тем лучше разделение

Визуализация:

- **Дендрограмма:** иерархическая структура слияния кластеров
- **UMAP 2D:** проекция эмбедингов на плоскость

Результат: файлы `clustering_results.json`, `dendrogram.png`, `umap_clusters.png`.

3.2.6 ЭТАП 6: Интерактивная 3D визуализация

Модули:

- `src/visualization/generate_clustering_cache.py` - генерация кэша
- `src/visualization/app.py` - веб-приложение
- `run_visualization.sh` - автоматический запуск

Технологический стек:

- **Backend:** Python, Dash
- **Frontend:** Plotly.js (WebGL 3D), React (через Dash)
- **UI:** Dash Bootstrap Components
- **Снижение размерности:** t-SNE (1024D $\rightarrow$ 3D)

Архитектура кэширования:

1. Генерация 79 уровней кластеризации (от 1 до 1948 кластеров)
2. Для каждого уровня: центроиды, цвета, названия, топ-10 тегов
3. Сохранение в JSON ($\sim$800 MB)
4. Время загрузки: 2-3 секунды

5. Переключение уровней: <100 мс

Математика визуализации:

1. **t-SNE проекция 3D:**

Минимизация дивергенции Кульбака-Лейблера:

$$KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

где P - распределение в высокоразмерном пространстве

Q - распределение в 3D проекции

2. **Размер сферы кластера:**

Формула 1 (по умолчанию): $r = \frac{\sqrt{n} \cdot \log(n+1)}{scale}$

Формула 2: $r = \frac{\sqrt[3]{n}}{scale}$

Формула 3: $r = \frac{\log(n+1)}{scale}$

где n - количество документов в кластере

$scale$ - масштабный коэффициент (настраиваемый)

3. **Наследование цветов в иерархии:**

При слиянии кластеров C_1 и C_2 :

$color(C_{1+2}) = \frac{color(C_1) + color(C_2)}{2}$ (покомпонентно в RGB)

Интерфейс (3 панели):

- **Левая (25%):** список всех кластеров с цветами и названиями
- **Центральная (50%):** 3D визуализация кластеров-сфер
- **Правая (25%):**
  - Логарифмический слайдер количества кластеров
  - Слайдер масштаба сфер

- Выбор формулы радиуса
- Информация о выбранном кластере (топ-5 тегов, количество документов, ID)

Адаптивные названия кластеров:

```
if n_clusters < 5:          = "Tag1, Tag2, Tag3, ..."
elif n_clusters < 10:       = "Tag1, Tag2, ..."
else:                       = "Tag1"
```

Результат: веб-приложение http://144.124.229.26:8050

3.3 Используемые модели и технологии

Программное обеспечение: Python 3.11, PyTorch 2.0+, Sentence-Transformers 2.2+, scikit-learn 1.3+, Dash 2.14+, Plotly 5.17+

Модели машинного обучения:

- **LLM:** OpenAI GPT-4o-mini (извлечение тегов)
- **Эмбединги:** BAAI/bge-large-en-v1.5 (векторизация текстов)
- **Кластеризация:** Scipy Hierarchical Clustering (Average linkage, Cosine metric)
- **Визуализация:** t-SNE (снижение размерности для 3D)

Формат данных: JSONL (построчный JSON), NumPy compressed .npz (float32), JSON (кластеризация), PNG (визуализация)

4 Аналитическая часть: Результаты экспериментов

4.1 Эксперимент с 4 кластерами

Цель: выявление крупномасштабной структуры научной деятельности ОИЯИ через агрегацию документов в 4 основных направления.

Результаты кластеризации:

Таблица 1: Результаты кластеризации (4 кластера)

Кластер	Название	Документов	%	Интерпретация
#1	Nuclear physics, Accelerator Physics, High Energy Physics	1821	93.5%	Основная миссия ОИЯИ - фундаментальные исследования
#0	Numerical Analysis, Software Engineering, Grid Computing	125	6.4%	Поддерживающая инфраструктура - обработка данных, моделирование
#2	Supersymmetric Integrable Spin Chains	1	0.05%	Теоретическая математическая физика
#3	Fiber Optic Sensors	1	0.05%	Инструментальные разработки

3D визуализация (4 кластера):

