

Искусственный интеллект и геологические архивы: новые подходы для автоматической индексации

О.А. Ядрышникова^{1,2} к.т.н.
А.Ф. Тенюнин¹
М.Л. Бычков¹

¹ФГБОУ ВО «Тюменский индустриальный университет»

²ООО «Тюменский нефтяной научный центр»

Адрес для связи: oyadrishnikova@tnnc.rosneft.ru,
aftenyunin@tnnc.rosneft.ru,
ml_bychkov2@tnnc.rosneft.ru

Ключевые слова: искусственный интеллект, геологические архивы, индексация данных, машинное обучение, обработка естественного языка (NLP), языковая модель, нейросети, трансформеры.

В работе представлено практическое применение новых подходов для автоматической индексации данных геологических архивов с помощью технологий искусственного интеллекта (ИИ). Рассматриваются преимущества использования ИИ в этой области, а также представлен обзор созданного сервиса для повышения эффективности извлечения информации из геологических данных.

Artificial Intelligence and Geological Archives: New Approaches for Automatic Indexing

O.A. Yadrishnikova^{1,2}
A.F. Tenyunin¹
M.L. Bychkov¹

¹Tyumen Industrial University

²Tyumen Oil Research Center

E-mail: oyadrishnikova@tnnc.rosneft.ru,
aftenyunin@tnnc.rosneft.ru,
ml_bychkov2@tnnc.rosneft.ru

Keywords: artificial intelligence, geological archives, data indexing, machine learning, natural language processing (NLP), language model, neural networks, transformers

This paper presents practical applications of new approaches to automatic indexing of geological archive data using artificial intelligence (AI) technologies. The benefits of using AI in this field are discussed, and an overview is provided of a developed service aimed at increasing the efficiency of extracting information from geological data.

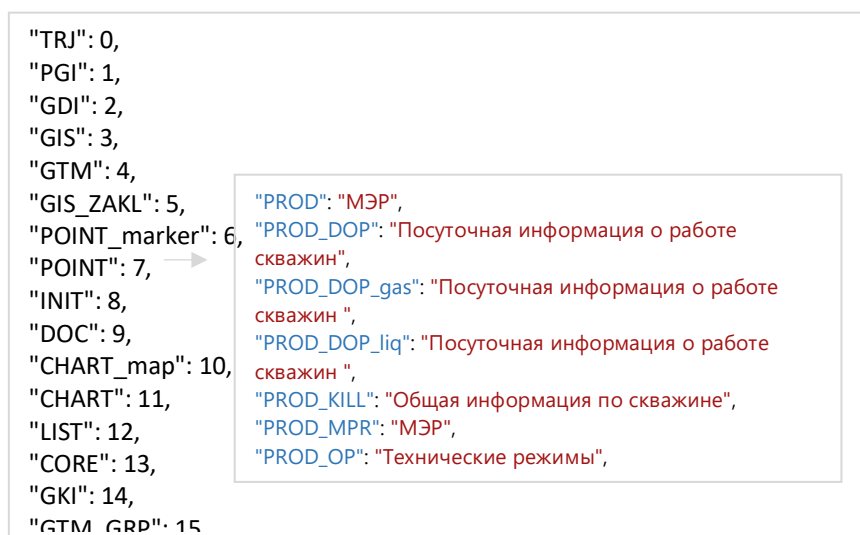
Введение

Актуальность темы обусловлена наличием огромных массивов данных в геологических архивах и базах данных нефтегазовых предприятий, накопленных десятилетиями. Геологические исследования играют важную роль в понимании природных ресурсов. Текстовые отчеты являются важным источником данных о результатах проведенных исследований. Поиск и извлечение информации из неструктурированных массивов является затруднительной работой, требуется систематизация для эффективного использования архивов. Ручная индексация данных является трудоемкой, долговременной и подвержена человеческим ошибкам.

Автоматизация индексации с помощью искусственного интеллекта открывает путь к быстрому и точному поиску информации, а генеративный искусственный интеллект уже сейчас может создавать новый контент и значительно облегчить работу аналитиков и инженеров [1], ускоряя принятие решений и повышая эффективность разведки и добычи.

Решены две задачи, такие как классификация и извлечение наименованных сущностей из текстов документов. Эти задачи важны для полной автоматической обработки и возможности анализа текстовых данных. Поиск именованных сущностей в текстах позволяет извлекать ключевую информацию из текстов, что важно для задач поиска информации. Классификация документов по их типам, в свою очередь, необходима для систем фильтрации информации, автоматической категоризации и структуризации больших объемов данных.

Для решения задачи классификации по 20-ти типам отчетов (рис.1) модель была обучена на выборке из ~ 200 000 файлов. Выборка была взята из уже существующего структурированного ранее архива предприятия. классификации F1 ~ 0,89. Метрика качества модели классификации.



**Рис. 1. Примеры геологических типов отчетов,
определяемых созданной моделью классификации**

Для решения задачи автоматической индексации была создана база размеченных данных на корпусах 20 800 файлов геологических отчетов нефтегазового общества. Создана NER-модель (Named Entity Recognition) с учетом используемого “БИО-подхода”

(Beginning - начальные, Inner - внутренние, Outer - внешние части наименованной сущности), которая индексирует найденную информацию по структуре: недропользователь, площадь, лицензионный участок, месторождение, куст, скважины. Словарь токенизатора был дополнен терминами и именами собственными без окончаний, например, «Ем-Еговская» => «ем-еговск». Модель предсказывает вероятности 15 классов для каждого токена, с точностью F1 ~ 0,94.

