

# 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

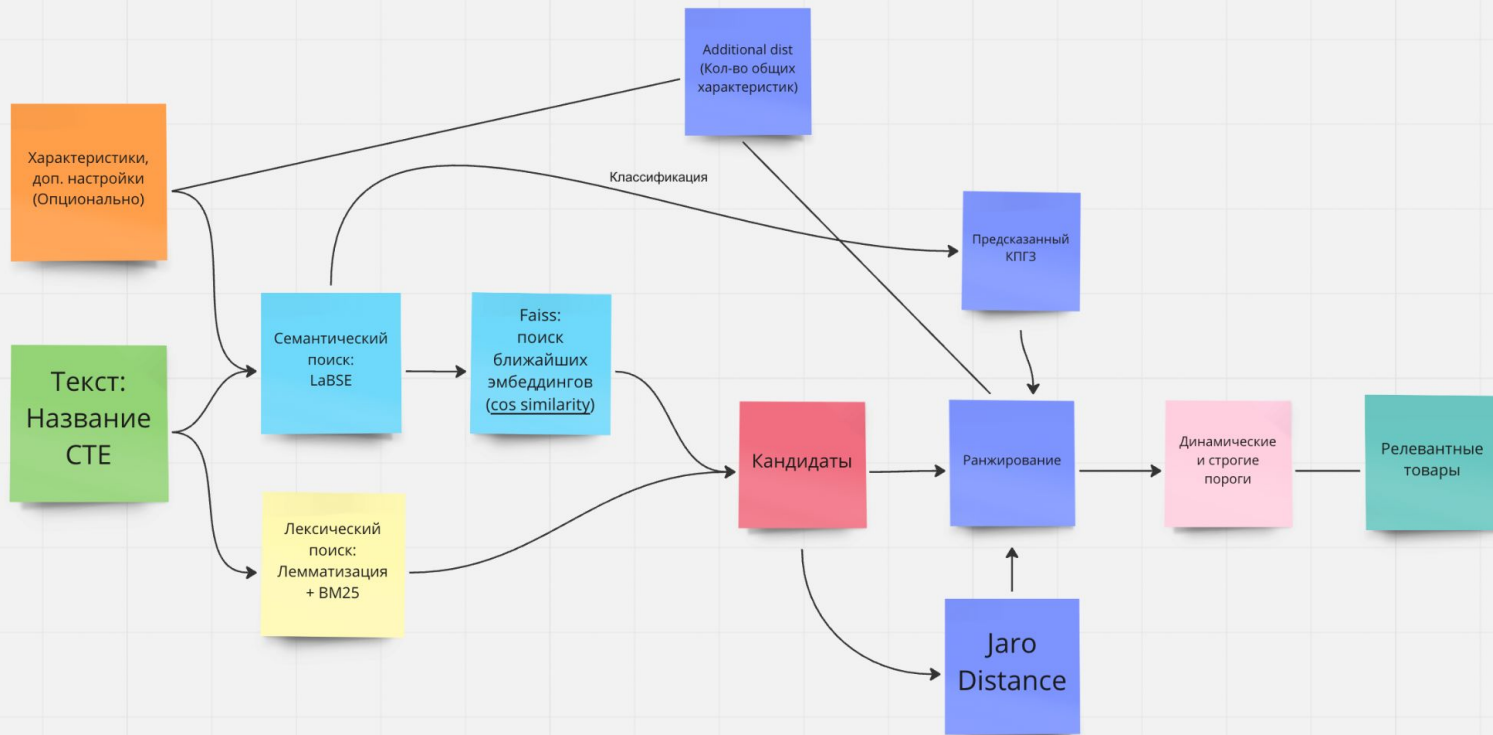
# Особенности данных

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 мультязычность (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

