Wildhack

Онлайн-хакатон Wildberries

Задача

Разработать алгоритм набора поисковых подсказок (тегов)

Команда WilderAI

Тихонов А. (Москва)
Утробин М. (Нижний Новгород)
Сергеев Ю. (Рязань)

93
% ru локаль

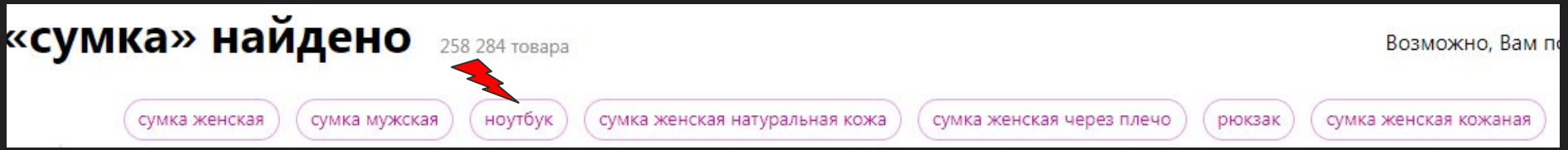
8.2
млн. клиентов

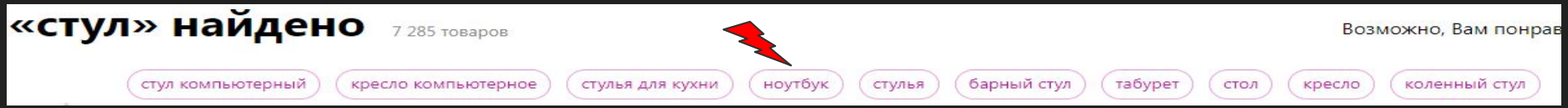
84
млн.
запросов

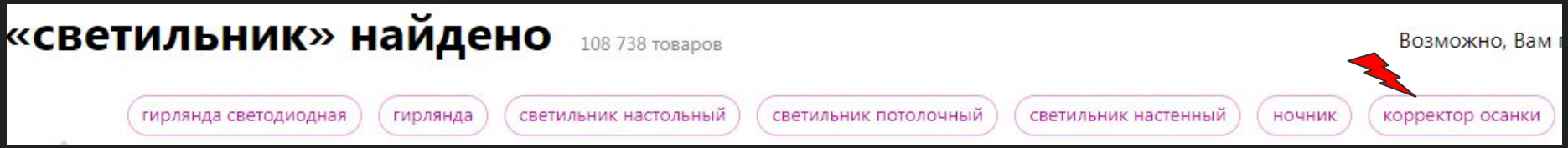
16
млн. уник.
запросов

Качество данных

- Текст -> токенизация
- Словарные ошибки - > лемматизация
- Разное написание -> приведение к одной форме
- .!,@ -> удаление пунктуации
- Некоторые теги ~30% простых запросов нерелевантны (ноутбук/iphone к разным товарам)







