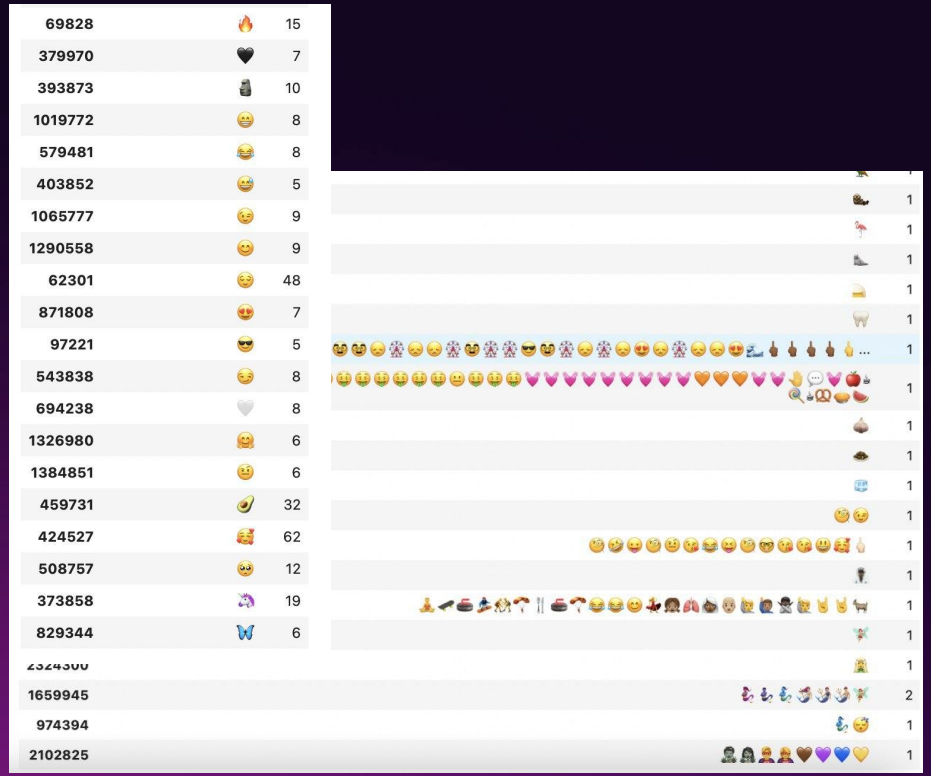
Wildhack

Поисковые теги

Решение представлено
командой Rabies

Предобработка данных

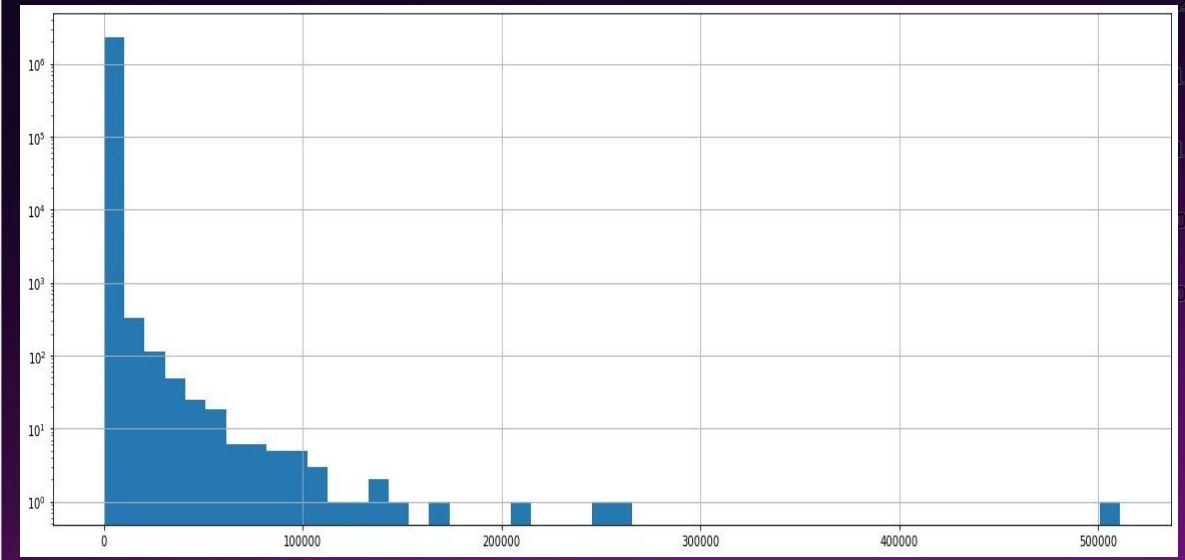
- Удаление пропусков в поисковых запросах
- Удаление единичных запросов
- Очистка небуквенных данных
- Коррекция порядка слов
- Приведение слов к нормальной форме