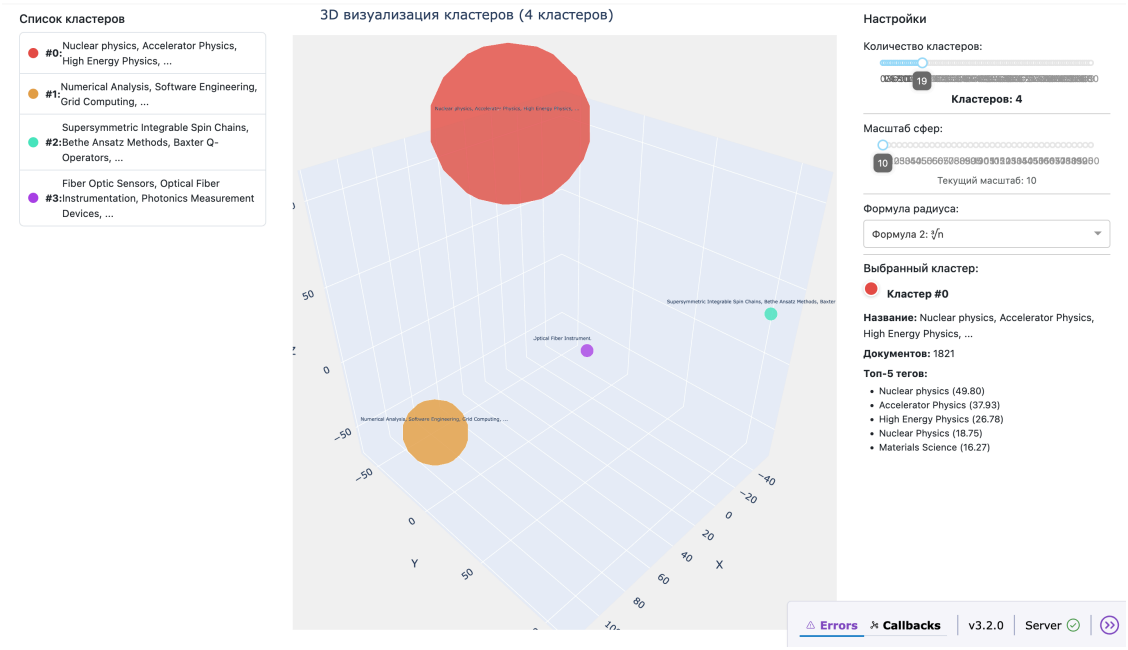

Рис. 1: 3D визуализация 4 кластеров (t-SNE проекция)

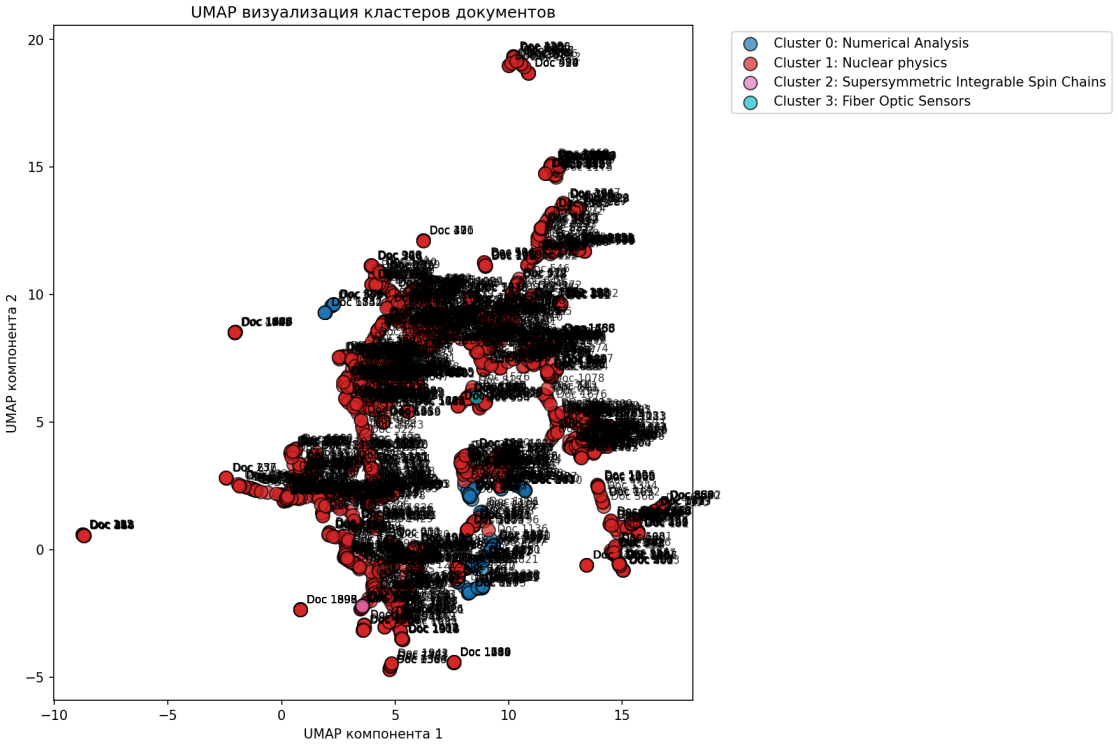

Рис. 2: UMAP 2D проекция 4 кластеров

Анализ результатов:

1. **Доминирование ядерной физики (Кластер #1):** 93.5% всех документов. Суммарный вес тега “Nuclear physics”: 49.8 - показывает концентрацию на фундаментальных исследованиях. Тесная связь с Accelerator Physics (37.93) и High Energy Physics (26.78). **Интерпретация:** основная миссия ОИЯИ.

2. **Вычислительные науки (Кластер #0):** 6.4% - значительная, но вторичная область. Фокус: численный анализ (3.25), разработка ПО (3.05), Grid-вычисления (2.0). **Интерпретация:** поддерживающая инфраструктура для экспериментальной физики.

3. **Уникальные нишевые исследования (Кластеры #2 и #3):** по 1 документу (0.05%). Теоретическая математическая физика и волоконно-оптические сенсоры. **Интерпретация:** узкоспециализированные направления с высоким патентным потенциалом.

Метрики качества: Silhouette Score: 0.334 | Davies-Bouldin: 1.037 | Calinski-Harabasz: 2.82

Выводы:

- Ядерная физика - абсолютное ядро (93.5%)
- Вычислительные науки - вторичная область (6.4%)
- Есть уникальные нишевые направления (0.1%)

4.2 Эксперимент с 10 кластерами

Цель: детализация научной структуры ОИЯИ путём выделения специализированных подобластей.

Ключевые кластеры:

Таблица 2: Ключевые кластеры (10 кластеров)

Кластер	Название	Документов	%	Значение
#1	Nuclear physics	1662	85.3%	Ядро ОИЯИ (снизилось с 93.5%)
#8	Neutron Activation Analysis	129	6.6%	Выделился из основного - аналитическая химия, экомониторинг
#2	Software Engineering	62	3.2%	Grid-вычисления, управление ИС
#4	Numerical Analysis	59	3.0%	Математическая база моделирования
#3	Power Electronics	28	1.4%	Высокий патентный потенциал - электроника для ускорителей
#0	Magnetic drug delivery	2	0.1%	Перспективное направление - биомедицина
#9	Stepper Motor Control	3	0.15%	Автоматизация установок

3D визуализация (10 кластеров):

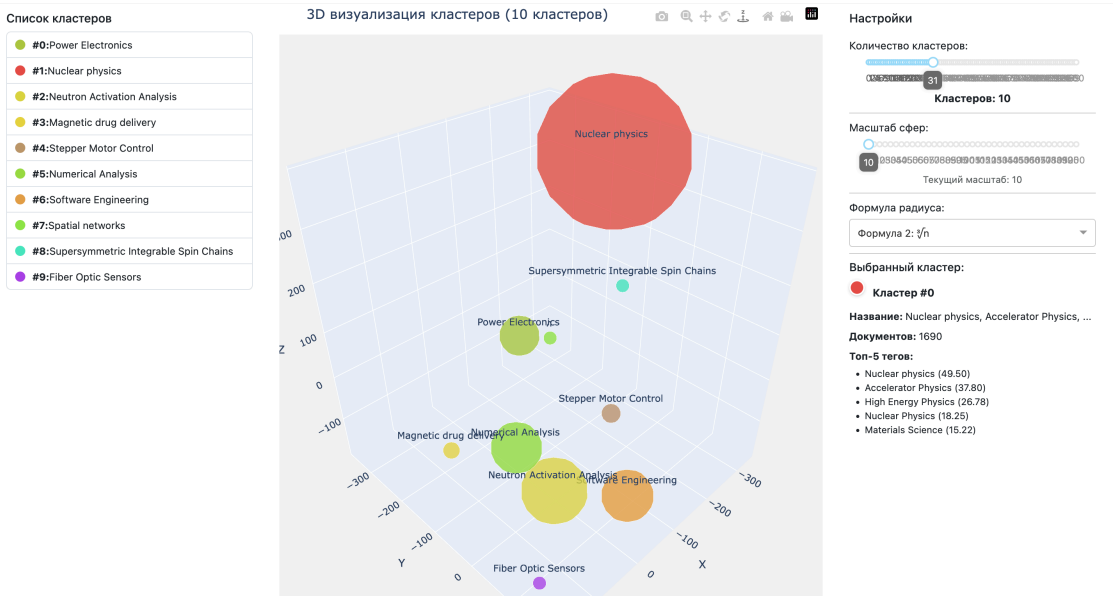

Рис. 3: 3D визуализация 10 кластеров (t-SNE проекция)