"0": "O", - класс обозначающий токен НЕ целевого класса
"1": "B-FIELD", - класс начала сущности месторождение
"2": "I-FIELD", - класс продолжение сущности месторождение
"3": "B-KUST", - класс начала сущности месторождение
"4": "I-KUST", - класс продолжение сущности месторождение
"5": "B-LICENSE", - класс начала сущности ЛУ
"6": "I-LICENSE", - класс продолжение сущности ЛУ
"7": "B-ORG", - класс начала сущности организация
"8": "I-ORG", - класс продолжение сущности организация
"9": "B-PLAST", - класс начала сущности Пласт
"10": "I-PLAST", - класс продолжение сущности Пласт
"11": "B-WELL", - класс начала сущности скважина
"12": "I-WELL", - класс продолжение сущности скважина
"13": "B-AREA", - класс начала сущности площадь
"14": "I-AREA", - класс продолжения сущности площадь

Рис. 2. Примеры предсказываемых NER-моделью классов

Данные вручную размечались на текстовых файлах определенных обществах групп, месторождениях, участков, скважинах. Что делать если подключить архив с другого общества и другими названиями объектов? Обученная модель про них ничего не знает. В решении помогла идея создания шаблонов (рис. 3). В уже размеченных данных были автоматически замены названия сущностей и произведено обучение модели на преобразованных файлах.

*8a7ded51aac6458b7dd813dc8efc207c.txt – Блокнот

Файл Правка Формат Вид Справка

Таблица

АО "Башнефтегеофизика"

ООО НПЦ "Геостра" отдел ГИРС

Утверждаю:

Главный геолог по промышленной геофизике

ООО НПЦ "Геостра"

Иванов И.А.

"17" февраля 2020 г.

Данные инклинометрии

Месторождение: {field}

Скважина № {well}

Категория скважины {og} добывающая

Заказчик: {area}

Подрядчик: АО БНГФ

Площадь: {area}

Сведения о проведенных исследованиях

Альтитуда стола ротора, м: 139.72 (данные 2020г.)

Магнитное склонение, градус: 13.83 (данные 2020г.)

Диаметр скважины по долоту, м: 0.126

Зарезка ствола в интервале, м: 1133-1136 (по заявке)

Диаметр тех.колонны, м: 0.146

Глубина башмака тех.колонны, м: 1136.00

Текущий забой, м: 1513.60

Примечание: В интервале 10 - 1140 м азимуты взяты с замера ИГН-73 №0509225 от 25/10/19

Интервал, Тип прибора, Характеристика замера

№ количество точек и номер Ствол ЛБТ УБТ

дата исследования номер дата поверки

1 10.0-1500.0 ИММН N 849 откр.ств 0

150 2/17/20

2 10.0-1500.0 ИММН N 768 откр.ств 0

150 2/17/20

Результат обработки

Глубина Зенитный Азимут Азимут Удлине- Смеще- Дирекционный угол смещения, градус

м угол магн. геогр. ние ние

градус градус градус м м

10.00 0.30 308.07 321.90 0.00 0.00 321.90 129.72 10.00 0.00

20.00 0.32 239.68 253.51 0.00 0.10 299.04 119.72 20.00 0.00

30.00 0.41 234.66 248.49 0.00 0.10 276.90 109.72 30.00 0.00

40.00 0.41 231.78 245.61 0.00 0.20 266.13 99.72 40.00 0.00

50.00 0.38 232.26 246.09 0.00 0.30 260.80 89.72 50.00 0.00

60.00 0.30 239.48 253.31 0.00 0.30 258.72 79.72 60.00 -0.10

70.00 0.32 245.98 259.81 0.00 0.40 258.41 69.72 70.00 -0.10

Рис. 3. Пример шаблона отчета для генерации синтетических данных

С применением созданных моделей машинного обучения были проиндексированы миллионы текстовых файлов из корпоративных архивных источников геологических отчетов. Теперь пользователи могут легко и быстро найти и получить нужную информацию из множества архивных файлов контекстным или пространственным поиском.

Созданное решение имеет ряд практических преимуществ для нефтегазовой отрасли, а именно:

- Ускорение доступа к информации: быстрый поиск нужных данных по различным параметрам сокращает время сбора данных для геологических исследований.
- Оптимизация процессов: автоматическая индексация освобождает специалистов для решения более сложных задач, аналитики и стратегического планирования.
- Повышение точности интерпретации: автоматизированная обработка большого объема данных способствует более полному и объективному пониманию геологических особенностей.

- Создание единой цифровой платформы: объединение разнородных архивов в структурированную базу данных с ИИ-поддержкой для комплексного анализа и принятия решений.

Выводы

1. Подходы с применением ИИ к индексации геологических архивов уже сейчас являются эффективными инструментами для автоматизации работы с большими массивами данных в нефтегазовой отрасли, способствуют более оперативной и точной работе специалистов.
2. Перспективы и дальнейшее развитие данной технологии заключается в создании интеллектуальной систем нового уровня, способных помогать экспертам в изучении недр земли, что приведет к новым открытиям и улучшению процесса разведки.

Список литературы

1. Арифуллин А. Перспективы применения генеративного искусственного интеллекта в нефтегазовой отрасли / А. Арифуллин, В. Маркин // Энергетическая политика. – 2024. – No 5(196). – С. 50–59. – DOI 10.46920/2409-5516_2024_5196_50.
2. Ядрышникова О.А., Воропаев Е.Н., Гнилицкий Р.А., Плиткин И.Б. Информационная система «Экспресс-оценка участков нераспределенного фонда недр» // Нефть. Газ. Новации. – 2018. – №12. – С. 47-50