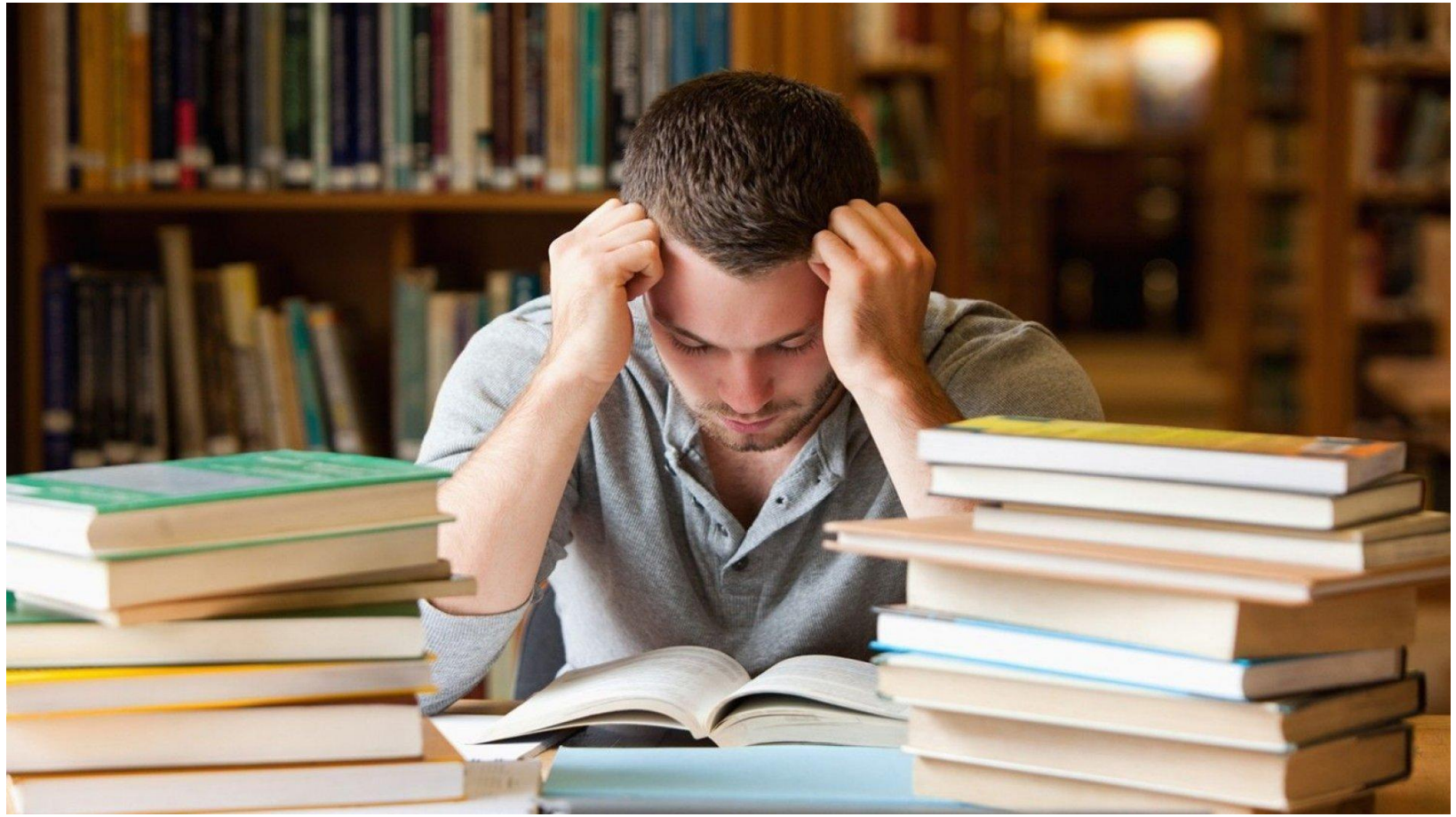
АНАЛИЗ ТЕКСТОВ ГЕОЛОГИЧЕСКИХ ПУБЛИКАЦИЙ С ИСПОЛЬЗОВАНИЕМ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Патук М.И., Наумова В.В.