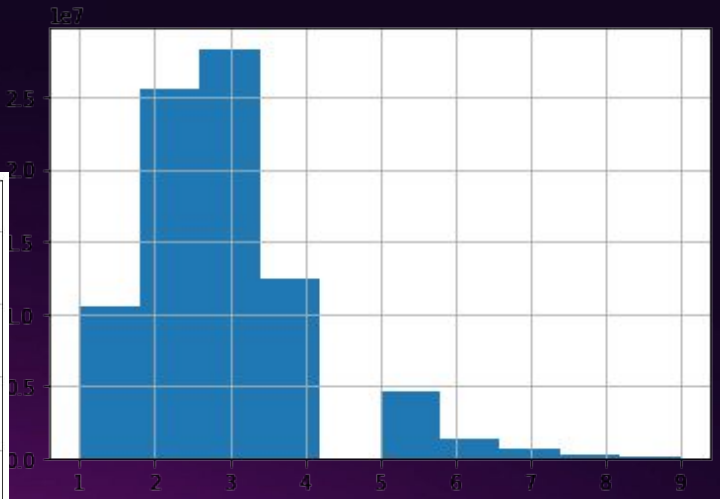
69828	🔥	15
379970	🖤	7
393873	🤳	10
1019772	😁	8
579481	😂	8
403852	😅	5
1065777	😉	9
1290558	😊	9
62301	😌	48
871808	😍	7
97221	😎	5
543838	😏	8
694238	🤍	8
1326980	🤗	6
1384851	🙂	6
459731	🥑	32
424527	🥰	62
508757	🥺	12
373858	🦄	19
829344	🦋	6
1659945		2
974394		1
2102825		1

Wildhack

Предобработка данных



Частотность употребления слов



Количество слов в запросе

Wildhack

Алгоритм

1. Подготовка списка **всевозможных потенциальных тегов** по историческим данным: словосочетания из 2 или 3 слов.
2. Выделение **ключевых слов**. Обрезание по 0.95 квантили.
3. Выдача **релевантных тегов** по новому запросу.
   a. Поиск для запроса ключевых слов по метрике: 1 ключ для 1 слова запроса

$$isKeyWord = 0.01 * similarityScore * \log UQCount$$

   b. В зависимости от величины запроса определяется количество теговых слов по новому запросу - наиболее часто используемые слова совместно с вычисленными ключевыми
   c. Для формирования запроса вычисляется декартово произведение теговых слов ко всем словам запроса
   d. Для получившегося множества запросов ищутся теги из полученных исторических словосочетаний с ранжированием по метрике, где t - степень похожести

$$metric = t * currentWeight * \log\log UQCount$$

Wildhack

Применение ML

- Решение представляет из себя физическую модель, которую можно будет заранее обучить на данных
- В решении использованы библиотеки pymorphy2 и nltk основанные на задачах nlp

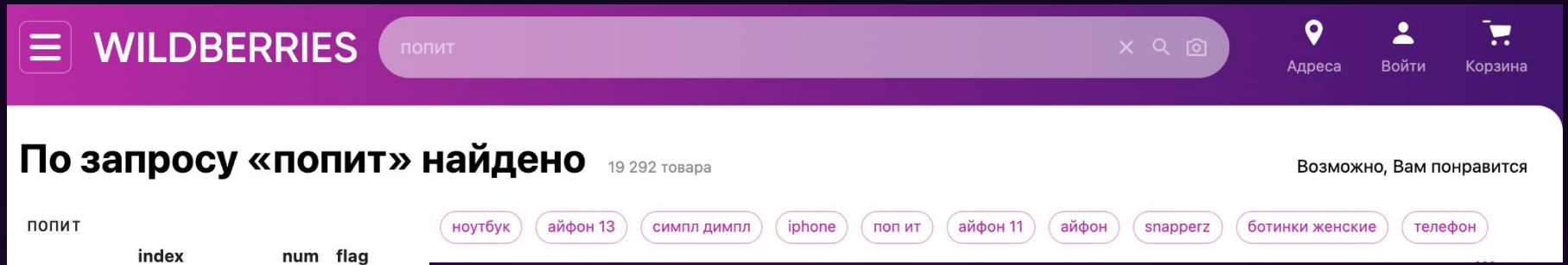
Wildhack

Достоинства решения

- Решение воспринимает любые запросы, даже если они новые
- Список тегов динамически меняется, может свободно расширяться или сужаться
- Опечатки и язык не важны - все равно найдутся теги!
- В среднем 7-8 тегов из 10 уточняют запрос пользователя и помогают сориентироваться в поиске
- Решение использует простые и логичные метрики, которые легко можно подстроить под новые данные

