Tender Hack Search Engine

RUT4

Максим Герасимов,
Штейнман Александр,
Игорь Васильев,
Цыганок Андрей

Задача

Проблемы поиска по базе данных:

- Отсутствие семантического поиска (поиск по смыслу). Незнание пользователем номеров код КПГЗ, точных названий.
- Отсутствие фичей: автодополнения, транслитерации, учета синонимов, других языков
- Отход от правил, поиск универсального метода, масштабируемость

Особенности данных

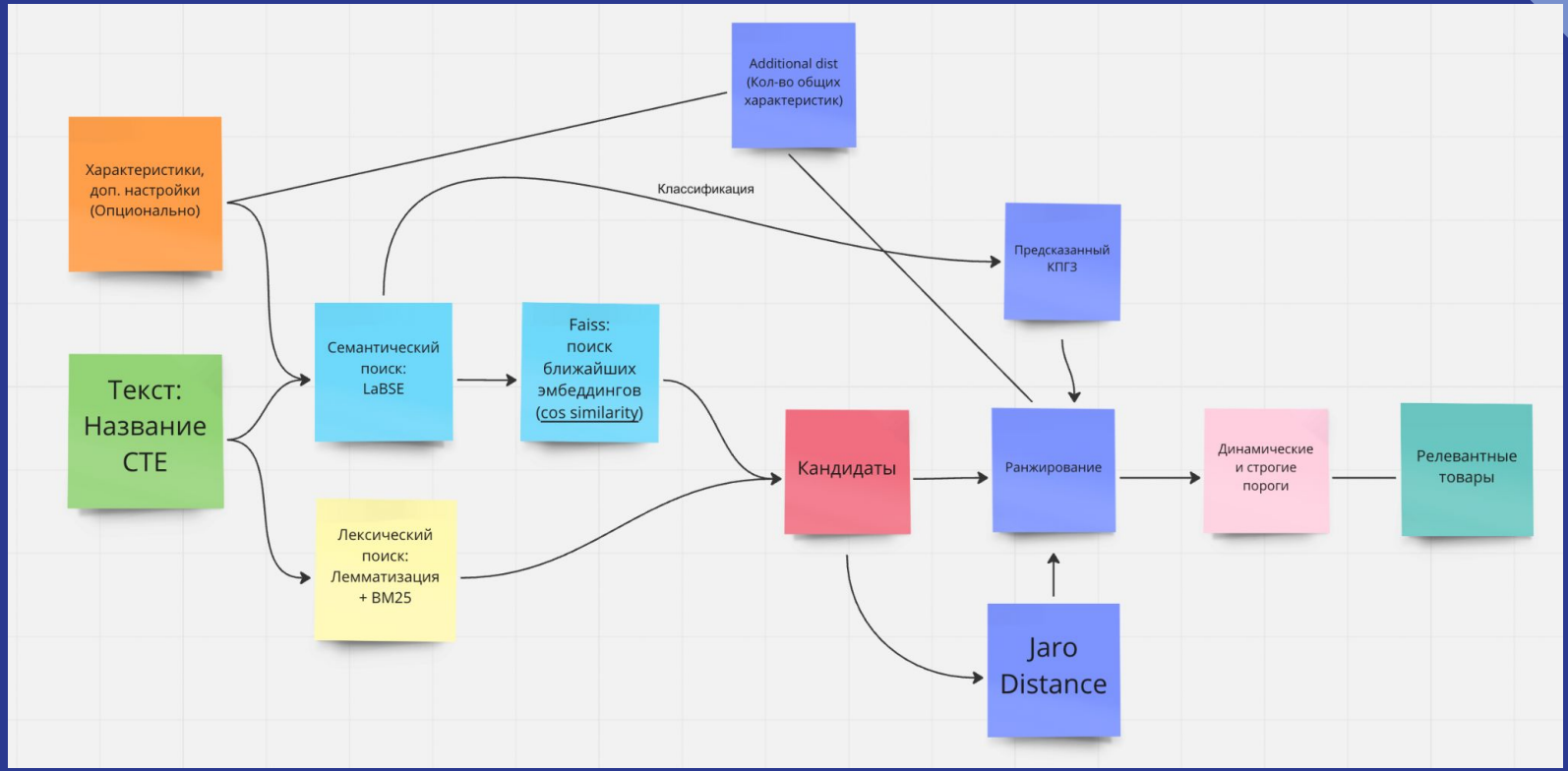
1. Частично структурированные данные (не унифицированные, нет структуры в характеристиках)

