Платформа для поиска похожих кроссовок

Состав проекта

Куратор: Георгий Панчук

Распределение ролей участников:

Литвинов Вячеслав: парсинг, eda, модели

Моисеев Даниил: парсинг, исследование моделей, телеграм-бот

Цели и задачи проекта

- Сбор данных (изображения кроссовок) с различных источников
- Обработка, чистка и объединение данных
- Разведочный анализ данных
- Получение эмбеддингов для картинок с помощью различных алгоритмов
- Исследование качества моделей классификации и поиска похожих изображений
- Обертка этих моделей в сервис и телеграмм-бот

Датасеты и парсинг

- Спарсили 5 интернет-магазинов кроссовок
- Парсили скрапингом и используя публичные API
- Храним данные с помощью DVC на S3
- Описали данные в .md файлах

POLO RALPH LAUREN
Masters Sport "White / Blue"
€149,95

MERRELL
Moab Speed GTX "Black"
€169,95

MERRELL
Agility Peak 5 "Black"
€159,95

MERRELL
Moab Speed GTX "White"
€169,95

	brand	title	pricecurrency	price	images_path
1	Nike	Blazer Mid Next Nature "White"	EUR	49.99	data/raw/images/sneakerbaas/category-kids/nike/b…
2	New Balance	550 PS "Lavender"	EUR	99.95	data/raw/images/sneakerbaas/category-kids/new-ba…
3	New Balance	550 "astro dust"	EUR	79.99	data/raw/images/sneakerbaas/category-kids/new-ba…
4	Reebok Classics	CLUB C CARDI "Blue"	EUR	34.99	data/raw/images/sneakerbaas/category-kids/reebok…
5	Nike	Air Max TW "University Red"	EUR	59.95	data/raw/images/sneakerbaas/category-kids/nike/a…
6	New Balance	2002 "TIMBERWOLF"	EUR	94.95	data/raw/images/sneakerbaas/category-kids/new-ba…
7	Nike	Nike Blazer '77 "Jade Ice"	EUR	69.95	data/raw/images/sneakerbaas/category-kids/nike/n…
8	Reebok Classics	Cardi B Club C "Red"	EUR	34.99	data/raw/images/sneakerbaas/category-kids/reebok…
9	Karhu	Albatross 82 "Navy"	EUR	29.99	data/raw/images/sneakerbaas/category-kids/karhu/…
10	Nike	Huarache Run GS "Jade Ice"	EUR	89.95	data/raw/images/sneakerbaas/category-kids/nike/h…
11	New Balance	550 PS "Orange Sea"	EUR	99.95	data/raw/images/sneakerbaas/category-kids/new-ba…
12	New Balance	550 PS "Marina"	EUR	99.95	data/raw/images/sneakerbaas/category-kids/new-ba…
13	New Balance	550 "Lavender"	EUR	89.95	data/raw/images/sneakerbaas/category-kids/new-ba…
14	New Balance	2002 "ECLIPSE"	EUR	94.95	data/raw/images/sneakerbaas/category-kids/new-ba…
15	New Balance	2002 "BLACK"	EUR	99.95	data/raw/images/sneakerbaas/category-kids/new-ba…
16	Jordan	Air Jordan 6 Retro GS "Dutch Blue"	EUR	99.99	data/raw/images/sneakerbaas/category-kids/jordan…
17	New Balance	2002 "SLATE GREY"	EUR	99.95	data/raw/images/sneakerbaas/category-kids/new-ba…
18	Reebok Classics	CLUB C CARDI "Blue"	EUR	49.99	data/raw/images/sneakerbaas/category-kids/reebok…
19	Vans	Old Skool "Powder Pink"	EUR	39.99	data/raw/images/sneakerbaas/category-kids/vans/o…
20	New Balance	550 "WHITE"	EUR	79.99	data/raw/images/sneakerbaas/category-kids/new-ba…
21	Nike	Air Max Motif "WHITE"	EUR	59.99	data/raw/images/sneakerbaas/category-kids/nike/a…
22	Reebok Classics	CLUB C CARDI "Purple"	EUR	24.99	data/raw/images/sneakerbaas/category-kids/reebok…
23	New Balance	2002 "ECLIPSE"	EUR	99.95	data/raw/images/sneakerbaas/category-kids/new-ba…

