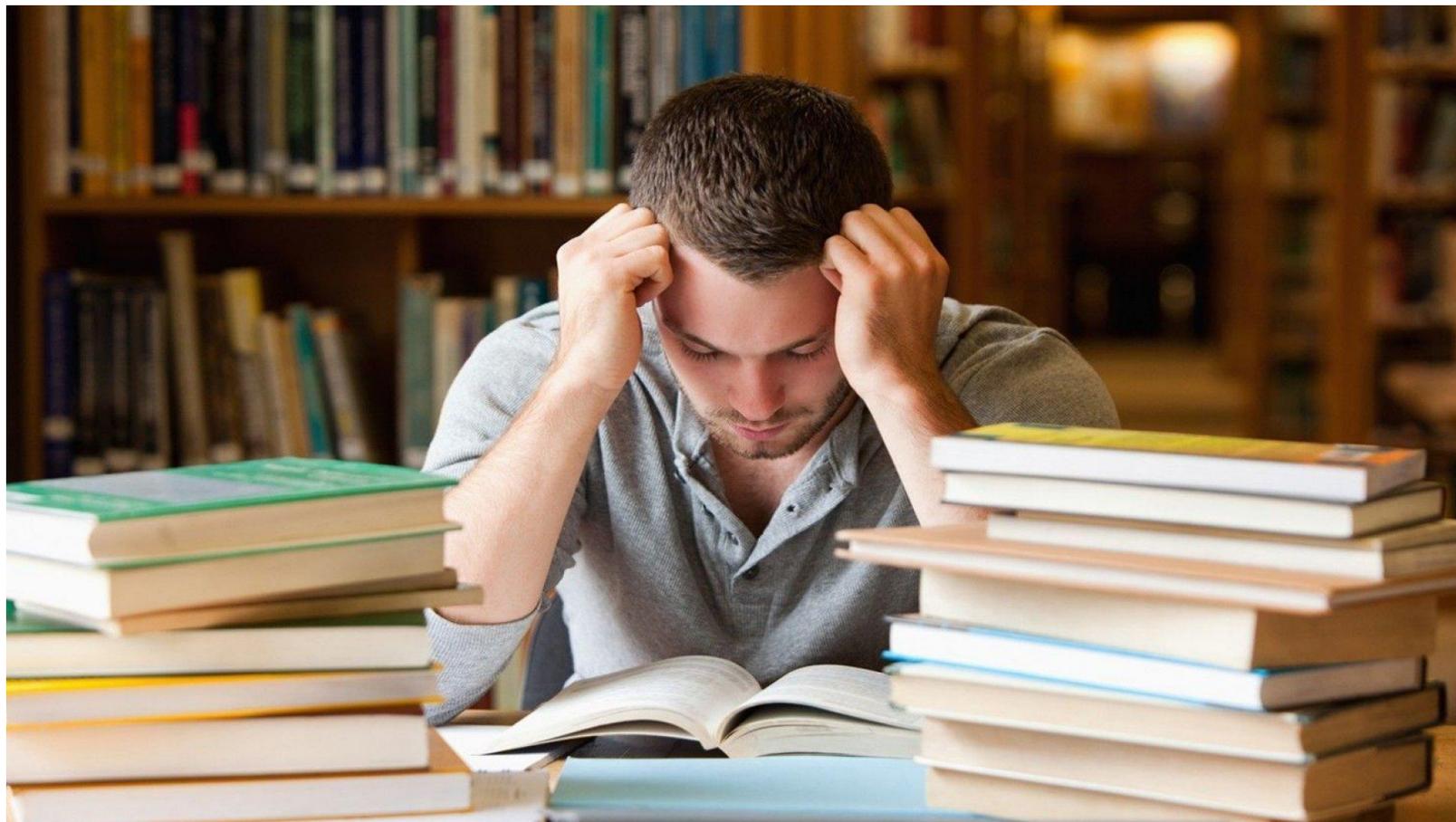


АНАЛИЗ ТЕКСТОВ ГЕОЛОГИЧЕСКИХ ПУБЛИКАЦИЙ С ИСПОЛЬЗОВАНИЕМ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

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 – 244М 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)

«Поиск семантически близких публикаций»

на основе дополнительно тренированной нейросетевой языковой модели [gpt/ru-target-en-ru](#)

[Патук М.И. ГГМ РАН](#)

Строка запроса	Кол-во результатов
<input type="text"/>	5 ▾

Найти



Строка запроса → результат

золото в дукатском месторождении серебра

Наименование статьи	Сходство
Редкоземельные элементы в метасоматитах и рудах золото- серебряного месторождения дукат (северо-восток россии)	0.79
Петрофизические условия локализации оруденения на золото-серебряном месторождении дукат	0.76
Дисперсное золото, ассоциирующие минералы рассеянной минерализации лейкогранитов дукатского рудного поля – индикаторы условий генерации магматогенных золотоносных флюидов	0.748
Глубинное строение дукатского рудного района	0.738
Метаморфизм вулканогенных толщ и серебряных руд месторождения дукат	0.73
Garnet-bearing zones of postmagmatic rhyolite alteration at the dukat ore field and their relation to the high-grade gold-silver ores	0.727
Условия формирования золото-серебряных месторождений северного приохотья, россия	0.717
Изотопные ($\delta^{34}s$, $\delta^{13}c$, $\delta^{18}o$) характеристики вкрапленной минерализации магматических пород дукатского рудного поля (северо-восток россии)	0.709
Изотопно-геохимические исследования уникального золото-серебряного месторождения дукат как ключ к пониманию процессов вулканогенного рудообразования	0.708
Гранатсодержащие зоны послемагматических изменений риолитов дукатского рудного поля и их соотношения с богатыми золото-серебряными рудами	0.708