1. Множество разреженных категорий (недостаточно данных, чтобы классифицировать по нижним уровням КПГЗ с помощью правил)

Реализованный функционал продукта

1. Семантический поиск (поиск по смыслу).
2. Частичная мультиязычность (ru-en)
3. Лексический поиск (лемматизация)
4. Поиск с учетом характеристик
5. Поиск сопутствующих товаров
6. Автодополнение
7. Автокоррекция
8. Транслитерация
9. Анализ рынка
10. Масштабируемость

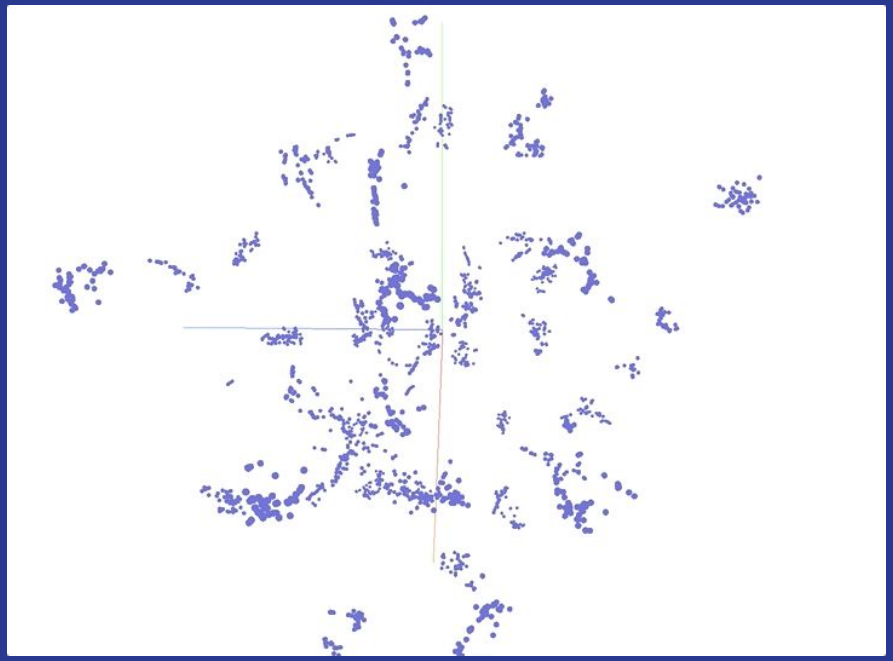
Как мы решаем задачу поиска