Государственный геологический музей им. В.И.Вернадского РАН

Анализ текстов человеком



Обработка естественного языка (NLP) - история

- 1 этап (середина 1950-х – конец 1980-х)
  - Подход основанный на правилах
    - Переводчики
    - Чат-боты
- 2 этап (конец 1980-х – настоящее время)
  - Подход основанный на статистике
  - 3 этап (2017 - появление нейросетей – BERT)
    - Большие языковые модели (LLM)

Обработка естественного языка (NLP) - задачи

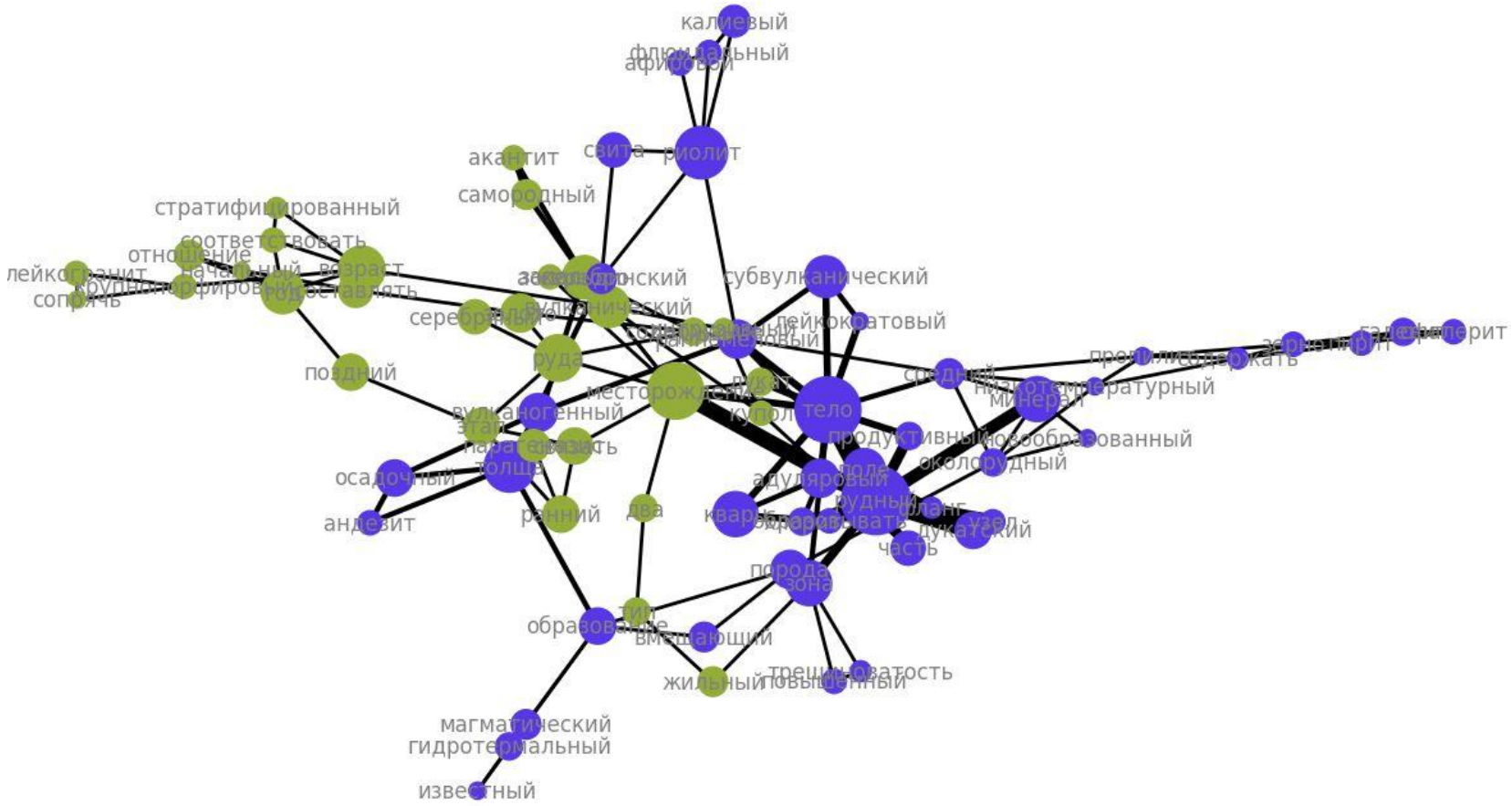
- Машинный перевод;
- Распознавание именованных сущностей;
- Классификация и кластеризация;
- Суммаризация;
- Семантическая близость;
- Генерация текста;
- …