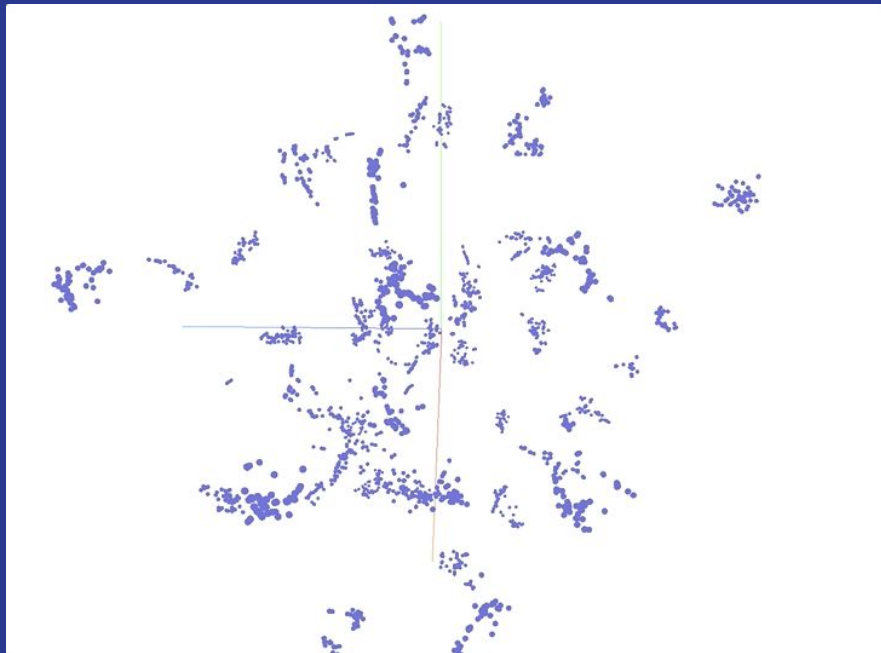
# Как мы решаем задачу поиска



# Эмбединги

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

# Эмбединги



# Что дает LaBSE

## Tender Search Engine

Введите слова для поиска:

water

Дополнительные настройки

Автодополнение:

Анализ рынка

	ID CTE	Название CTE	Характеристики	
<input type="checkbox"/>	34506778	Вода "Пилигрим"	0,25 л ^дм 3, 0.10000 грамм ...	Filters Columns
<input type="checkbox"/>	34527354	Вода "Фарватер"	10, 110 м, 19 л ^дм 3, 2 милл...	
<input type="checkbox"/>	34831156	Вода жемчужина байкала	0.25000 л ^дм 3, 12 мес, 199 ...	
<input type="checkbox"/>	18033079	Вода Аква Минерале газир...	0.50000 л ^дм 3, 12 мес, газ...	
<input type="checkbox"/>	28529714	Вода Онлайн" 0,5 л	0.50000 л ^дм 3, 12 шт, арте...	
<input type="checkbox"/>	23885517	Негазированная вода "Вод...	1.50000 л ^дм 3, 6 мес, 6 шт,...	
<input type="checkbox"/>	34312588	Вода "Королевская вода"	203 м, 450 мг, 7 рh, артезиа...	
<input type="checkbox"/>	35183893	Вода дистиллированная 5л ...	5.00000 л ^дм 3, вода, высо...	
<input type="checkbox"/>	18018336	Вода Аква минерале негаз...	0 до +35 0 с, 0.50000 л ^дм 3...	
<input type="checkbox"/>	1233745	Вода питьевая Главвода 5,...	12, 4 шт, 5.00000, n/a, аром...	

## Tender Search Engine

Введите слова для поиска:

laptop

Дополнительные настройки

Автодополнение:

Анализ рынка

	ID CTE	Название CTE	Характеристики	
<input type="checkbox"/>	20508737	Ноутбук ASUS	1 шт, 1.00000 шт, 14.00000 д...	Filters Columns
<input type="checkbox"/>	19563723	Ноутбук ASUS	0.00000 гбайт, 1 шт, 100 гба...	
<input type="checkbox"/>	20150138	Ноутбук Asus	1 шт, 1.00000 шт, 1.90000 кг...	
<input type="checkbox"/>	34758136	Ноутбук	15.60000 дюйм, 16.00000 д...	
<input type="checkbox"/>	20420736	Ноутбук	0.00000 гбайт, 1 шт, 1.0000...	
<input type="checkbox"/>	20416556	Ноутбук	0.00000 гбайт, 1 шт, 1.0000...	
<input type="checkbox"/>	18379575	Ноутбук	1 шт, 1,6 3,9, 1.00000 шт, 1.7...	
<input type="checkbox"/>	20369141	Ноутбук	1 шт, 100 гбайт, 10\100\10...	
<input type="checkbox"/>	24037862	Ноутбук	0.00000 гбайт, 1,6, 1.86000 ...	
<input type="checkbox"/>	19342956	Ноутбук	0.00000 гбайт, 1 шт, 14 дюй...	

