Наша команда

Романов Асхат

Role

Епанов Семён

Role

Сазанов Иван

Role

Наша задача

Преобразование каталога товаров ОАО «РЖД»

цифровой прорыв

сезон: ИИ

Российские железные дороги

Проблематика

Текущее разбиение не достаточно точное для современных задач

Требуется разбить на более точные группы с уникальными свойствами

Dataset + Ideas

New table

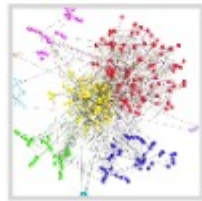
Наименование	Компрессор поршневой
Маркировка	FUBAG B 2800B/100CM 3
Регламенты ГОСТ	NaN
Параметры	2,2КВТ 220В 320Л/МИН 1570 ОБ/МИН
OKPD2_Name	Компрессоры прочие


```
{
 "ГРУППА ТОВАРОВ": "КОМПРЕССОР",
 "МОЩНОСТЬ": "2,2 кВт",
 "НАПРЯЖЕНИЕ": "220 В",
 "ПРОИЗВОДИТЕЛЬНОСТЬ": "320 л/мин",
 "СКОРОСТЬ ВРАЩЕНИЯ": "1570 об/мин",
 "ТИП": "ПОРШНЕВОЙ",
 "ОБЪЕМ РЕЦЕПТИВА": "100 л"
}
```

Мы пробовали

Парсинг
с маркетплейсов

Промптинг
пробовали решения от OpenAI, SberAI, Yandex Research, Anthropic



Кластеризация
на основе ONE, а также доп процессинга, такого как нормирование признаков, взятие моды признаков.
Также, на основе эмбедингов

Технические условия
Находили по ГОСТу и записывали параметры ГОСТа

Плюсы и минусы. Алгоритм:

Точно определяет подгруппу

Хорошо разбивает параметры по столбцам

Могут встретиться доп. специфические признаки из источников

Частая практика - разное название для одной и той же категории; разные ед. изм-я, разные названия характеристик

Работа над минусами

01 Стандартизируем названия специфичных признаков выбираем только из уже существующих названий признаков

02 Делаем предварительный отбор признаков, используя метрики важности слов

Парсер

Для каждого товара извлекает
возможные его характеристики
на Яндекс Маркете

Пример:

Источник
бесперебойного
питания

[выходная мощность,
число выходных
разъемов,
интерфейсы]

LLM

Получает набор характеристик товара и, если они указаны в тексте описания и параметрах, записывает их

Пример:

Источник бесперебойного питания + параметры

[выходная мощность : 480
число выходных разъемов: 2,
интерфейсы: Nan]

Итоговый алгоритм

Собираем данные о товаре (маркетплейсы, ГОСТы, маркировка) Собираем признаки

01

С помощью LLM находим признаки товара, определяем их к существующим признакам

02

Кластеризуем для каждой подкатегории N групп с K уник. свойствами

03

Итоговый алгоритм

Собираем данные о товаре (маркетплейсы, ГОСТы, маркировка) Собираем признаки

01

С помощью LLM находим признаки товара, определяем их к существующим признакам

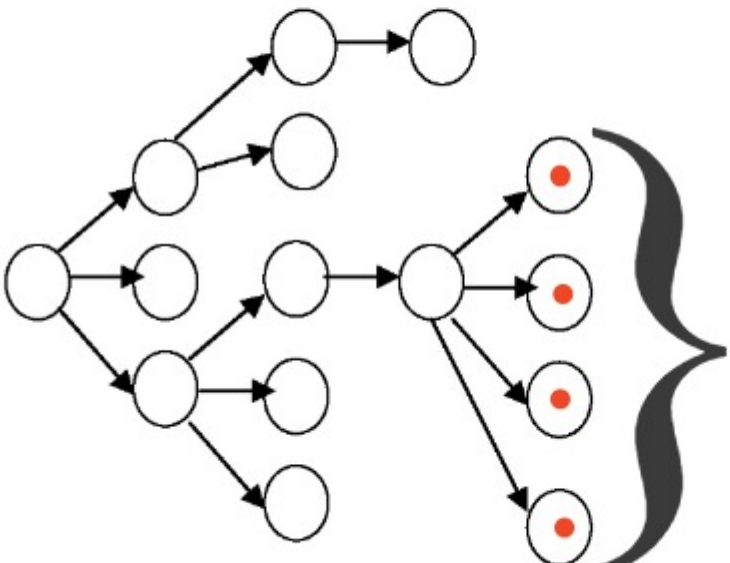
02

Кластеризуем для каждой подкатегории N групп с K уник. свойствами

03

Итоговый вид данных

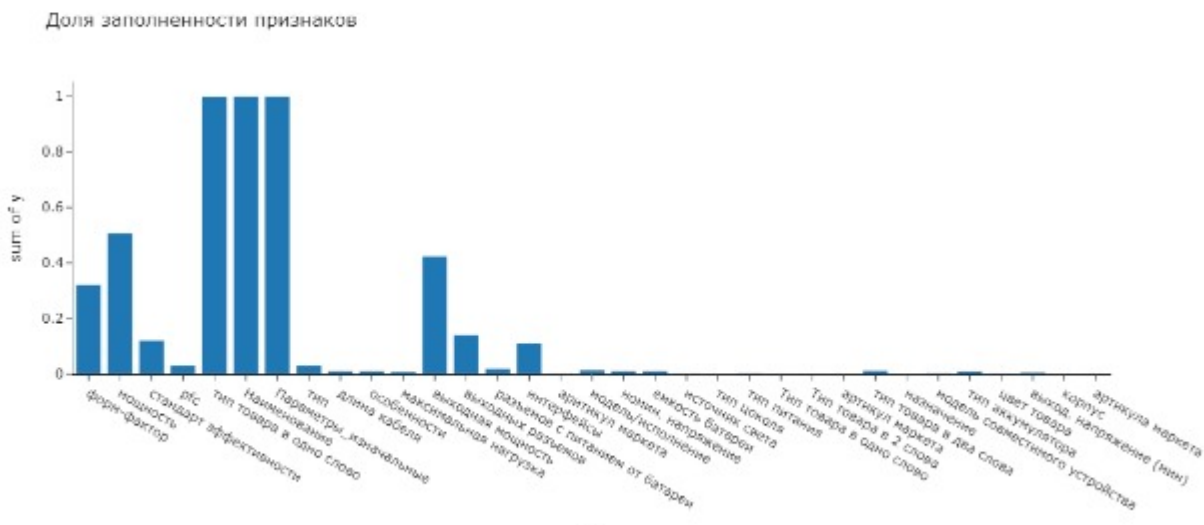
ОКПД



Наши группы, в каждой из которых специфичные для этой группы свойства

Работа со свойствами

После такой разметки появляются свойства, характерные нескольким крупным группам



Выделение групп

Выделяем такие группы, для которых существует большее соответствие по признакам

Названия групп:
Предварительно в LLM получали краткое название объекта, называем группы по самому частому

Улучшения

01 Прогнать через алгоритм весь датасет

02 Давать больше контекста про ОКПД2 для модели

03 Использовать самые новые модели для разметки

04 Использовать парсинг из большего числа источников

05 Повышать скорость инференса

Были бы рады ответить на ваши вопросы

Сазанов Иван

Role

Епанов Семён

Role

Романов Асхат

Role

Время инференса

01 Парсинг строк – 0.2 секунды на товар

02 Разметка от LLM – от 0.1 до 1 секунды на товар (зависит от модели)

03 Дальнейшая работа с группами – обрабатывается сразу