Классификация хостов

Хакатон МТС.Тета

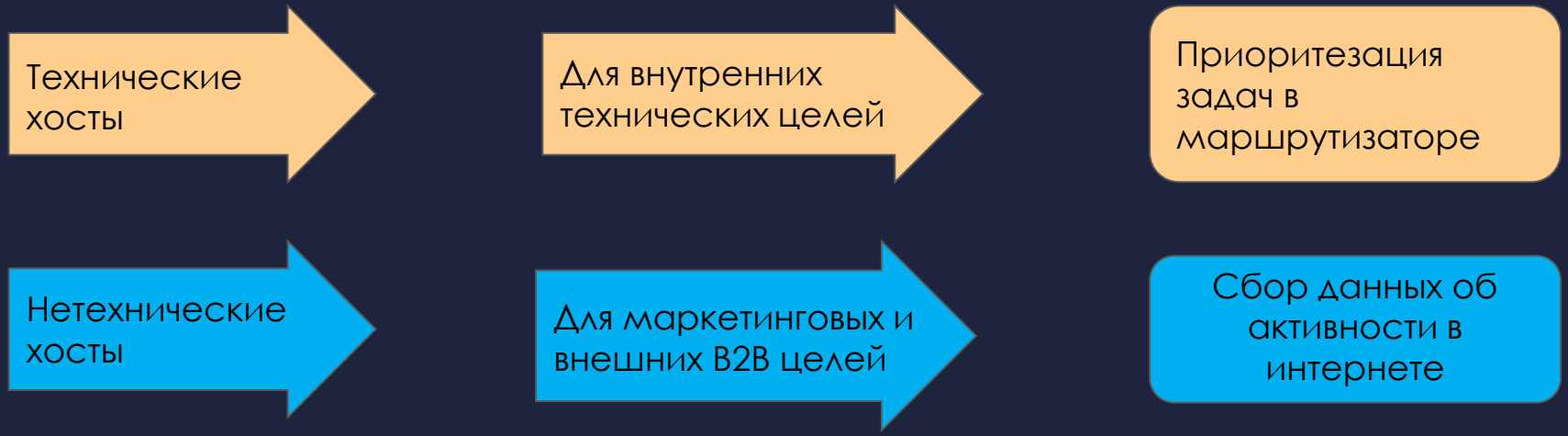
Команда “Машинисты”
Юлия Фокина
Рафаэль Рамазанов
Анастасия Мердалимова

Задача

Бинарная классификация хостов

Задача: разделять хосты на **технические** и **нетехнические**

Зачем?



Задача

Зачем?

Допустим, эти данные собираются с базовых станций
=> есть их **геоданные**. Тогда:

Нетехнические хосты, на которые осуществлялся вход

Геолокация базовых станций

Данные, ГДЕ (в каких регионах) КУДА (на какие сайты), КОГДА и как часто заходят
=> сравнительный анализ

Настроения, предпочтения, ценности местных жителей

Задача

Так и зачем?

Настроения, предпочтения, ценности местных жителей

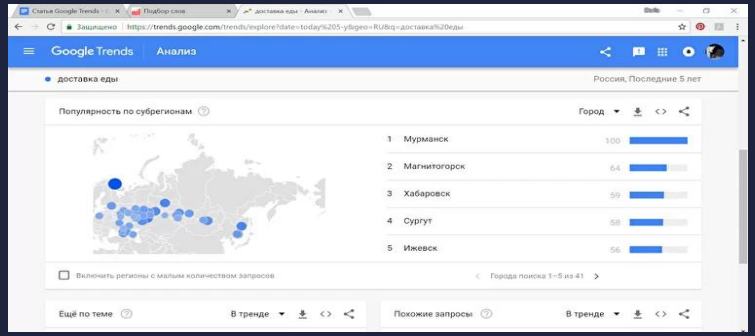
Политтехнологии	Бизнес, торговля	Реклама, маркетинг	Контроль доступа
Оценка **популярности кандидатов** и **госпрограмм**	Предиктивная аналитика: ●Оценка рентабельности **выхода на рынок** в новом регионе ●Предсказание изменения спроса по **динамике** запросов	Оценка целесообразности и **стоимости** размещения **рекламы:** ● на разных сайтах ● в разных регионах (в т.ч. офлайн)	Оценка безопасности сайтов Детский контент

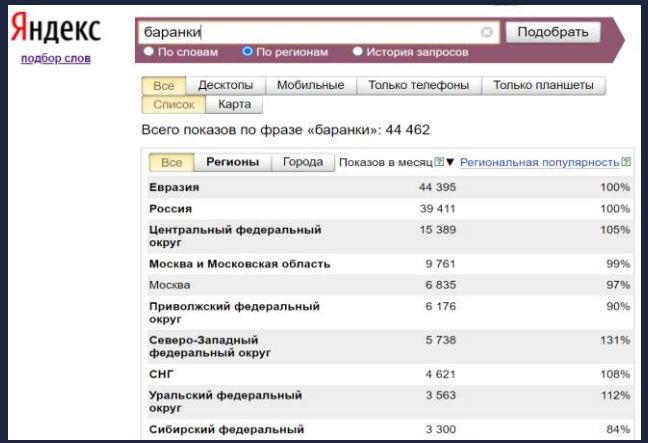
Задача

Аналоги

- **Wordstat.yandex**
- **Google Trends**

- анализируют именно Поисковые запросы конкретного поисковика





А если пользователь заходит на сайт по памяти/в Закладках/Google - не учитывается