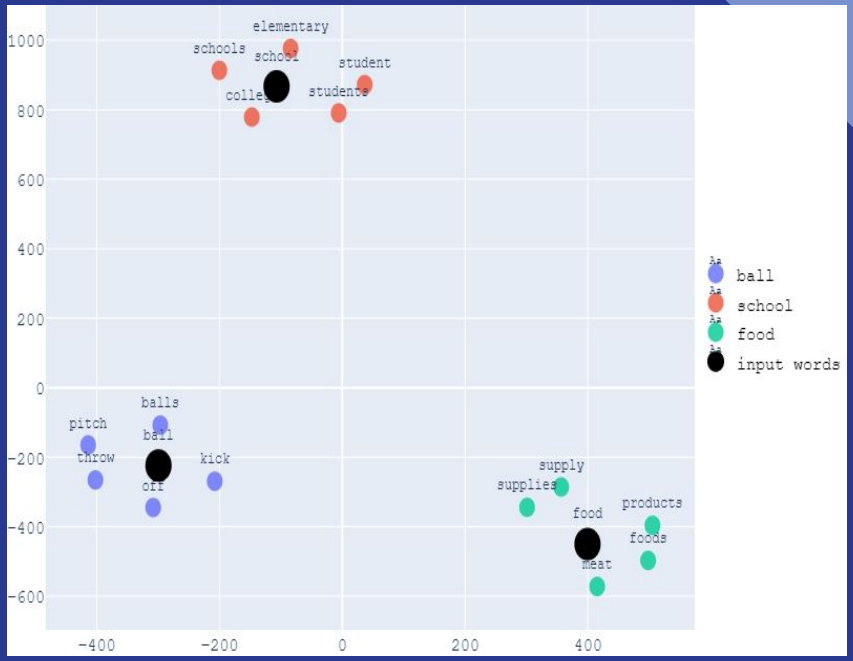


Эмбеддинги

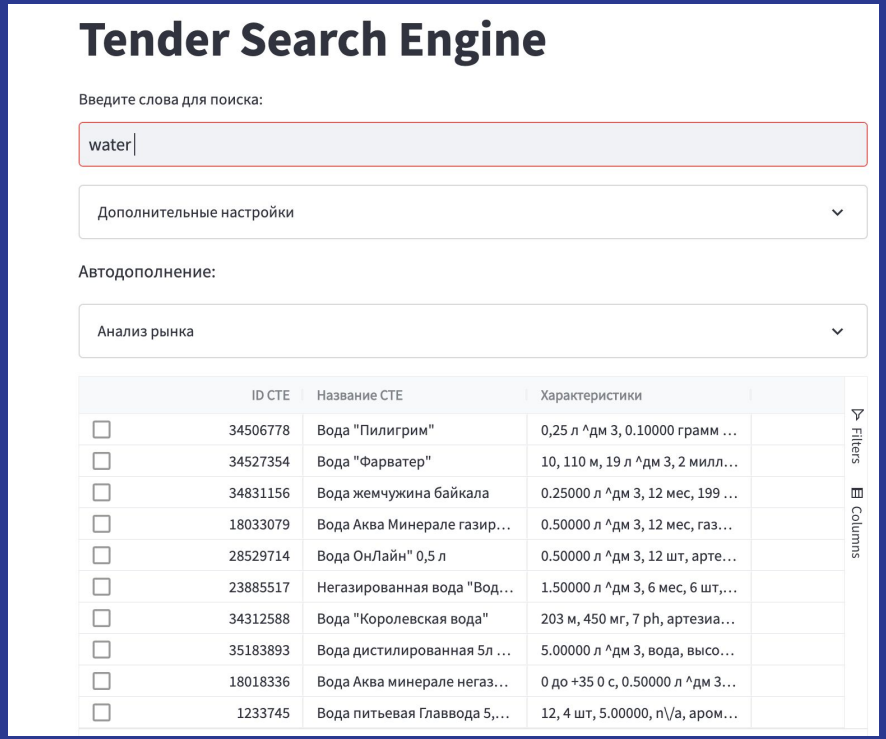
1. Близкие по смыслу товары располагаются рядом в кластерах
2. Кластера связаны с КПГЗ по смыслу
3. Мультиязычные (LaBSE) -> “Samsung” и “самсунг” близки по смыслу

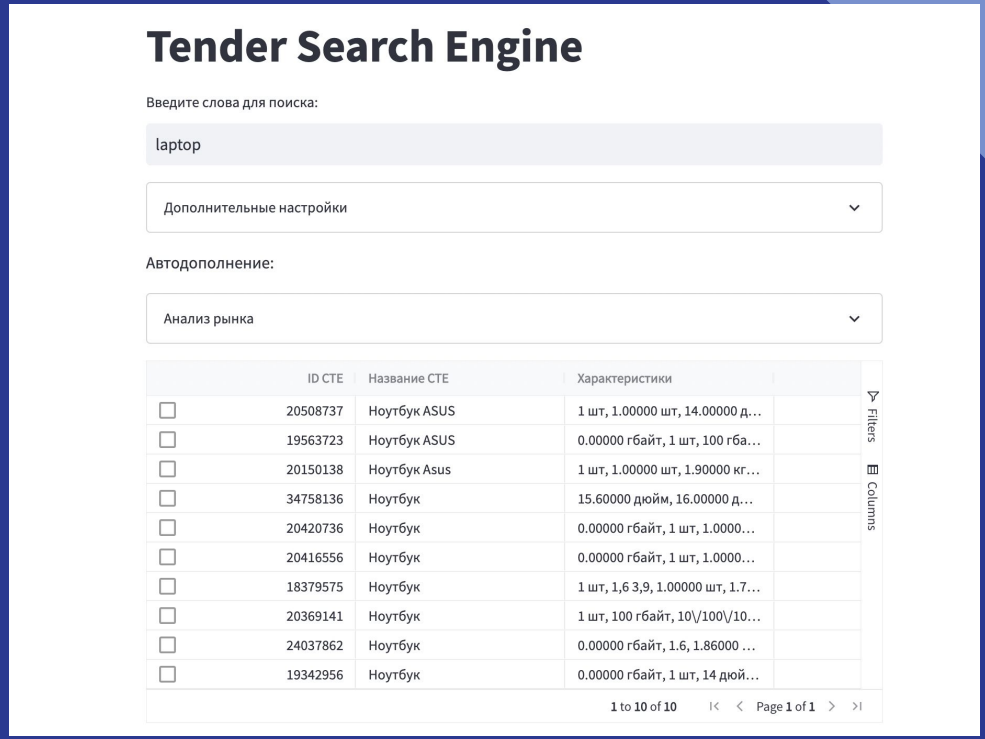
Эмбеддинги




elementary
schools
school
student
college
students
balls
pitch
ball
throw
kick
off
supply
supplies
products
food
foods
meat
1000
800
600
400
200
0
-200
-400
-600
-400
-200
0
200
400
ball
school
food
input words