1 to 10 of 10

< < Page 1 of 1 > >



# 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

## Tender Search Engine

Введите слова для поиска:

Дополнительные настройки



Автоисправление

бутилированн| было заменено на бутилированная

# Фичи (Автодополнение)

- Частота + startwith

## Tender Search Engine

Введите слова для поиска:

Дополнительные настройки

Автоисправление  
бутилирования было заменено на бутилированная

Автодополнение:  
бутилированная вода  
бутилированная питьевая вода "родник прикамья", 19 литров  
бутилированная питьевая вода амелия 19 л (возвратная тара)

# Фичи (Транслитерация)

- Если при поиске ничего не вывелось -> делаем транслитерацию

## Tender Search Engine

Введите слова для поиска:

ruka

Дополнительные настройки

Автодополнение:

Анализ рынка

	ID CTE	Название CTE	Характеристики	
<input type="checkbox"/>	22223636	Перчатки для защиты рук	10, 150 текс, 6, 60 г, белый, ...	
<input type="checkbox"/>	24107604	Комплект жидкости для рук	12 мес, 50.00000 л ^дм 3, ан...	
<input type="checkbox"/>	34815626	Защита рук	2 шт, 700 г, otom, запчасть ...	
<input type="checkbox"/>	1361480	Накладка на руку	1 шт, 140 г, 210x140x80 мм, ...	

Filters  
Columns

# 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 эмбединги были хуже

# Масштабируемость

1. Модель умеет обобщать за счет эмбедингов, 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

Введите слова для поиска:

конфеты

Дополнительные настройки

Автоисправление  
конфеты было заменено на конфеты

Автодополнение:  
конфеты бабаевские трюфельный крем 200г  
конфеты шоколадные  
конфеты красная шапочка 250г

Анализ рынка

	ID CTE	Название CTE	Характеристики
<input type="checkbox"/>	34312009	Конфеты	275x242x28 смотреть, 3.000...
<input type="checkbox"/>	1251909	Конфеты Ромашки 250г	24 шт, 250 г, 419, 9, молочн...
<input type="checkbox"/>	1251907	Конфеты Васильки 250г	24 шт, 250 г, 419, 9, молочн...
<input type="checkbox"/>	18528183	Конфеты Ксюша вес	4500.00000 г, 8.00000 мес, к...
<input type="checkbox"/>	22128707	Конфеты Столичные	1.00000 мес, нежный сливо...
<input type="checkbox"/>	18698130	Конфеты шоколадные	0.00000 г, 5 шт, 9.00000 мес,...
<input type="checkbox"/>	19671856	Конфеты шоколадные	1000.00000 г, 9.00000 мес, б...
<input type="checkbox"/>	1233283	Конфеты Белочка Бабаевс...	200 г, 24 шт, 6.00000 мес, м...
<input type="checkbox"/>	22202291	Конфеты Вдохновение 250...	250.00000 г, 9.00000 мес, ш...
<input type="checkbox"/>	17874125	Конфеты Кремлина груша ...	100 г, 5 шт, 6, груша, молоч...

Введите слова для поиска:

konfety

Дополнительные настройки

Автодополнение:  
Анализ рынка

	ID CTE	Название CTE	Характеристики
<input type="checkbox"/>	34312009	Конфеты	275x242x28 смотреть, 3.000...
<input type="checkbox"/>	18528183	Конфеты Ксюша вес	4500.00000 г, 8.00000 мес, к...
<input type="checkbox"/>	18499141	Конфета Коровка вкус шок...	2000.00000 г, 6.00000 мес, к...
<input type="checkbox"/>	1251895	Конфеты Коровка любима...	100 г, 419, 5 шт, 8, варёный ...
<input type="checkbox"/>	18501019	Конфеты Марсианка тира...	1000.00000 г, 9.00000 мес, к...
<input type="checkbox"/>	18502446	Конфеты Марсианка чизке...	4000.00000 г, 9.00000 мес, к...
<input type="checkbox"/>	1251909	Конфеты Ромашки 250г	24 шт, 250 г, 419, 9, молочн...
<input type="checkbox"/>	17894590	Конфеты Мишка на севере...	100 г, 6, 6 шт, молочный, п...
<input type="checkbox"/>	18425340	Конфета Коровка молочна...	2000.00000 г, 6.00000 мес, к...
<input type="checkbox"/>	1251913	Конфеты Цитрон 250г	24 шт, 250 г, 419, 9, молочн...