Основные столбцы полученных метаданных

Предобработка и объединение данных

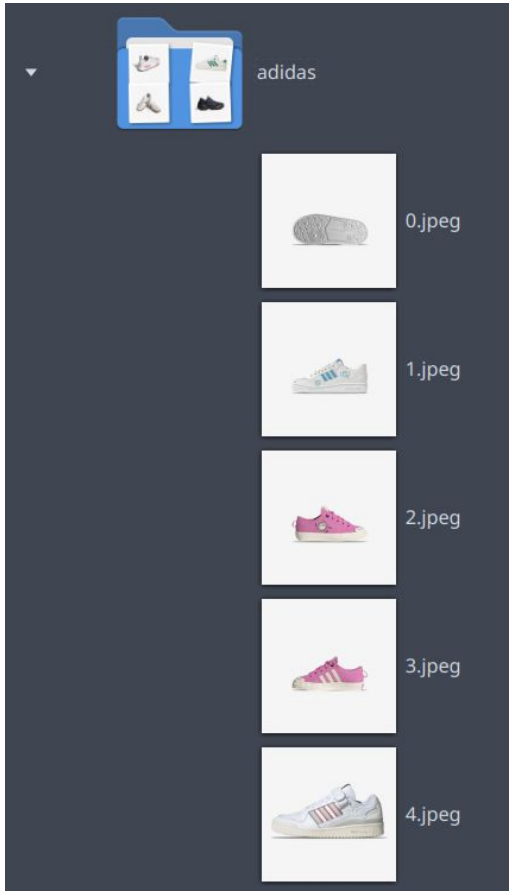
- Для каждого датасета провели базовую предобработку
  - Колонки: удалили ненужные, привели к одному названию
  - Строки: удалили дубликаты, почистили от лишних символов
- Привели название и бренд к одному формату
- После объединения получили два датасета