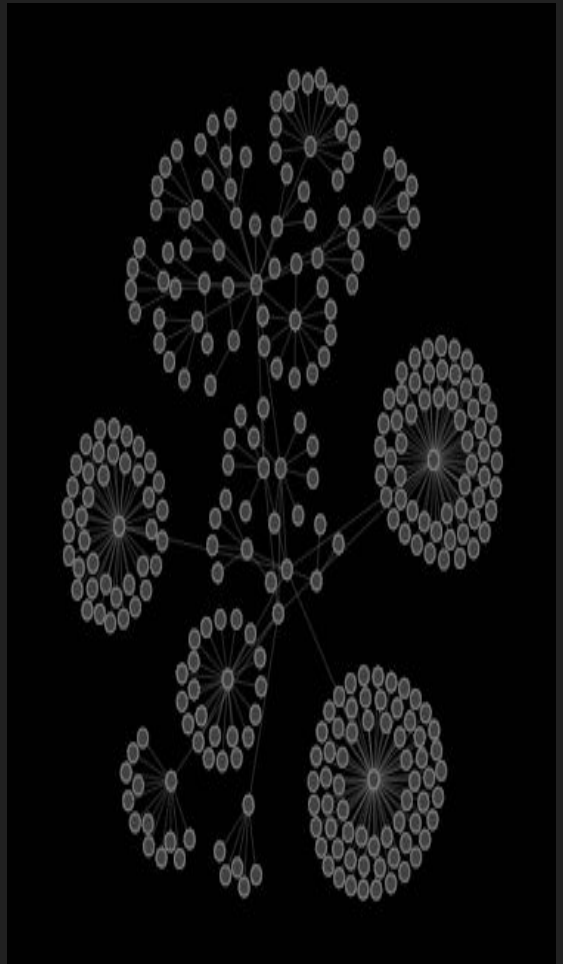
Основная идея

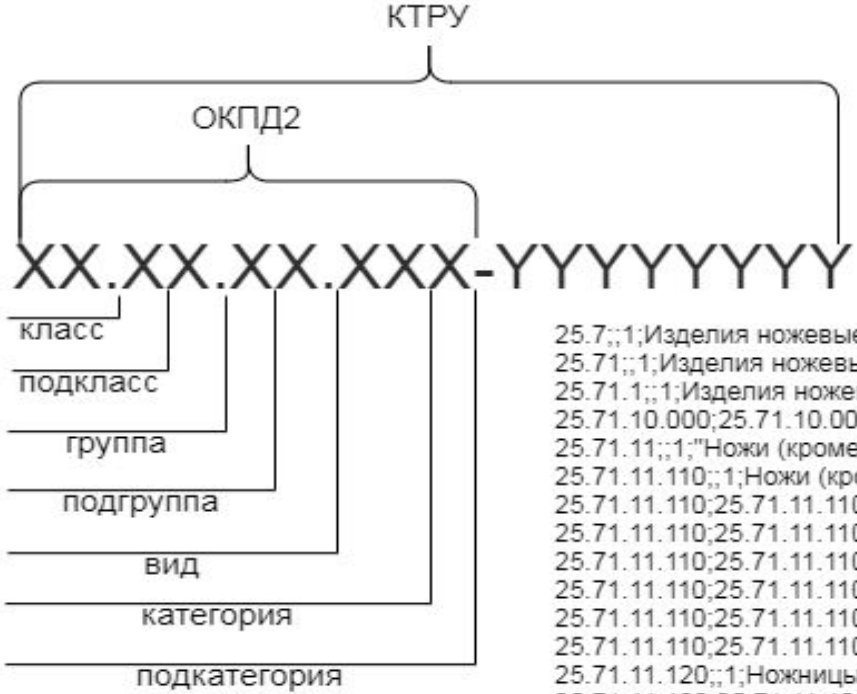
- Распределение истории запросов на классификаторы
- Выбор тегов в пределах узла классификатора

Итог

Вектор запросов увязанный на ОКПД2 и КТРУ
Классификаторы ОКПД2 и КТРУ - общедоступные

- ОКПД2 - Общероссийский классификатор продукции по видам экономической деятельности
- КТРУ - Каталог товаров работ и услуг (дополняющий ОКПД2)





25.7;;1;Изделия ножевые, инструмент и универсальные скобяные изделия
25.71;;1;Изделия ножевые и столовые приборы
25.71.1;;1;Изделия ножевые и столовые приборы
25.71.10.000;25.71.10.000-00000001;2;Изделия ножевые и столовые приборы
25.71.11;;1;"Ножи (кроме ножей для машин) и ножницы; лезвия для них"
25.71.11.110;;1;Ножи (кроме ножей для машин)
25.71.11.110;25.71.11.110-00000011;2;Кухонный нож-дозатор
25.71.11.110;25.71.11.110-00000016;2;Набор ножей хозяйственных
25.71.11.110;25.71.11.110-00000010;2;Нож макетный
25.71.11.110;25.71.11.110-00000014;2;Нож хозяйственный
25.71.11.110;25.71.11.110-00000005;2;Ножи специальные
25.71.11.110;25.71.11.110-00000006;2;Ножи хозяйственные
25.71.11.120;;1;Ножницы
25.71.11.120;25.71.11.120-00000012;2;Ножницы Зигзаг
25.71.11.120;25.71.11.120-00000010;2;Ножницы закройные/портновские
25.71.11.120;25.71.11.120-00000001;2;Ножницы канцелярские

Архитектура ML

Решение

Подготовка

- Этап 1 - Формирование векторов каталога ОКПД + КТРУ
- Этап 2 - Формирование векторов запросов пользователей
- Этап 3 - Объединение запросов и каталога
- Этап 4 - Хранение в БД