Что дает LaBSE





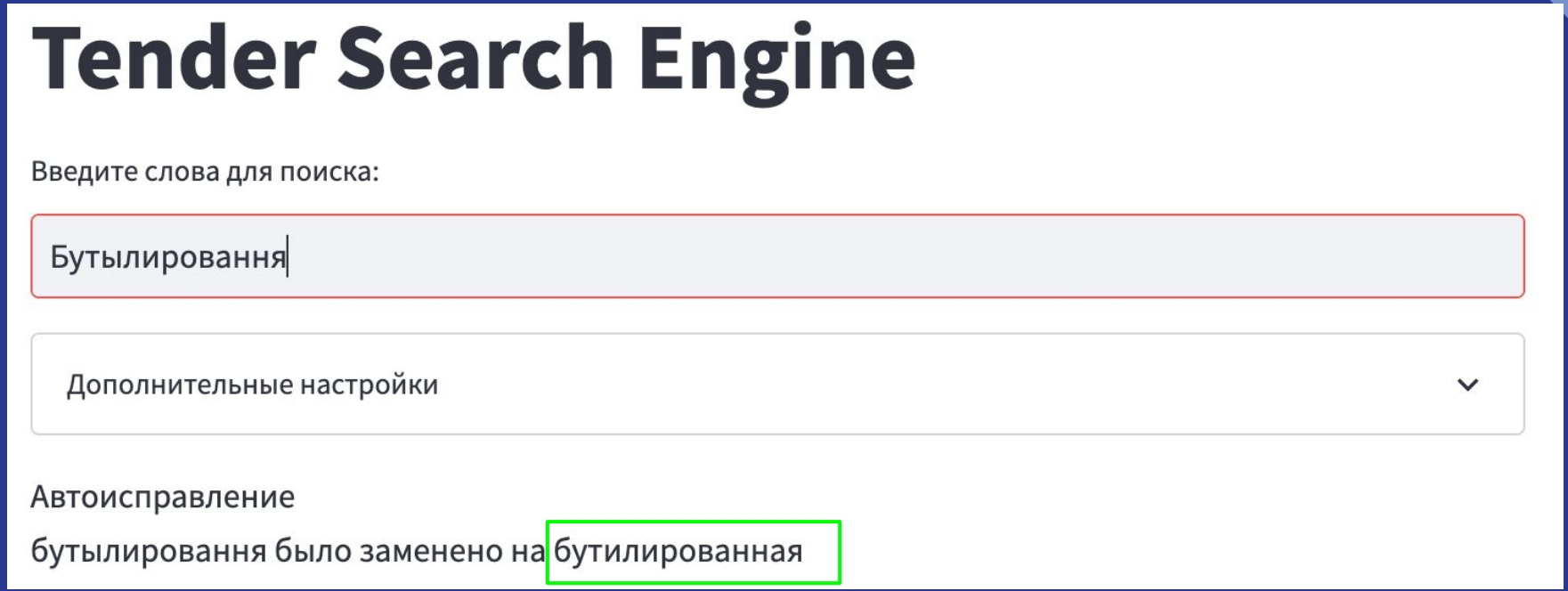
Сопутствующие товары

Идея:

1. Выбираем продукт, для которого хотим найти сопутствующие товары (Example: КПГЗ: 1.2.3.4)
2. Выбираем похожие категории, отличные от КПГЗ продукта (НЕ 1.2.3.4)
   Метрика: максимальная последовательная схожесть КПГЗ:
   (Example: 1.2.5.4 -> 2 последовательных совпадения, 2.2.5.4 -> 0 совпадений)
3. Из пула с наибольшим последовательным совпадением (Example: 1.2.3.5, 1.2.3.7 и тд) выбираем топ-100 с наибольшим косинусным расстоянием (ближайшие в семантическом смысле)