Объединили по моделям: 2067 моделей кроссовок объединились в 877

Объединили по брендам: 30 брендов в итоге

Предобработка изображений

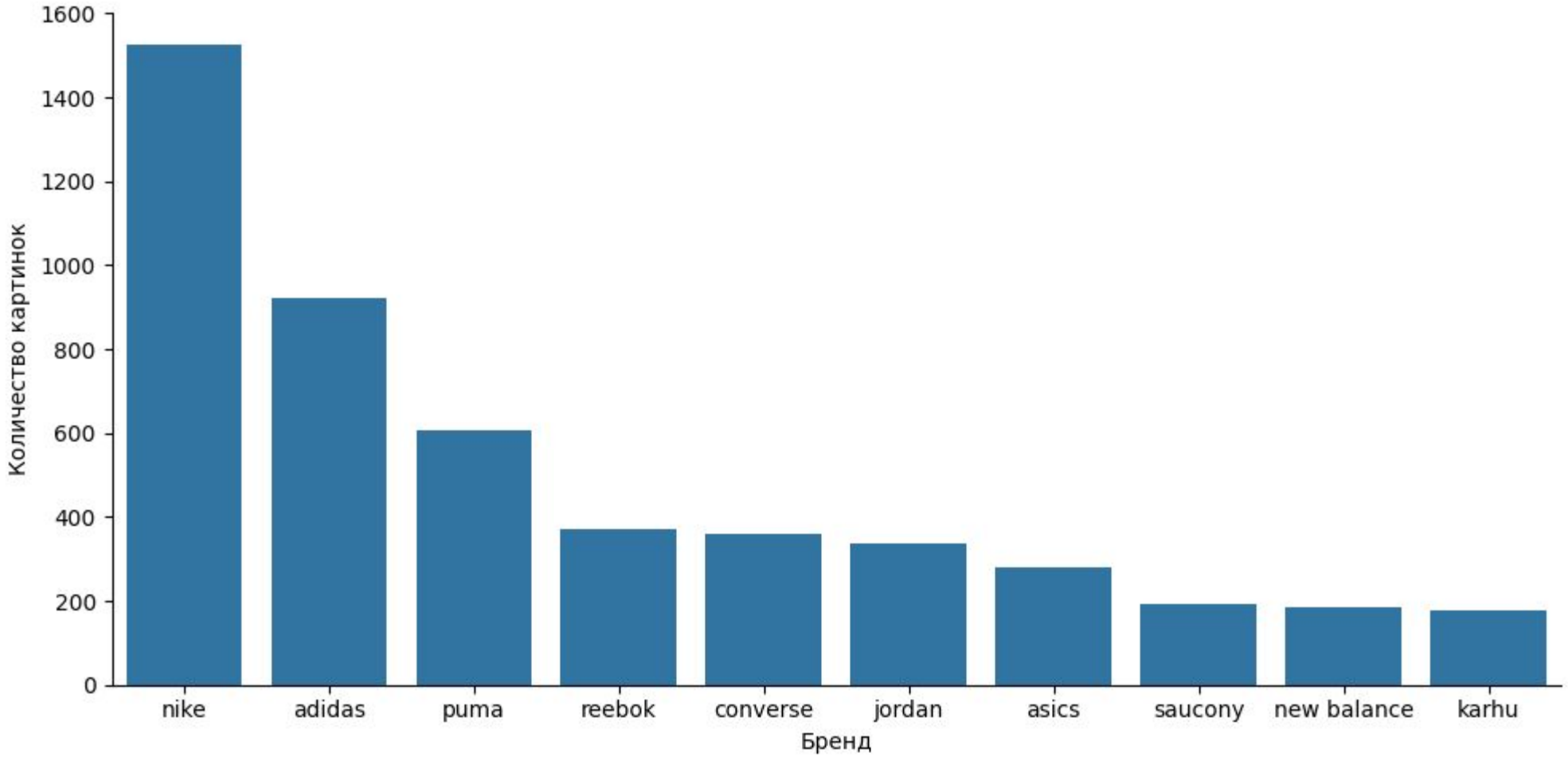
- Удалили идентичные картинки
- Перевели в палитру RGB
- Привели к одному формату .jpeg



EDA

- Вывели как выглядят метаданные и их основные статистики
- Убедились, что в данных нету пропусков
- Убедились, что все картинки в одинаковом формате
- Посмотрели размеры картинок - они различные, от 400x400 до 2000x2000
- Общее количество картинок - 5852