Входящий запрос

- Формируем вектор запроса, определяем 3 наиболее подходящих категории
- Выдаем наиболее релевантные запросы другие пользователей из данных категорий и подкатегорий
- Выдаем наиболее релевантные категории

Wildhack

Демонстрация

Wildhack | Тэги | API logs | Версии | Справка | Source Code

Введите запрос:

таблетки для стирки

Покажите теги

☑ Персональное тегирование (в разработке)

Введите id пользователя:

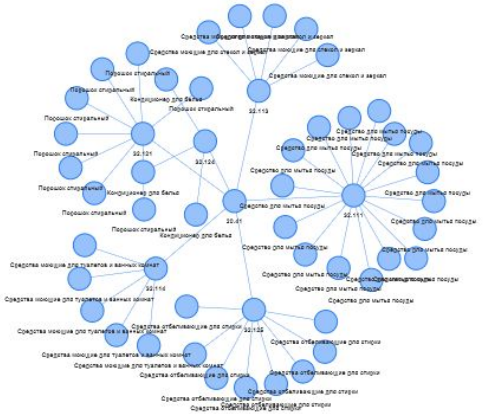
24254

Теги

1	таблетки для посудомойки
2	средства отбеливающие для стирки
3	салфетки для стирки
4	порошок стиральный
5	пластины для стирки
6	кондиционер для белья
7	самат таблетки для посудомоечной
8	средство для мытья посуды
9	гель для стирки япония
10	средства моющие для стекол и зеркал

Граф связи узлов

Select by id



Созданные признаки

- Категория запроса товара по Классификатору ОКПД+КТРУ
- Уровень иерархии по Классификатору ОКПД+КТРУ
- Получение частей речи текстов запросов
- Векторы текстовых запросов

Модель ML

fastText - векторное представление слов

- Готовые вектора обученные на Википедии РУ
- Работает с различным написанием
- Способна к сжатию без потери качества*

Код	Категория	Запрос
14.20.10.412	Куртки женские нагольные	куртка женская осенняя
14.13.10.000	Брюки спортивные	штаны женские спортивные
14.19.12.000	Костюм спортивный	костюм мужской спортивный
26.30.22.000	Мобильный телефон (смартфон)	iPhone 11
14.13.14.140	Платье женское	платье
14.20.10.352	Пальто женские нагольные	пальто женское демисезонное стеганое
26.30.22.000	Мобильный телефон (смартфон)	айфон 12
26.30.22.000	Мобильный телефон (смартфон)	iPhone 12
14.19.12.000	Костюм спортивный	спортивный костюм женский теплый