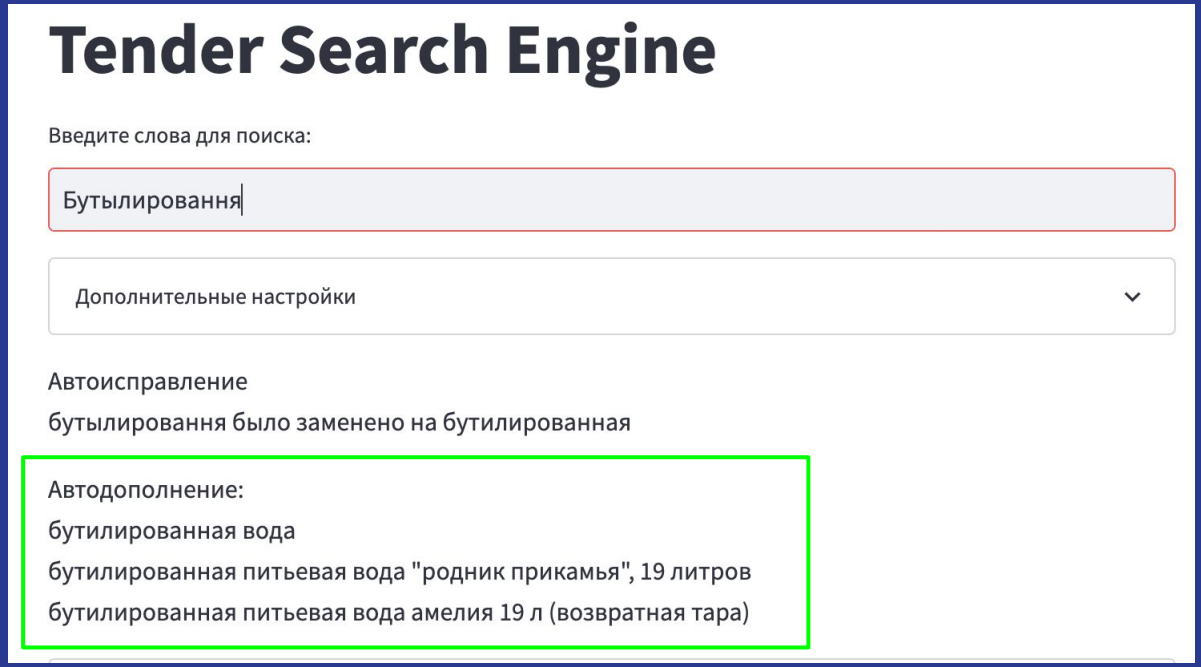
Фичи (Автоисправление)

- Jamspell



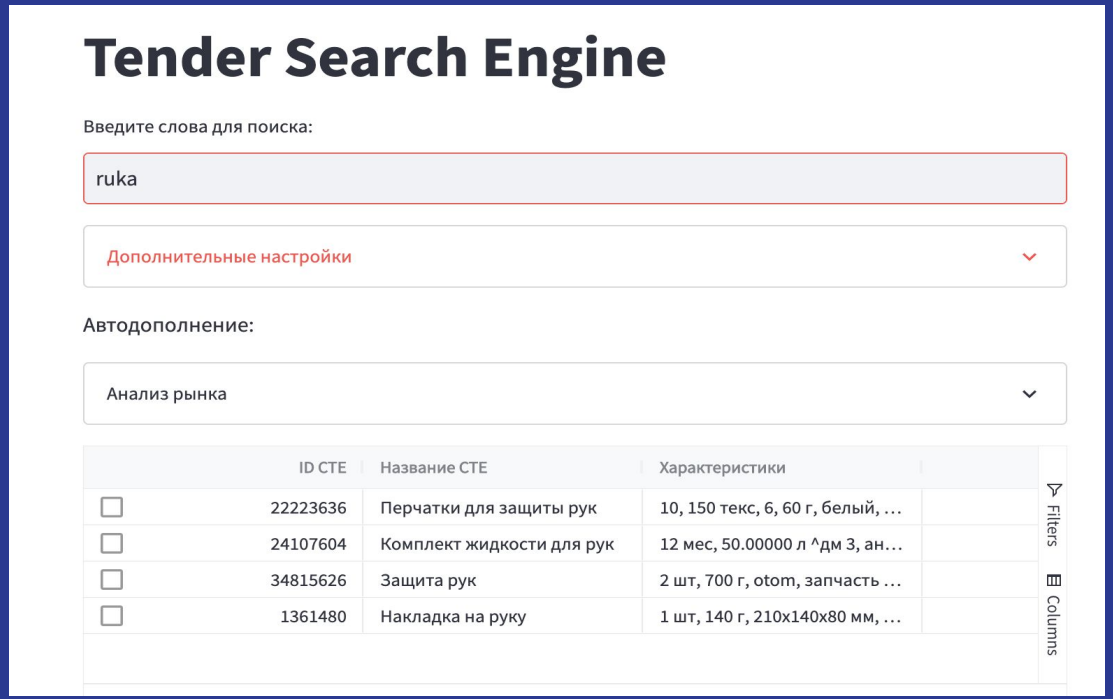
Фичи (Автодополнение)