«Определение близости двух текстов геологической направленности»



Определение близости двух текстов геологической направленности

на основе дополнительно тренированной нейросетевой языковой модели [d0rj/e5-large-en-ru](#) определяется косинусная близость двух текстов (1.0 - тексты максимально близки, 0.0 - тексты не совпадают)

[Патук М.И. ГГМ РАН](#)

Наименование 1	Наименование 2
Месторождение алмазов кимберлитовой трубки Мир: основные этапы изучения, особенности и результаты разведки	DISCOVERY AND MINING OF THE ARGYLE DIAMOND DEPOSIT, AUSTRALIA
Абстракт 1 Месторождение алмазов – кимберлитовая трубка «Мир» – открыто 13 июня 1955 г. Трубка расположена в Мало-Ботуобинском алмазоносном районе Якутской алмазоносной провинции. К настоящему времени это одно из крупнейших и наиболее известных в России и мире месторождений алмазов. Своевременная разведка месторождения во второй половине XX в. позволила начать его отработку карьером, которая продолжалась более полувека на глубину свыше 500 м, и тем самым заложить основу алмазодобывающей промышленности в России. В статье приведены результаты поэтапного изучения геологического строения и отработки месторождения алмазов – кимберлитовой трубки Мир, включая новые данные разведки глубоких горизонтов. По полученным данным о геологическом строении месторождения на глубину свыше 1500 м	Абстракт 2 In 1983, the Argyle mine was established as the first major diamond-mining operation in Australia. Almost immediately, it became the world's largest source of diamonds in terms of the volume (carats) produced. The discovery, development, and operation of this mine challenged conventional beliefs about diamond geology, mineral processing, and the marketing of gem diamonds. In its peak year, 1994, the mine produced over 42 million carats (Mct) of rough diamonds, which represented 40% of the world's production. A large proportion of this staggering output consists of small brown-to-yellow as well as some near-colorless and colorless-rough diamonds. A major cutting industry developed in India to process these diamonds into cut gems. The Argyle mine is

Рассчитать



Введите наименования статей и абстракты в каждое из окон. Для расчета используются первые ~250 слов абстракта. Допускаются русский и английский языки.

«Определение близости двух текстов геологической направленности» - результат

Наименование 1

Наименование 2

Абстракт 1

Абстракт 2

Рассчитать

Абстракт 1	Абстракт 2
<p>Месторождение алмазов – кимберлитовая трубка «Мир» – открыто 13 июня 1955 г. Трубка расположена в Мало-Ботуобинском алмазоносном районе Якутской алмазоносной провинции. К настоящему времени это одно из крупнейших и наиболее известных в России и мире месторождений алмазов. Своевременная разведка месторождения во второй половине XX в. позволила начать его отработку карьером, которая продолжалась более полувека на глубину свыше 500 м, и тем самым заложить основу алмазодобывающей промышленности в России. В статье приведены результаты поэтапного изучения геологического строения и отработки месторождения алмазов –</p>	<p>In 1983, the Argyle mine was established as the first major diamond-mining operation in Australia. Almost immediately, it became the world's largest source of diamonds in terms of the volume (carats) produced. The discovery, development, and operation of this mine challenged conventional beliefs about diamond geology, mineral processing, and the marketing of gem diamonds. In its peak year, 1994, the mine produced over 42 million carats (Mct) of rough diamonds, which represented 40% of the world's production. A large proportion of this staggering output consists of small brown-to-yellow—as well as some near-colorless and colorless—rough diamonds. A major cutting industry developed in India to process these</p>

1: " Месторождение алмазов кимберлитовой трубки Мир: основные этапы изучения, особенности и результаты разведки "

2: " DISCOVERY AND MINING OF THE ARGYLE DIAMOND DEPOSIT, AUSTRALIA "

Косинусная близость: 0.485

[Подробнее](#)

«Определение близости двух текстов геологической направленности» - развернутый результат

Различия:

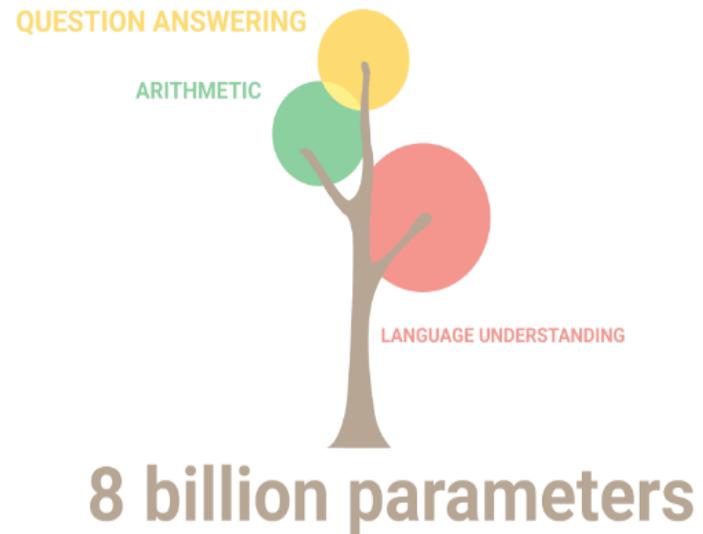
1. Тип месторождения и геология:

- **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/>

Сервисы нейросетевого анализа геологических текстов

[Патук М.И. ГГМ РАН](#)

Список сервисов:

1. Поиск семантически близких публикаций
2. Косинусная близость двух текстов
3. Сравнительный анализ текстов геологических публикаций

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».