Первые эксперименты по анализу геологических текстов

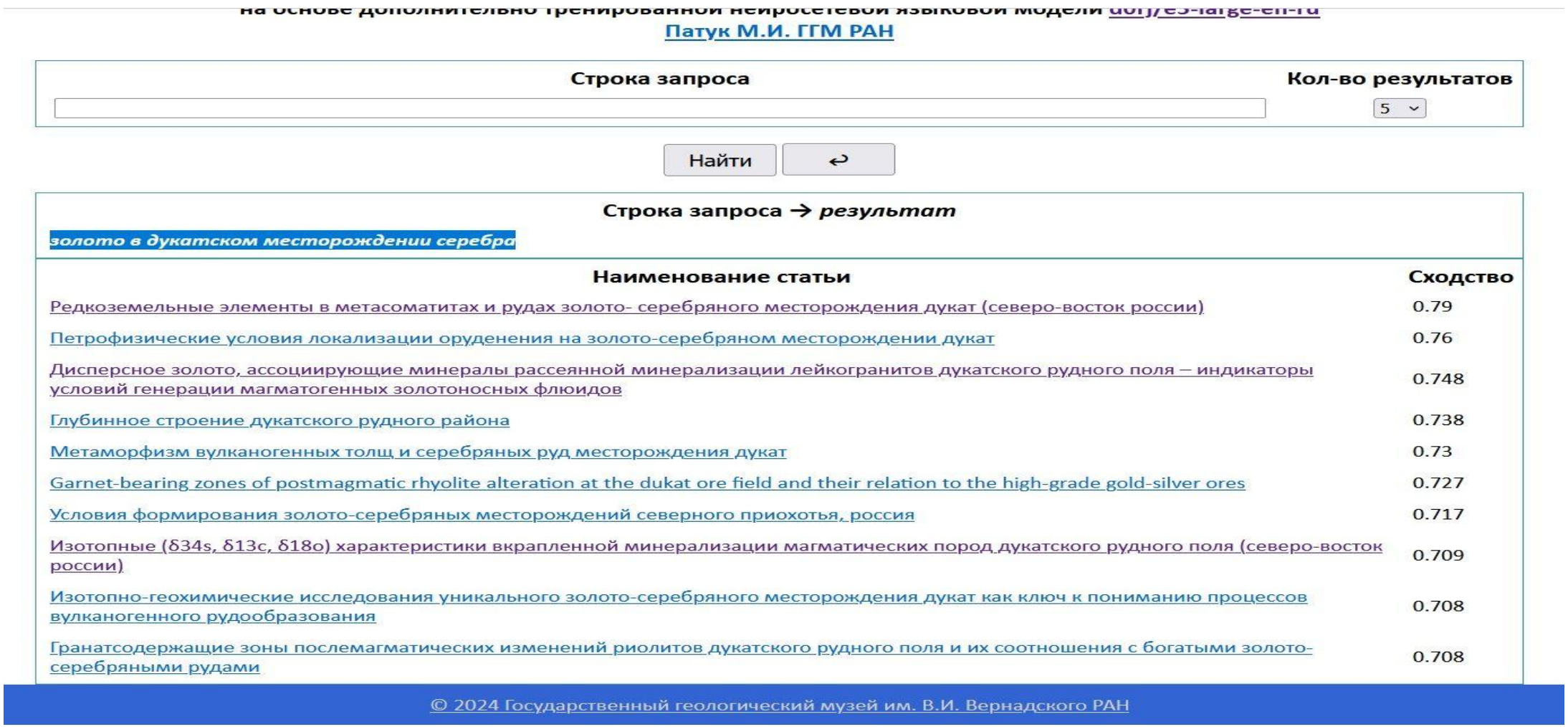
- Задача бинарной классификации описания геологических объектов
- Задача реферирования (извлечение ключевых слов) из текстов научных публикаций

Задачи решаются на основе дополнительно тренированной ЯМ (T5 – 244M параметров) с помощью текстов научных публикаций из архива публикаций ГГМ РАН (**43 000** публикация) и Wiki-педии ГГМ РАН (**959** страниц текстов на русском языке)