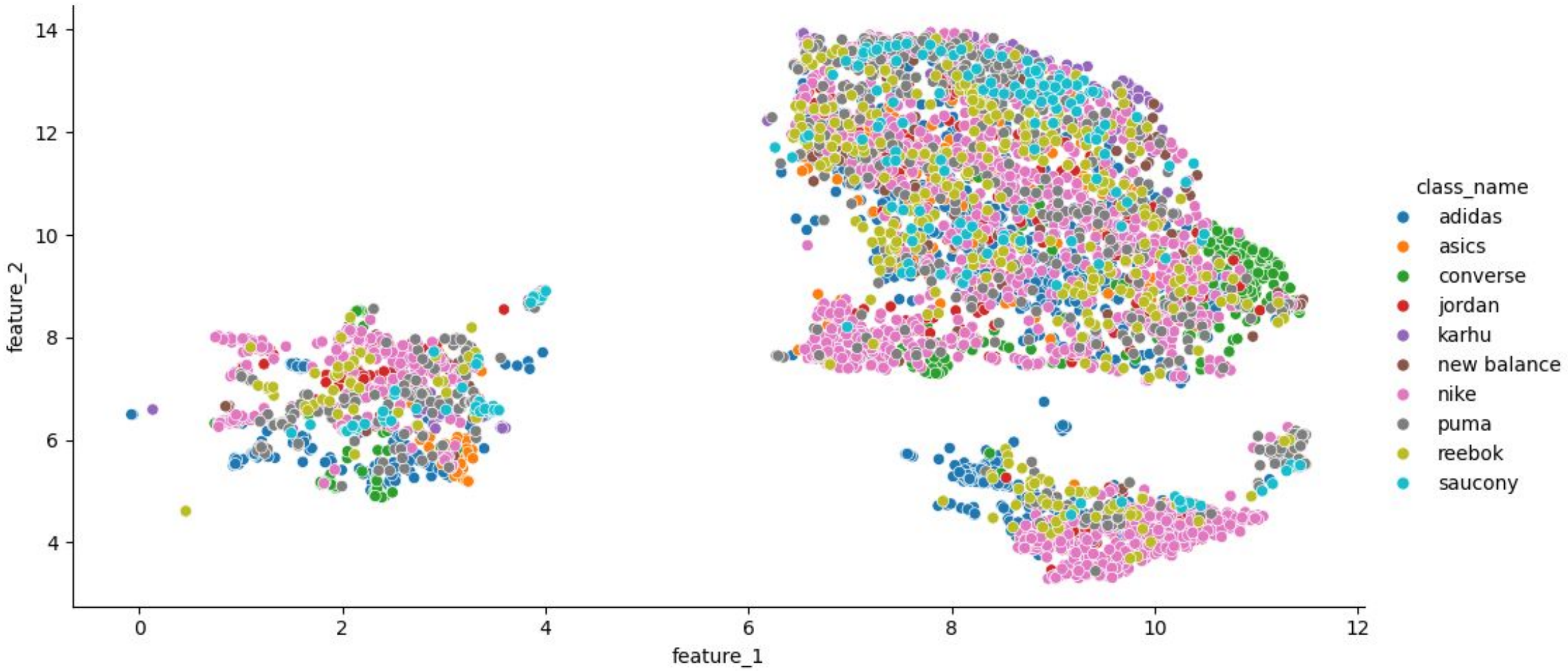
Классификация брендов



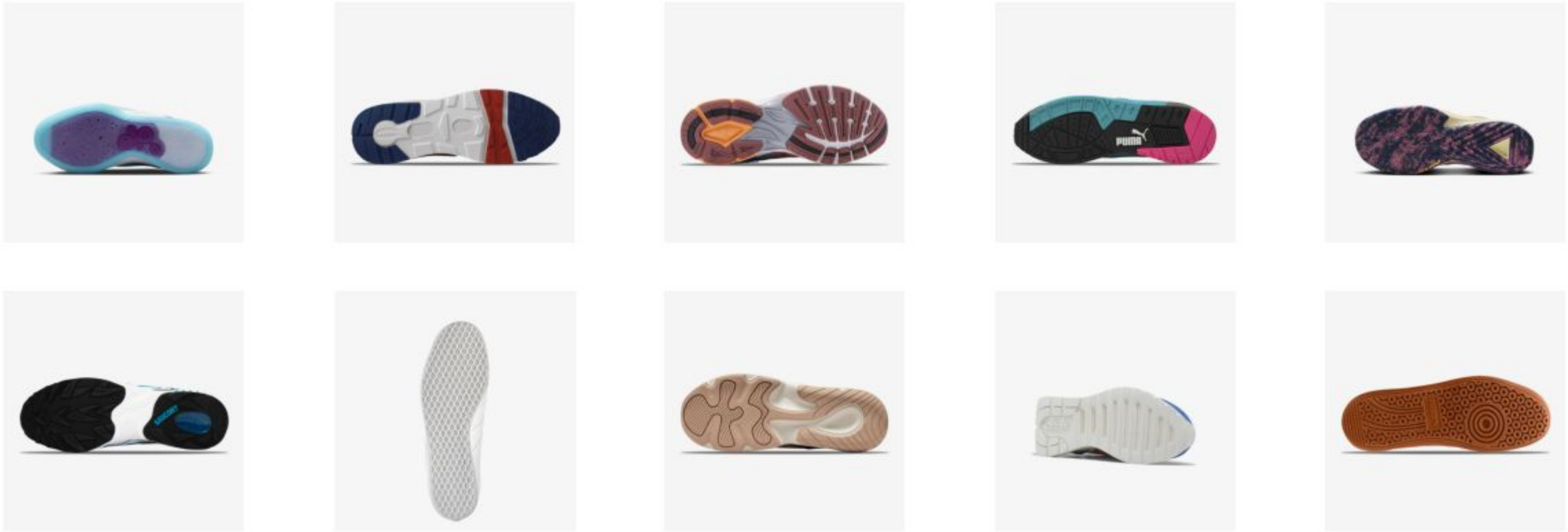
Зависимость числа картинок от бренда

Features

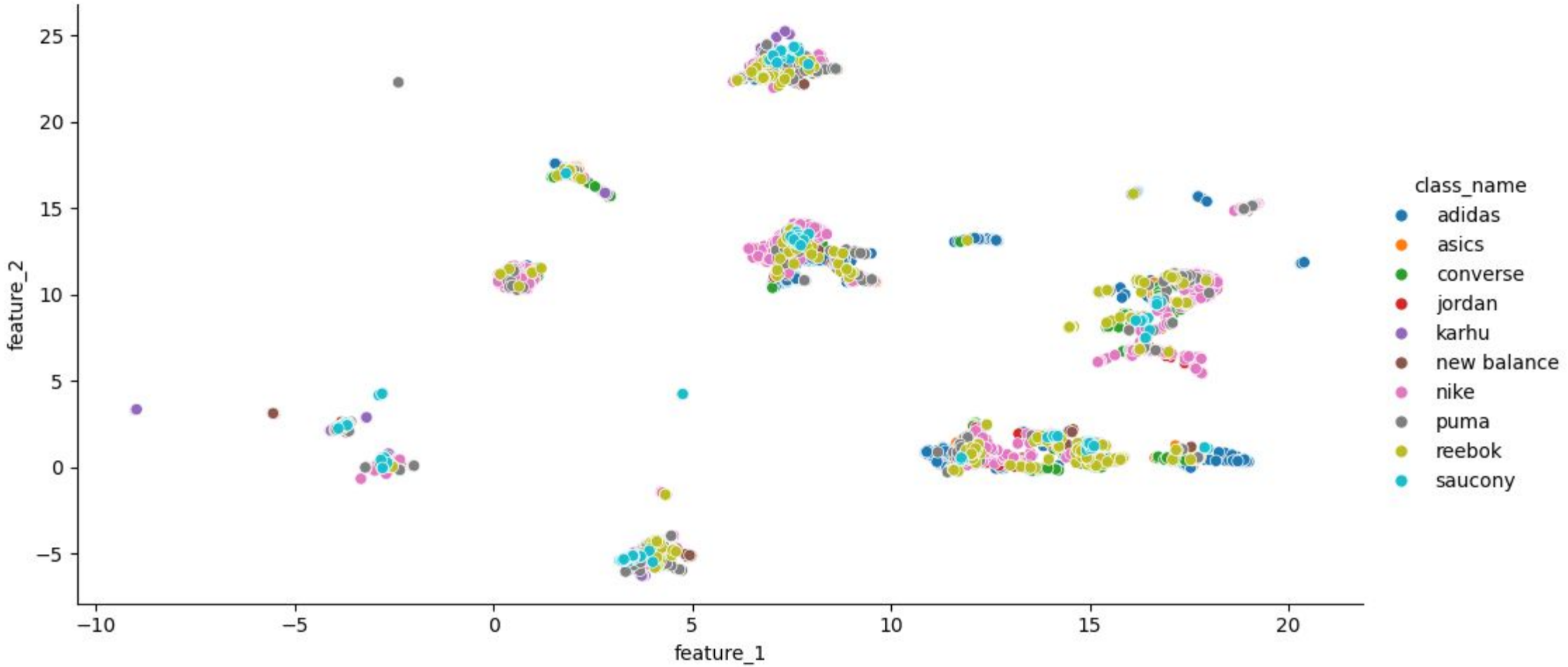
- Поделили датасет на train/val/test в пропорциях 60/20/20
- Отобрали только те бренды, у которых больше 100 картинок
- Получилось 13 брендов для классификации
- Для всех картинок взяли эмбеддинги с помощью ResNet, HOG, SIFT
- Для визуализации эмбеддингов использовали SVD, T-SNE, PCA, UMAP