Wildhack

Улучшения

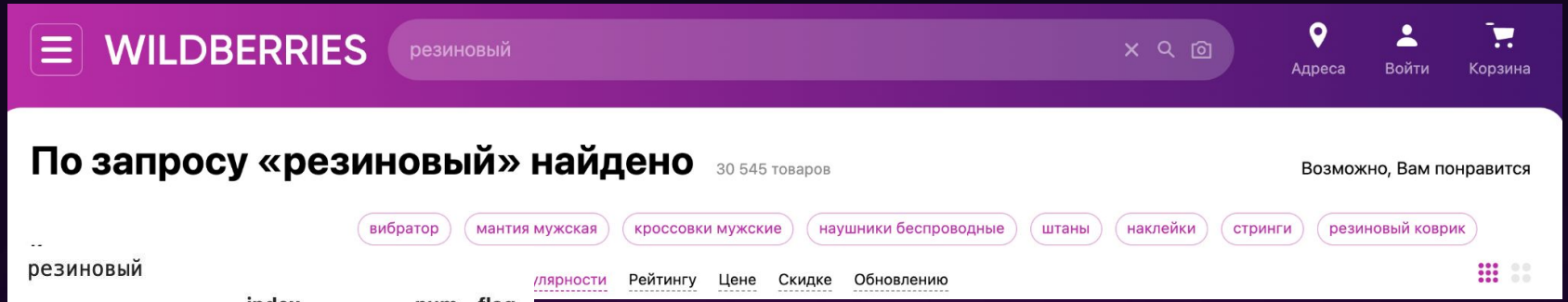


попит

	index	num	flag
10	поп туб	10283.698304	0
2	поп ыт	9799.914312	0
11	вода питьевая	9540.075335	0
0	ткань поплин	9300.444059	0
8	чехол попит	9218.689948	0
6	поп тубс	8614.385002	0
9	купить коврик	8548.305985	0
4	пи пи	8157.217628	0
7	купить уф	8116.793612	0
5	поп сокет	7997.210378	0

Wildhack

Улучшения



резиновый

	index	num	flag
79	резиновые сапоги	14379.368984	0
75	резиновый коврик	13701.231564	0
39	мяч резиновый	12977.217794	0
78	новый год	11838.344102	0
23	резиновый член	11479.391945	0
72	резиновые тапочки	11115.470168	0
77	протеиновый батончик	10944.762890	0
74	перчатки резиновые	10916.375576	0
64	резиновые ботинки	10798.619436	0
45	тапки резиновые	10742.661235	0

Wildhack

Использованный стек

- Pymorphy2 и nltk для работы с лексикой
- Vaex для работы с большими данными
- FuzzyWuzzy для поиска расстояний между слов
- Pandas, numpy, re и прочее для обработки данных

Wildhack

Специфика применения

Время формирования списка тегов ~ 10-15 мин

Время вычисления результата ~ 5-15 сек

Данные для хранения - база тегов, словарь запросов

Данные для обновления - теги, словарь, данные новых запросов

Wildhack

Как накатить в продакшн?

- Провести предрасчет похожести для словаря ключевых слов, примерно 120000 значений
- Сделать матрицу похожести, к которой можно обращаться в лайв-режиме - это обеспечит покрытие не менее 95% запросов
- Обогащать новыми данными базу знаний и дополнять словарь ключевых слов

Wildhack

Дальнейшее развитие

Очистка ошибок транслитерации

Очистка несостоятельных запросов

Включение более сложных запросов

Оптимизация словаря запросов

Оптимизация времени вычисления

Подстройка выдаваемых тегов под поведение пользователя в последнее время (рекомендательная система)

Wildhack

Бонус: что изменить в сборе данных

- Автоматическая очистка ошибок и транслитерации на этапе запроса
- К запросам пользователя записывать заказанный товар или его отсутствие, чтобы понимать какие теги оказались искомыми
- Потенциально - сделать мэтчинг товаров с тегами

Wildhack

Wildhack

Спасибо за внимание!

Решение представлено
командой Rabies