- Частота + startwith



Фичи (Транслитерация)

- Если при поиске ничего не вывелось -> делаем транслитерацию



Метрики

Классификация категорий КПГЗ на 5307 классов:

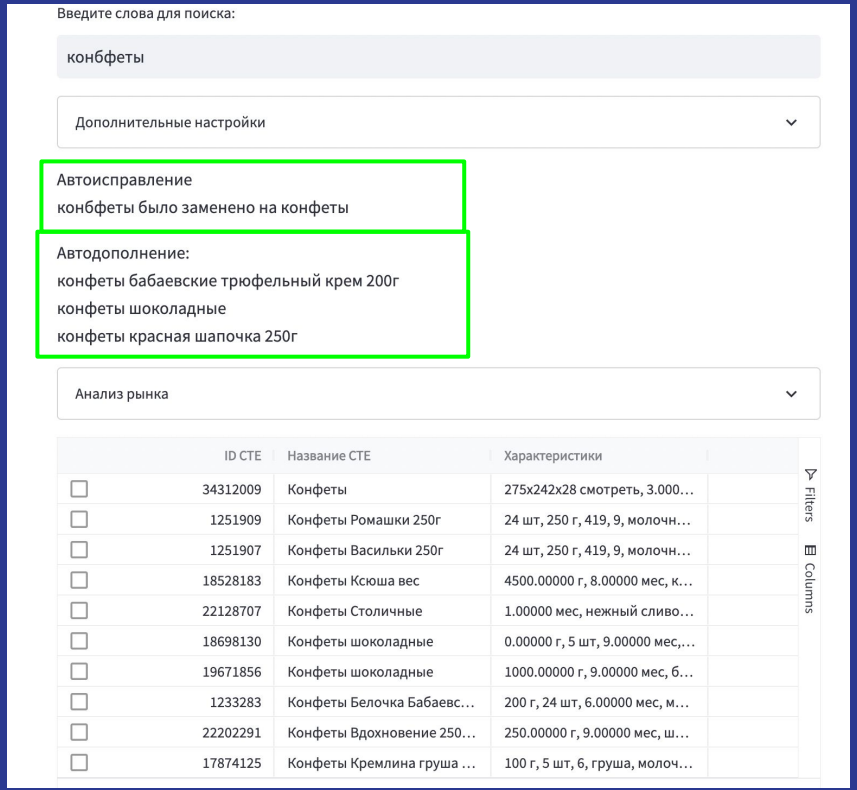
- Accuracy: 0.7 (Лемматизация)
- F1_weighted: 0.71 (Лемматизация)

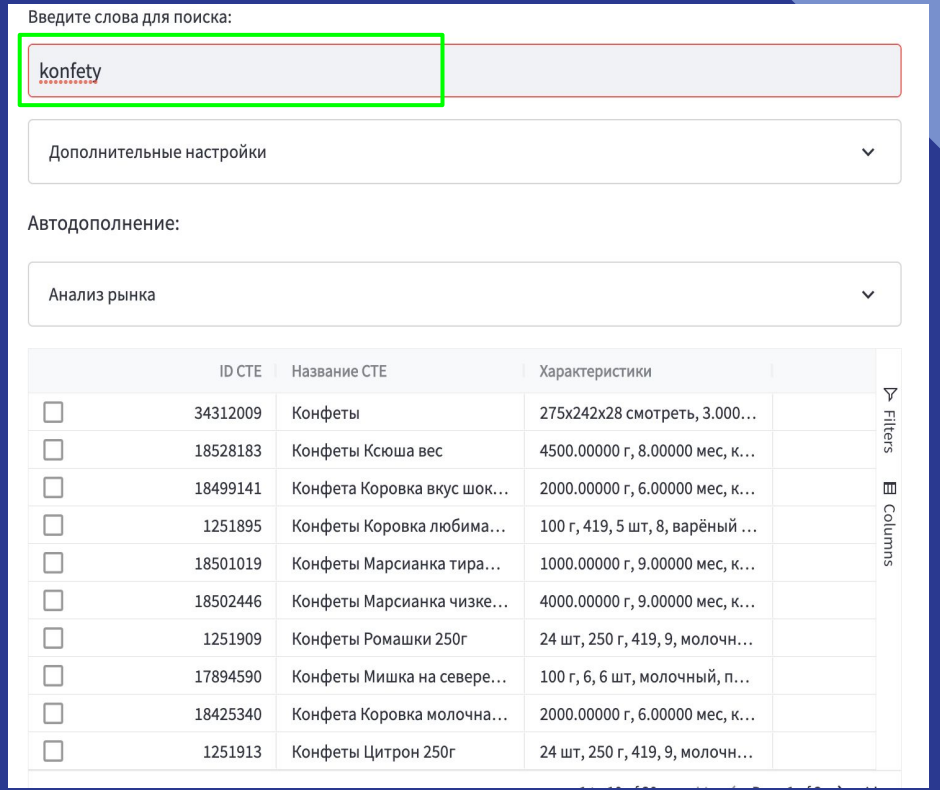
Также, была модель с метриками 0.9, но мы отказались от нее, т.к эмбеддинги были хуже