Предобработка

Исходные данные:
список хостов **1.000.000 записей**

Предобработка:
< 200.000 записей
Убраны дубликаты записей

	host
0	api.youla.io
1	favicon.yandex.net
2	w-74721.fp.kaspersky-labs.com
3	questtime.net
4	passport-authproxy.taxi.yandex.net
...	...
999996	pull-hls-f16-sg01.tiktokcdn.com
999997	m5.ehgeqxn.me
999998	sun9-6.userapi.com
999999	m9.igoxzza.com
1000000	

1000001 rows × 1 columns

Предварительные исследования

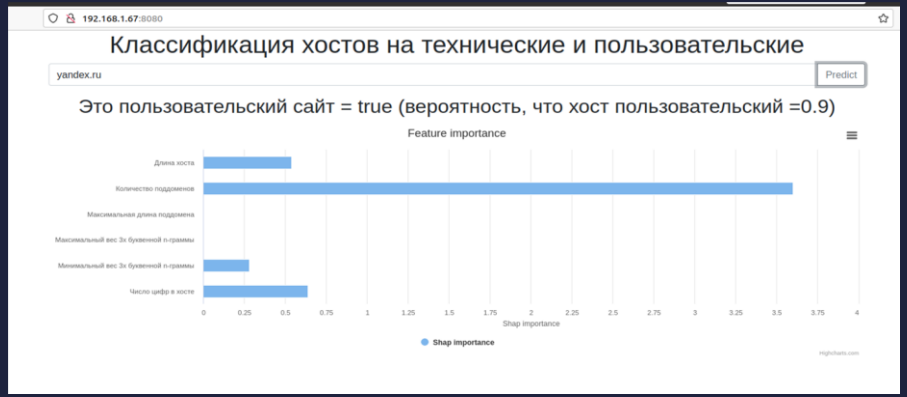
Разметка

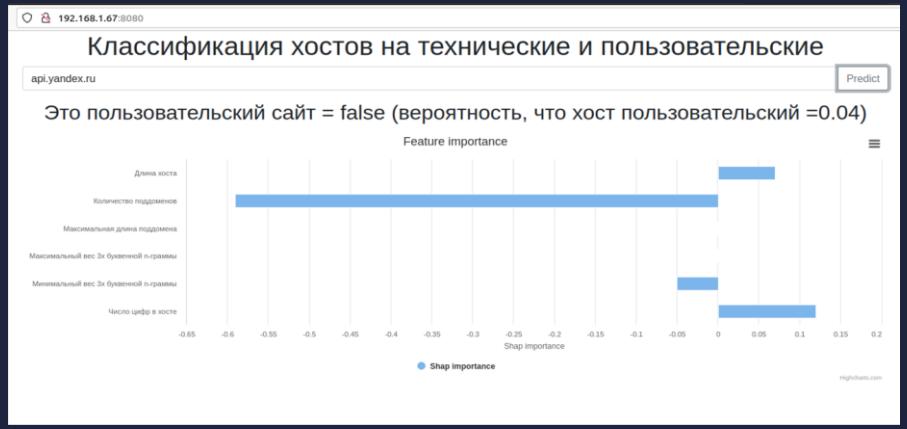
Парсинг => robots.txt в корне

Технические	**Нетехнические**
• Есть **ключевые слова** (буквосочетания), конкретные для начала/конца/любого положения • **IP** адреса • словарь английского (если ни одного слова)	• вспомогательные датасеты (пользовательские URL) • Ключевые слова, сочетания букв

https://data.iana.org/TLD/
https://www.kaggle.com/shawon10/url-classification-dataset-dmoz?select=URL+Classification.csv
https://github.com/dwyl/english-words

ML-модель: демо





ML-модель: концепция

Тип модели: CatBoost

Переменная	Признак	Важность на тестовой выборке
digits_count	число цифр	28.7
max_domain_level	число точек	22.5
url_len	длина адреса	19.7
ngram_max	макс. вес 3-букв. сочетания	19.4
ngram_min	мин. вес 3-букв. сочетания	8.1
max_domain_part_len	макс. длина имени домена	1.6

ML-модель: метрика

Валидация: разметкой вручную (200 хостов)

Функция потерь: LogLoss

Метрики качества:

Precision = 72% хостов из предсказанных, как нетехнические определены верно

Recall = 39% реальных нетехнических хостов найдена

Precision = TP / (TP + FP)
Recall = TP / (TP + FN)

TP = кол-во нетехнических хостов, которые предсказали верно как нетехнические
FP = кол-во технических хостов, которые предсказали неверно, как нетехнические
FN = кол-во нетехнических хостов, которые предсказали неверно, как технические

Что дальше?

Выделили пользовательские хосты

Дальше - их классифицировать **по содержанию**:

- Просматриваем содержимое полученных пользовательских хостов и классифицируем их с помощью модели, классифицирующей тексты
- Берем размеченный датасет с пользовательскими url и их категориями*
- Используем как обучающую выборку для классификатора

*Например, https://www.kaggle.com/shawon10/url-classification-dataset-dmoz?select=URL+Classification.csv