# Пример работы системы

## Tender Search Engine

Введите слова для поиска:

tea

Дополнительные настройки

Ключевые характеристики

зеленый

Введите минимальную стоимость

0.00

- +

Введите максимальную стоимость

0.00

- +

Введите КППЗ код

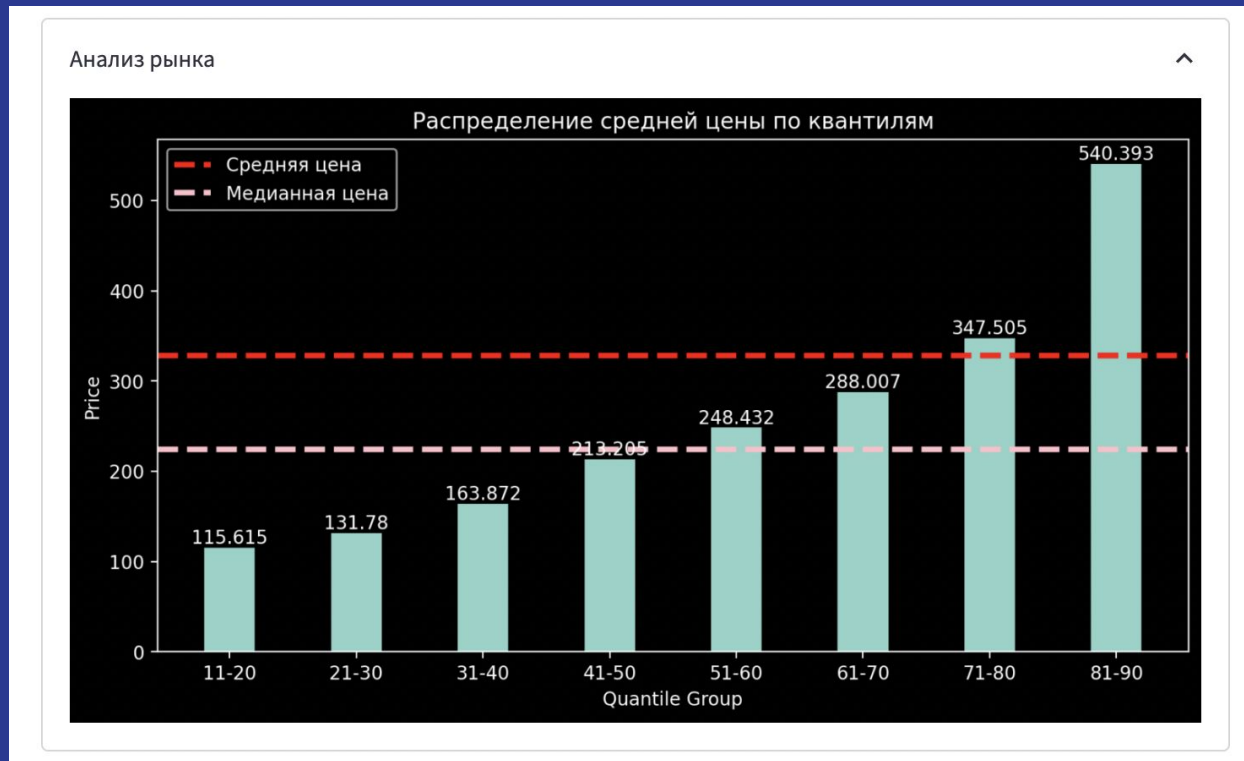
Автодополнение:

teamviewer business годовая лицензия

teamviewer corporate subscription



# Пример работы системы



# Пример работы системы

	ID CTE	Название CTE	Характеристики	Filters
<input type="checkbox"/>	1206640	Чай AHMAD (Ахмад) "Jasmi...	100, 230 г, 24, ahmad tea, зелёный, картонный	Columns
<input type="checkbox"/>	23126514	Чай принцесса ява	100 шт, 2.00000 г, 200 г, 36 мес, бакалея, зелёный	
<input type="checkbox"/>	20749411	Чай Greenfield Sweet Jasmi...	100 г, 24 мес, бакалея, в набор, жестяной бан	
<input type="checkbox"/>	1249538	Чай Greenfield коллекция п...	120 шт, 218, 24 мес, 240 г, 8 шт, ассорти, карто	
<input type="checkbox"/>	1250674	Чай Ahmad Tea Китайский ...	100 шт, 180 г, 36 мес, 8 шт, 870, зелёный, карто	
<input type="checkbox"/>	21295240	Чай MAITRE (Мэтр) "de The ...	130 г, 24 мес, 60 шт, для весь, наличие подарс	
<input type="checkbox"/>	1206828	Чай AHMAD (Ахмад) "Chines...	300, 36 мес, 545 г, ahmad tea, зелёный, картон	
<input type="checkbox"/>	35008175	Чай	100.00000 г, 3.00000 мес, бакалея, индийский,	
<input type="checkbox"/>	18434026	Чай в пакетиках GREENFIEL...	100 шт, 2.00000 г, 200 г, бакалея, наличие кон	
<input type="checkbox"/>	34526007	Чай GREENFIELD	10.00000 мес, 12 мес, 25.00000 шт, 50.00000 г, (	
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	Filters Columns
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, 300.00000	