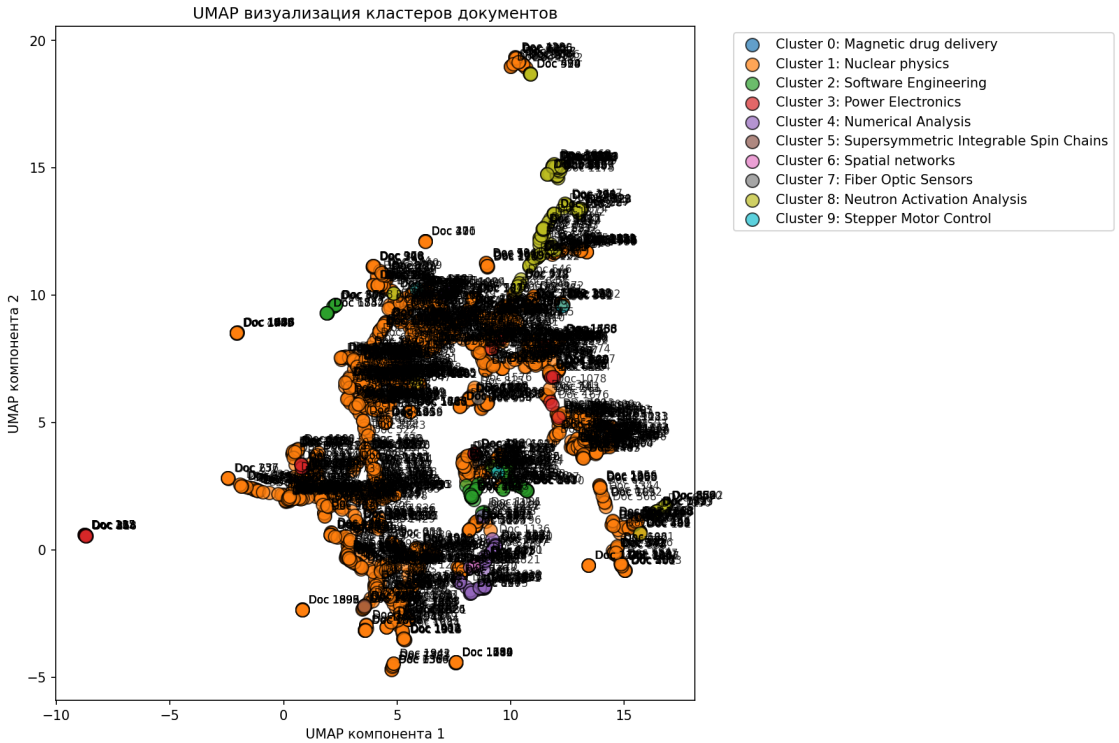

Рис. 4: UMAP 2D проекция 10 кластеров

Анализ результатов:

1. **Ядерная физика остаётся доминантой (Кластер #1):** 85.3% (снижение с 93.5% при 4 кластерах). При детализации из основного кластера выделились специализированные подобласти.

2. **Аналитическая химия и нейтронные методы (Кластер #8):** 6.6% - крупная специализированная область. Фокус: Neutron Activation Analysis (4.95), Analytical Chemistry (4.85). **Практическое значение:** высокий потенциал для прикладных патентов (анализ загрязнений, медицинская диагностика).

3. **Силовая электроника (Кластер #3):** 1.4% (28 документов). Специализация: Power Electronics (2.35), High Voltage Engineering (1.15). **Интерпретация:** разработка электронных систем для ускорителей и детекторов - высокий патентный потенциал.

4. **Биомедицинские приложения (Кластер #0):** 2 документа, но высокая специализация. Магнитная доставка лекарств, биофизика липидных мембран. **Интерпретация:** новое, перспективное направление на стыке ядерной физики и медицины.

5. **Программная инженерия (Кластер #2):** 3.2%. Software Engineering (2.8), Grid Computing (2.0), Intellectual Property Management (1.05). ОИЯИ разрабатывает сложное ПО для обработки данных.