*Статья участника по сжатию fastText https://habr.com/ru/post/582980/

Поиск FAISS

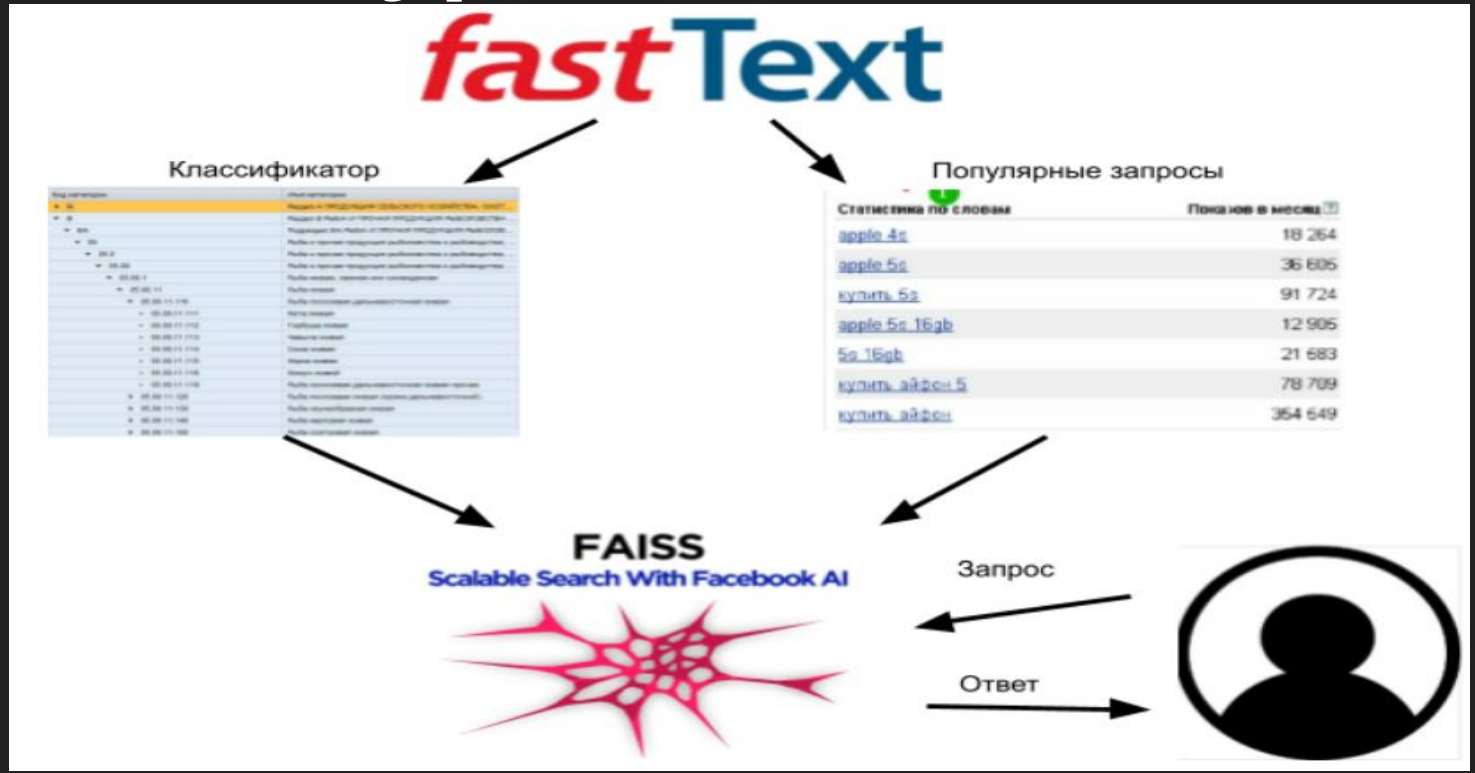
Facebook AI Research Similarity Search

- Поисковый движок в векторном пространстве
- Быстрый и масштабируемый
- Можно и на GPU
- Не требователен к оперативной памяти
- Различные методы поиска
- Работает на многомиллионных данных*