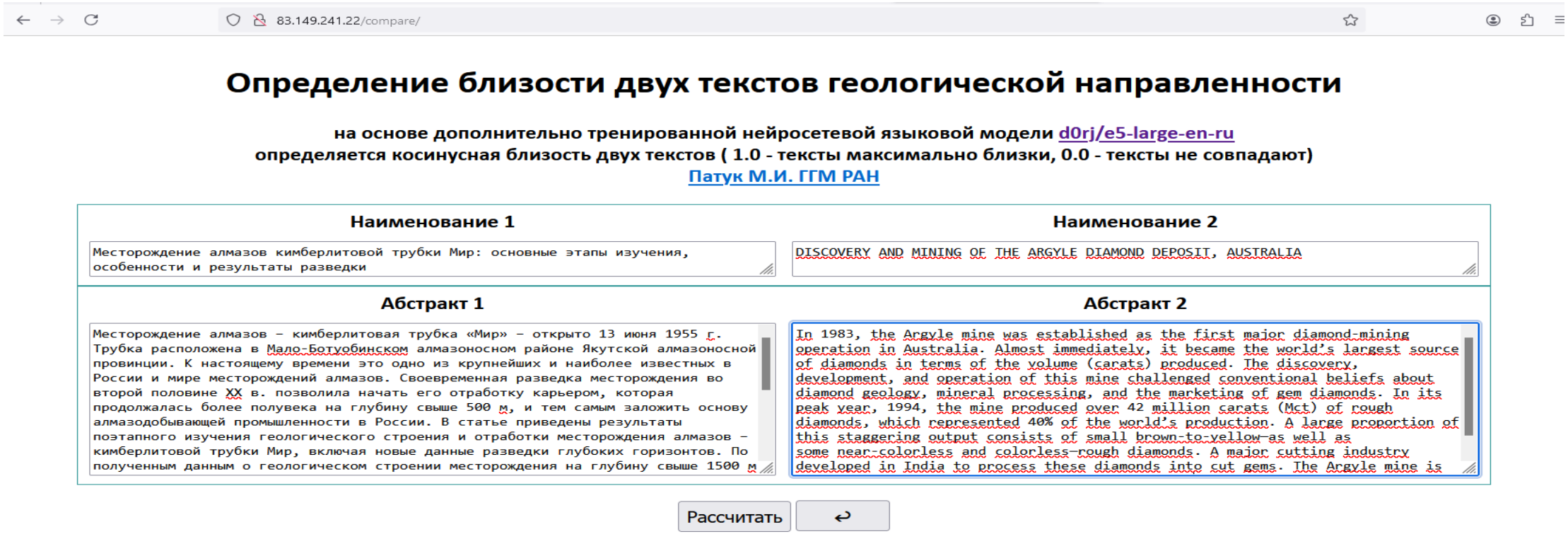
Ядро графа месторождения Дукат



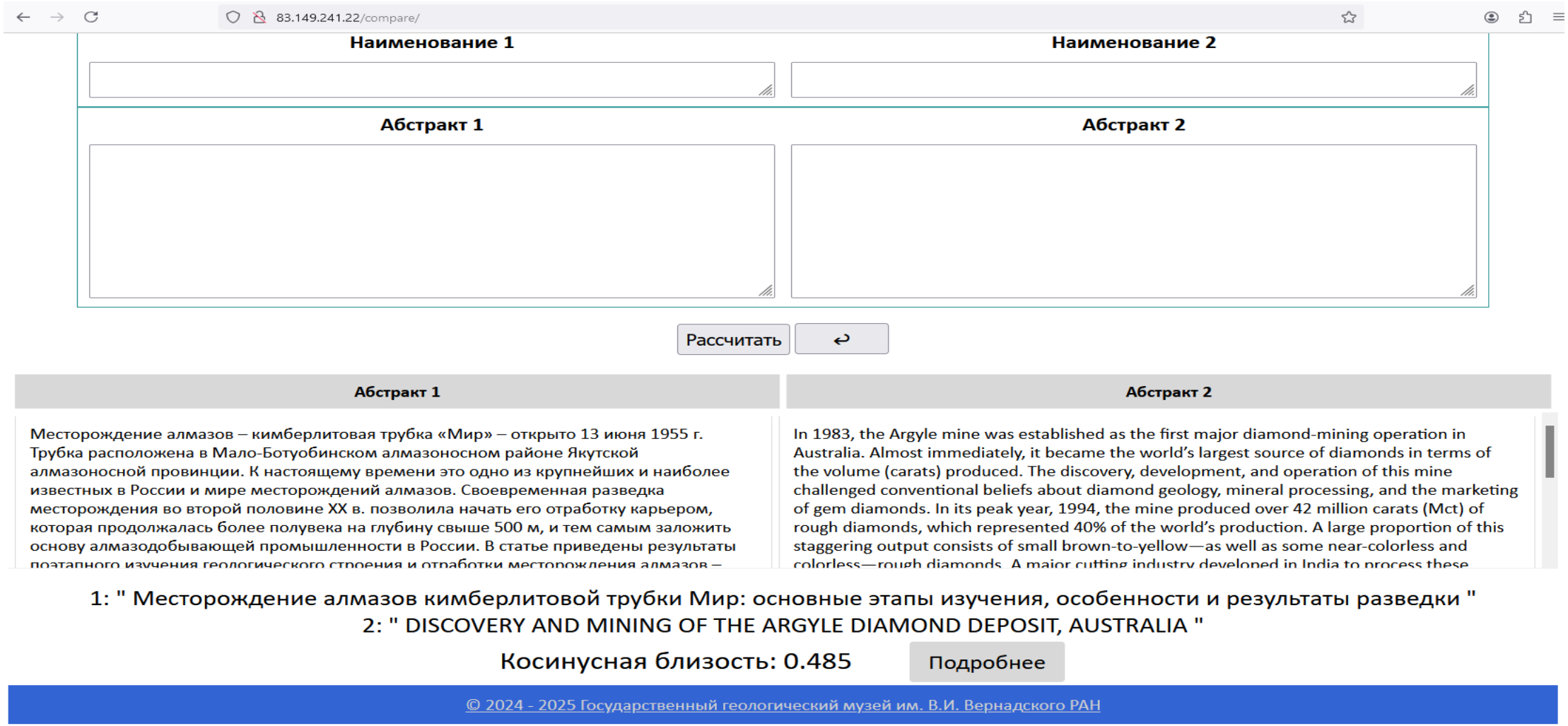
«Поиск семантически близких публикаций»



«Определение близости двух текстов геологической направленности»



«Определение близости двух текстов геологической направленности» - результат

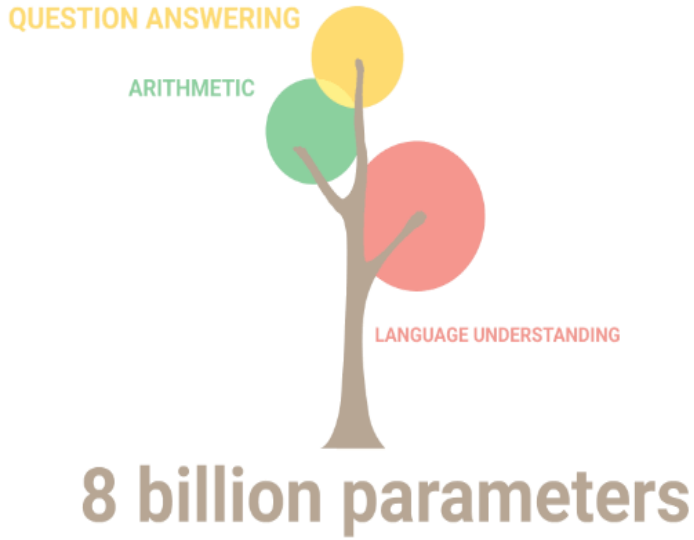


«Определение близости двух текстов геологической направленности» - развернутый результат

Различия:

1. **Тип месторождения и геология:**
   - **Мир:** Четко указан тип - **кимберлитовая трубка**. Текст фокусируется на **изменениях геологического строения с глубиной** (увеличение объема рудного тела >1500 м, стабильность состава).
   - **Аргайл:** Тип месторождения *не указан* в тексте (известно, что это **лампроитовая трубка**, что и было вызовом традициям). Текст **не содержит деталей геологии**, акцент на историю добычи и характеристики продукции.
2. **Характер продукции (Алмазы):**
   - **Мир:** В тексте **не описываются** конкретные характеристики добываемых алмазов (размер, цвет, качество). Упор на объем руды и запасы алмазов в каратах.