6. **Численный анализ (Кластер #4):** 3.0%. Чистая математика: теория аппроксимации, дифференциальные уравнения, функциональный анализ.

7. **Системы управления (Кластер #9):** 3 документа. Управление шаговыми двигателями, реакторами, квантовые нечёткие системы.

Метрики качества: Silhouette Score: $\sim$0.28 (снижение из-за дробления, но приемлемо)

Выводы:

- При детализации выявились **специализированные подобласти**
- Междисциплинарность: от теоретической физики до биомедицины
- **Патентный потенциал:** силовая электроника (#3), нейтронный анализ (#8), биомедицина (#0)

4.3 Сравнительный анализ

Динамика:

- 4 кластера: один гигант (93.5%) + периферия
- 10 кластеров: один доминант (85.3%) + спектр специализаций

При детализации:

- Nuclear physics → Nuclear physics (85.3%) + Neutron Activation (6.6%) + прочие (8.1%)
- Computational → Software (3.2%) + Numerical (3.0%) + Power Electronics (1.4%)

Дендрограмма:

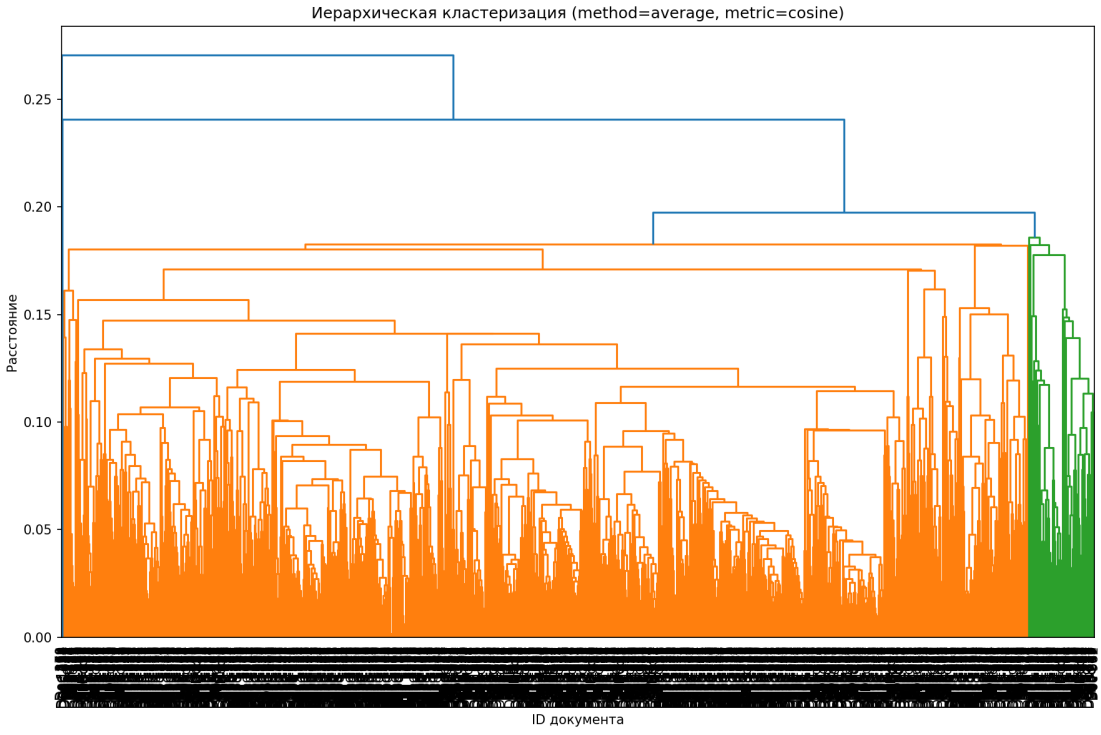

Рис. 5: Дендрограмма иерархической кластеризации

Дендрограмма показывает:

- Два суперкластера сливаются на расстоянии ∼0.24
- Высокая когерентность внутри кластеров (слияние на 0.03-0.15)

- Малая группа специализированных документов (зелёный кластер)

Ключевой вывод: ядерная физика устойчиво остаётся ядром, но детализация раскрывает скрытую междисциплинарную сложность.