Масштабируемость

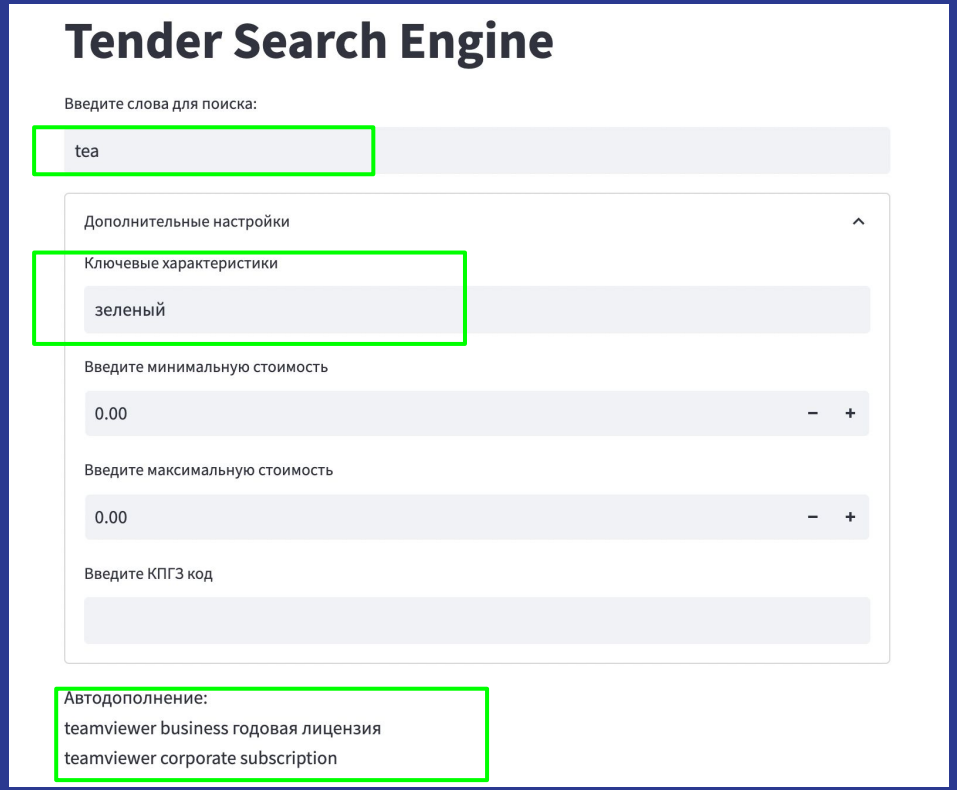
1. Модель умеет обобщать за счет эмбеддингов, работает с тем, чего не было обучающей выборке, в.т. ч с английским языком
2. Легко и быстро обучать, не нужна предобработка
3. Устойчива к редким позициям за счет лексического поиска