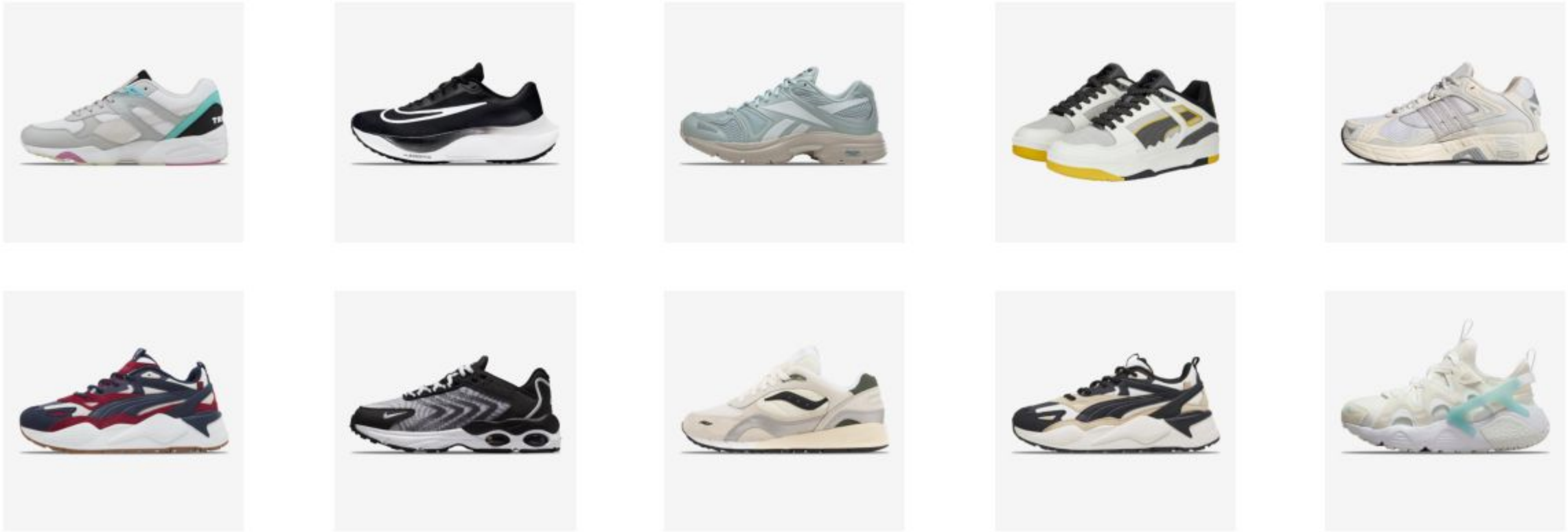
Визуализация ResNet features с помощью UMAP



Визуализация ResNet features с помощью UMAP - левый кластер



Визуализация HOG features с помощью UMAP



Визуализация HOG features с помощью UMAP - верхний кластер

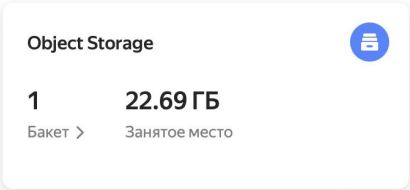
Результаты моделей

- Считали метрику F1-weighted и F1-macro
- Baseline модель предсказывает самый частый класс
- Использовали модели SVM, SGD, CatBoost
- Перебирали параметры по сетке с кросс-валидацией
- Лучшее качество показывает hog-svm