    - Нормализация векторов
    - Косинусное сходство

*Источник
https://habr.com/ru/company/dentsuRU/blog/509204/

Архитектура ML

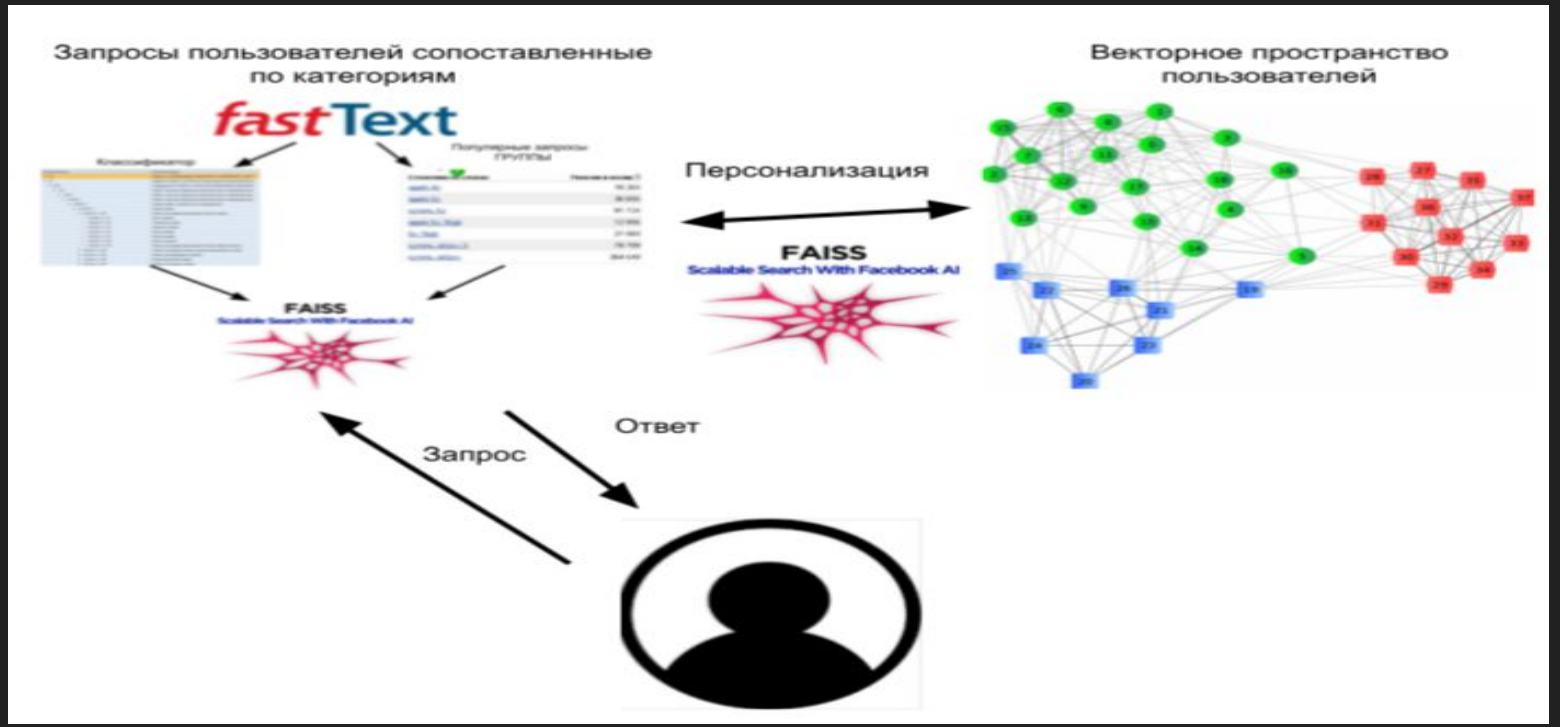


Персонализация

Общая схема предлагаемого решения

- История поиска
  - Храним векторное представление всех запросов пользователя
  - Формируем и храним вектор каждого пользователя
  - Кластеризуем пользователей в группы
- Входящий запрос
  - Определяем группу пользователя
  - Ищем уникальные запросы среди группы пользователя
    - Предварительно наложенные на классификатор
  - Формируем уникальную выдачу по интересам данной группы

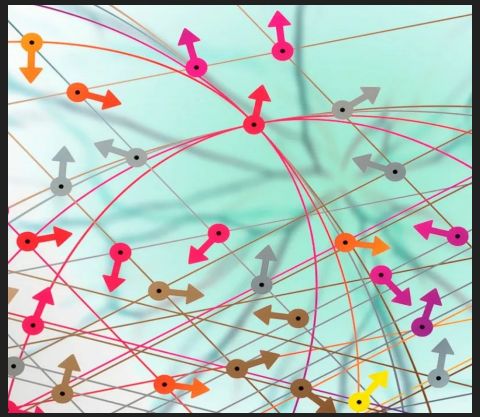
Персонализация



Персонализация

Вектор пользователя в виде суммы N последних запросов

- Вектор нового запроса прибавляем
- Вектор выбывающего N+1 запроса вычитаем
- С каждым запросом получаем немного обновленный вектор
- Миграция пользователя, в зависимости от изменения интереса



Развитие

Улучшения

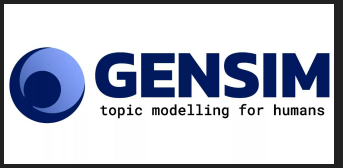
- Персональное тегирование
- Сезонность
- Дополнительные данные
- LSTM с весами FastText
- Работа с базой данных
- CI/CD

Используемые технологии

Решение полностью построено на открытом ПО с использованием языков R и Python

Список использованных библиотек и технологий:

- R - обработка данные, анализ, визуализация
- Shiny - веб-интерфейс
- Python - реализация модели
- spaCy - нормализация текста
- Gensim fastText - реализация алгоритма NLP
- faiss - поиск векторов
- numpy - операции с типами и массивами
- pandas - оперирование данными



Состав команды

Алексей Тихонов

mail: potom2007@yandex.ru

tg: @aitikhonov

Михаил Утробин

mail: utrobinmv@yandex.ru

tg: @utrobinmv

Юрий Сергеев

mail: jurbanhost@mail.ru

tg: @MentalSky

Спасибо за внимание!