Пример работы системы



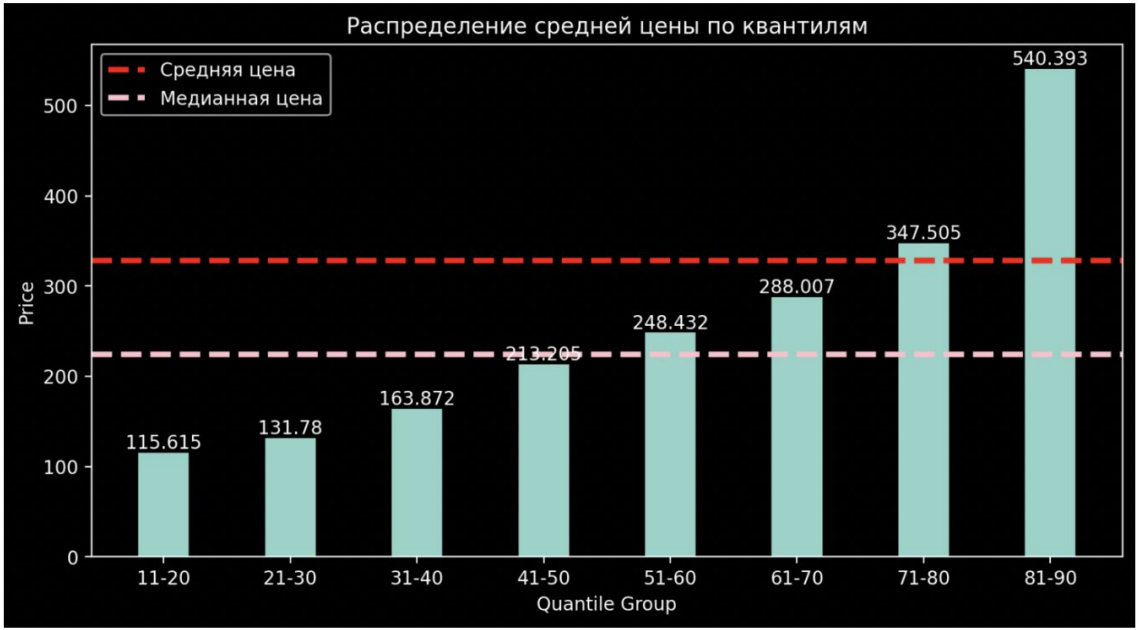


Пример работы системы



Пример работы системы

Анализ рынка



Пример работы системы

	ID CTE	Название CTE	Характеристики
☐	1206640	Чай AHMAD (Ахмад) "Jasmi…	100, 230 г, 24, ahmad tea, зелёный, картонный
☐	23126514	Чай принцесса ява	100 шт, 2.00000 г, 200 г, 36 мес, бакалея, зелён
☐	20749411	Чай Greenfield Sweet Jasmi…	100 г, 24 мес, бакалея, в набор, жестяной бан
☐	1249538	Чай Greenfield коллекция п…	120 шт, 218, 24 мес, 240 г, 8 шт, ассорти, карто
☐	1250674	Чай Ahmad Tea Китайский …	100 шт, 180 г, 36 мес, 8 шт, 870, зелёный, карт
☐	21295240	Чай MAITRE (Мэтр) "de The …	130 г, 24 мес, 60 шт, для весь, наличие подарс
☐	1206828	Чай AHMAD (Ахмад) "Chines…	300, 36 мес, 545 г, ahmad tea, зелёный, картон
☐	35008175	Чай	100.00000 г, 3.00000 мес, бакалея, индийский,
☐	18434026	Чай в пакетиках GREENFIEL…	100 шт, 2.00000 г, 200 г, бакалея, наличие кон
☐	34526007	Чай GREENFIELD	10.00000 мес, 12 мес, 25.00000 шт, 50.00000 г,

Filters Columns

1 to 10 of 30 Page 1 of 3

Пример работы системы

Полная информация

Название: Чай AHMAD (Ахмад) "Jasmine Green Tea", зелёный с жасмином, 100 пакетиков по 2 г

Категория: Чай черный (ферментированный)

Код КПГЗ: 01.01.01.19.01.04

Цена: 303.235 ₽

Характеристики:
100, 230 г, 24, ahmad tea, зелёный, картонный упаковка, классический, наличие конвертик, наличие ярлычок, цветочный

Пример работы системы

Сопутствующие товары

ID CTE	Название CTE	Характеристики
1254533	Торт Коровка вафельный ассорти топленное моло…	6 мес, 684, 9 шт, вафе
1351021	Кофе NORR Morkrost сублимированный 75 г, мягка…	100 арабика, 24 мес, 2
17914273	Зефир Лянеж ванильный 315г	12 шт, 315 г, ваниль, г
17953981	Консервация Персики Corrado отборные половин…	24 шт, греция, жестян
17968280	Какао Вкусвилл напиток растворимый, 375г	12 шт, 24, 375 г, пласт
34636793	Пряники имбирные	1.00000 мес, 300.0000

Filters Columns