5 Результаты и выводы

5.1 Структура научной деятельности ОИЯИ

Уровень 1: Фундаментальные исследования (85-95%)

Ядерная и физика высоких энергий, ускорительные технологии, квантовая хромодинамика

Уровень 2: Прикладные направления (5-10%)

- **Нейтронно-активационный анализ** (6.6%) - высокий патентный потенциал
- **Вычислительные науки** (3-6%) - Grid, численные методы, ПО
- **Силовая электроника** (1.4%, 28 документов) - готовые технологии для патентования

Уровень 3: Междисциплинарные направления (<1%)

- **Биомедицинские приложения** - перспективная область для развития
- **Теоретическая математическая физика** - точечные глубокие исследования

5.2 Патентная активность

Высокий потенциал (готовые технологии):

- **Силовая электроника** (28 док.): резонансные зарядные системы, радиационно-стойкие микросхемы
- **Нейтронный анализ** (129 док.): экомониторинг, анализ микропримесей, биомедицина
- **Оптические сенсоры** (1 док.): волоконно-оптические датчики для радиации

Средний потенциал:

- **ПО** (62 док.): Grid-системы, регистрация и лицензирование
- **Системы управления** (3 док.): интеллектуальное управление реакторами

Перспективный (долгосрочный):

- **Биомедицина** (2 док.): магнитная доставка лекарств - требует расширения и партнёрств

5.3 Практическая значимость визуализатора

Возможности (http://144.124.229.26:8050):

1. **Стратегическое планирование:** оценка распределения ресурсов, выявление недоразвитых направлений
2. **Анализ трендов:** динамическая детализация 1-1948 кластеров, обнаружение уникальных направлений
3. **Поддержка решений:** приоритеты финансирования, поиск коллабораций, оценка патентоспособности
4. **Коммуникация:** наглядная демонстрация для партнёров, грантовых заявок, обучения сотрудников

Технические преимущества:

- Масштабируемость, мгновенное переключение уровней (<100 мс)
- Веб-доступность без специального ПО
- Детерминированность (воспроизводимость результатов)

5.4 Рекомендации

Для патентования:

1. Приоритет: силовая электроника (28 патентов), нейтронный анализ (50+ применений)

2. Междисциплинарность: физика+медицина (нейтронная терапия), ускорители+энергетика

Для развития:

1. Усиление: биомедицина (<1% → 5%), нанотехнологии, квантовые технологии
2. Коллаборации: использовать визуализатор для поиска партнёров
3. Система анализа: добавить временную динамику, интеграцию с базами цитирований

6 Заключение

Ключевые достижения:

- Собран единый датасет из 1948 документов различных типов (патенты, публикации, ПО, БД) путём парсинга множественных источников
- Разработан оригинальный шестиэтапный пайплайн с алгоритмом Weighted Tag Frequency
- Построена иерархическая структура из 79 уровней детализации научных направлений ОИЯИ
- Создан высокопроизводительный интерактивный визуализатор с переключением между 1-1948 кластерами <100 мс

Аналитические результаты:

- Подтверждён фокус на ядерной физике (85-95%)
- Выявлены направления для патентования: электроника (28), нейтронный анализ (129), биомедицина (2)
- Обнаружены междисциплинарные связи фундаментальных и прикладных исследований

Практическая значимость:

Система применима для стратегического планирования, патентного анализа, поиска коллабораций, мониторинга трендов и коммуникации с партнёрами. Масштабируема для любых научных организаций.

Перспективы: временная динамика, интеграция с цитированиями, прогностические модели, сравнительный анализ институтов.

7 Список литературы

1. **Reimers N., Gurevych I.** Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // EMNLP 2019. – P. 3982-3992.

2. **van der Maaten L., Hinton G.** Visualizing Data using t-SNE // JMLR. – 2008. – Vol. 9. – P. 2579-2605.

3. **Rousseeuw P.J.** Silhouettes: A graphical aid to cluster analysis // J. Comp. Appl. Math. – 1987. – Vol. 20. – P. 53-65.

4. **Müllner D.** Modern hierarchical, agglomerative clustering algorithms // arXiv:1109.2378. – 2011.

5. **OpenRouter API Documentation** – `https://openrouter.ai/docs`

6. **Plotly Dash Framework** – `https://dash.plotly.com/`

7. **ОИЯИ** – Официальный сайт – `https://www.jinr.ru/`

8. **GitHub Repository** – JINR Hackathon Autumn 2025 – `https://github.com/MansurYa/JINR-hackathon-autumn-2025`

Интерактивная визуализация: http://144.124.229.26:8050

Репозиторий проекта: https://github.com/MansurYa/JINR-hackathon-autumn-2025