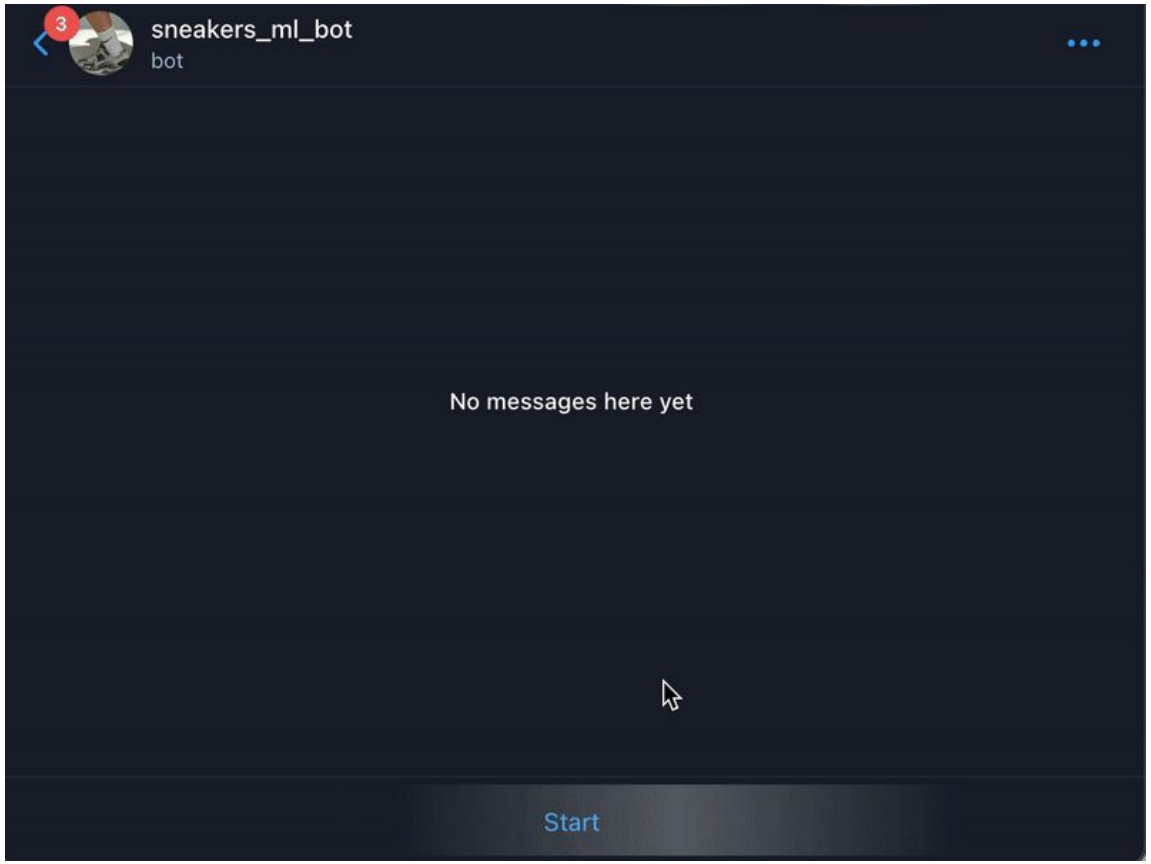
Модель	Baseline	hog-svm	hog-sgd	hog-catboost	resnet-svm	resnet-sgd	resnet-catboost
F1-weighted	0.13	**0.81**	0.76	0.72	0.76	0.69	0.68
F1-macro	0.03	**0.78**	0.72	0.68	0.76	0.65	0.64

Структура проекта

- Данные, эмбеддинги и модели храним с помощью DVC на S3
- Добавили различные линтеры и форматтеры
- В проекте настроили poetry и pre-commit
- Модели храним в onnx формате
- Реализовали базовый телеграмм-бот, CI/CD сборка Docker-образа
- Документация проекта в .md файлах в /notes



Демонстрация работы бота



Будущие задачи

- Улучшить качество классификации с помощью нейросетевых моделей
  - Fine-tune ResNet
  - Vision Transformer
- image2image поиск
- Similarity learning
- Обертка ML в FastAPI-сервис, использование его в телеграмм-боте и streamlit
- Сборка датасета из изображений пользователей боту в базу данных
- Флоу оценки о качестве модели
- Логирование, мониторинг и алерты production-окружений (streamlit, tg-bot)

Ссылки

- Репозиторий: https://github.com/miem-refugees/sneakers-ml
- Телеграм-бот: https://t.me/sneakers_ml_bot