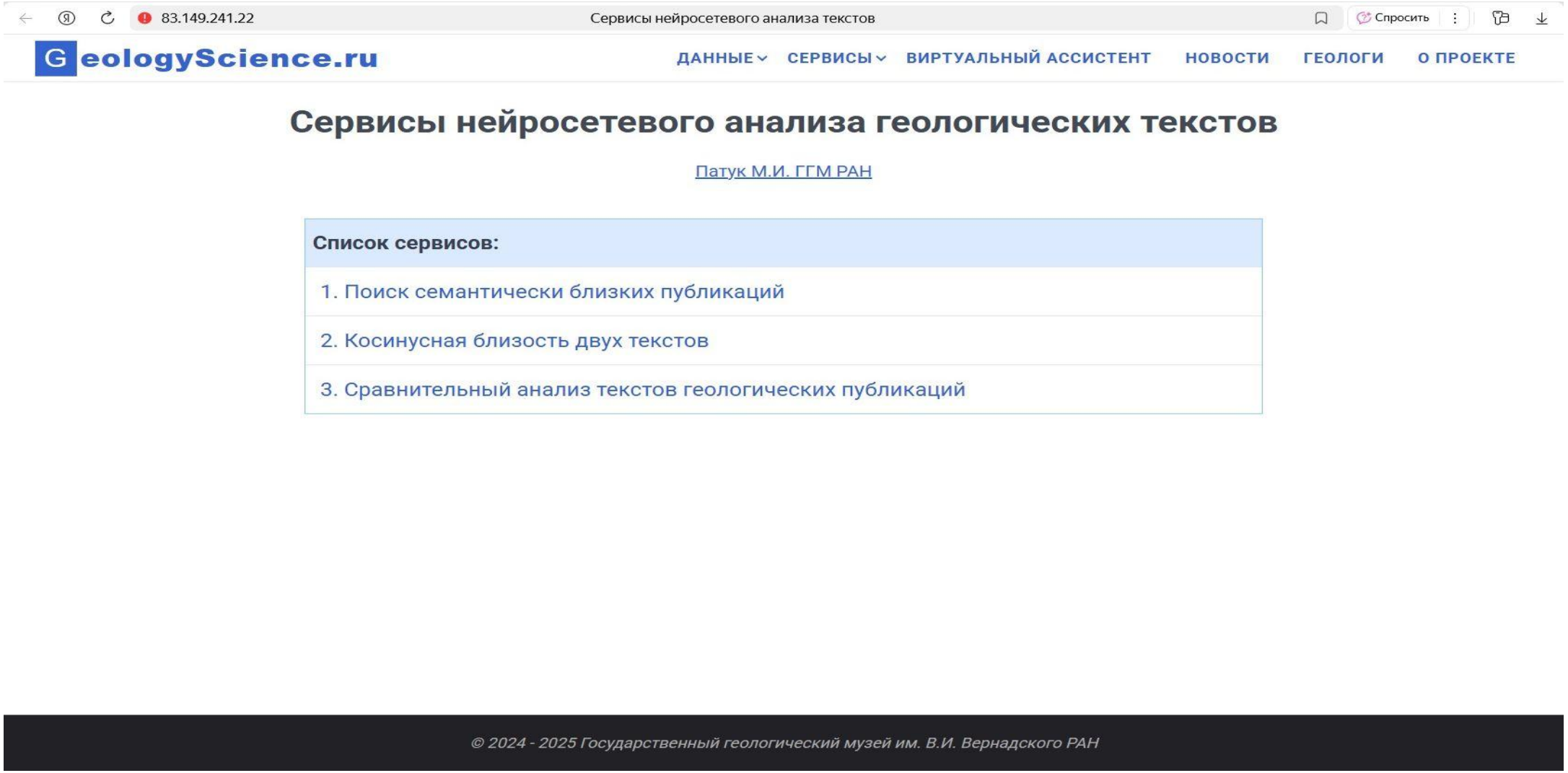
Большие языковые модели (LLM) - развитие



Анализ текстов с помощью LLM DeepSeek R1 - результат

Document

Сервисы анализа текстов http://83.149.241.22/



Заключение

- Современные Большие Языковые Модели являются лидирующим инструментом для анализа текстов научных статей.
- Возможности семантического поиска, анализа графов, методов суммаризации и сравнительного анализа значительно расширяют возможности анализа публикаций.

Спасибо за внимание!

Работы выполняются в рамках темы государственного задания ГГМ РАН No 1021061009468-8-1.5.1 «Цифровая платформа интеграции и анализа геологических и музейных данных».