ЛИДЕРЫ ЦИФРОВОЙ ТРАНСФОРМАЦИИ 2023

OZON OZON Tech

ТЕХНИЧЕСКОЕ ЗАДАНИЕ

ЗАДАЧА 05

Поиск одинаковых товаров на маркетплейсе

1. Актуальность задачи

Ozon — ведущая мультикатегорийная платформа электронной коммерции и одна из крупнейших интернет-компаний в России. На площадке представлено более 150 млн товарных наименований в 20 категориях: от книг и одежды до продуктов питания и товаров для здоровья.

Сейчас более 90% ассортимента площадки формируют партнеры маркетплейса. И в таком внушительном ассортименте товаров разные продавцы могут предлагать одинаковые товары по разной стоимости и с разными сроками доставки. Чтобы клиенты лучше ориентировались в предложениях продавцов, Ozon нужно постоянно совершенствовать алгоритм матчинга одинаковых товаров.

Сейчас благодаря матчингу:

- пользователь видит плашку «Есть дешевле» и выбирает наиболее подходящий по цене товар;
- предложениям продавцов присваивается выгодный, умеренный или невыгодный индекс цен (который определяется на основе сравнения предложения продавца с конкурентами и другими площадками);
- к товарам с неполными атрибутами добавляются недостающие — из тех карточек, которые описаны более качественно.

2. Постановка задачи

Участникам хакатона предстоит разработать ML–модель, которая по названиям, атрибутам и картинкам сможет ответить на вопрос являются ли два товара одинаковыми. Модель должна найти среди предложенных пар-кандидатов как можно больше одинаковых товаров с долей ошибок меньше 25%. Это решение поможет клиентам Ozon улучшить пользовательский опыт, а компании — оптимизировать ресурсы и сэкономить на закупке серверного оборудования.

3. Требования к решению

1. Участникам нужно получить разметку для тестовой выборки в формате csv-файла с заполненным полем score.
2. Язык Python и библиотеки в открытом доступе (pypi, conda, etc.)
3. Используются только данные, которые предоставляются организаторами соревнования **(только архив hackathon_files_for_participants_ozon.tar.gz)**
   Использование собственного датасета не разрешается.

4. Разметка получена автоматически без участия человека или GPT. Можно генерировать автоматически новую разметку на основе предоставленных данных или удалять неправильные пары (не вручную!), но код должен воспроизводиться. Нельзя вручную делать переразметку тренировочной выборки и разметку тестовой.
5. Код решения выкладывается в github/gitlab/bitbucket в закрытый репо. После дедлайна доступ на просмотр открывается, чтобы код можно было клонировать, воспроизвести окружение и запустить решение. Для воспроизводимости необходимо подготовить файл README с описанием самого решения, а также инструкцией по запуску пайплайна подготовки данных, обучения модели и ее применения.
6. Любая модель должна помещаться на карте с 40ГБ памяти без model parallelism-а.
7. Основной пайплайн (без обучения моделей) должен длиться не более 2-х часов.

4. Источники данных

- Тренировочная выборка: пары одинаковых и различных товаров (id товаров, target).
- Тестовая выборка: пары товаров (id товаров) без разметки (выборка для формирования лидерборда).
- Данные о товаре: id товаров, названия, векторные представления названий (эмбединги), атрибуты, векторные представления картинок (эмбединги) товаров.

Как был получен датасет?

С помощью разметки людьми в Ozon profit.

Baseline

Участникам предоставят baseline – jupyter notebook, содержащий:

- Чтение данных.
- Джоин таблиц (dataset join etl, test join etl).
- Расчёт простых фичей.
- Обучение модели.
- Расчёт метрики для отложенной из обучающей выборки.
- Применение модели к тестовой выборке.
- Сохранение результатов в submission_example.csv.

5. Требования к сдаче решений на платформе

- Ссылка на презентацию
- Ссылка на сопроводительную документацию
- Ссылка на репозиторий
- Приложить пример csv-файла (структура id1, id2, score).

variantid1	variantid2	scores
1234	3425	0.99
3456	5678	0.15

6. Критерии оценки

Для быстрой проверки и прозрачной оценки ML-моделей будет представлена платформа DS Works, где будут отображаться промежуточные результаты каждой Команды в публичной рейтинговой таблице (лидерборде). Лидерборд будет содержать: название команды, скор – рассчитываем PRAUC(Precision = 75%) на основе предсказаний участников из загруженного ими csv и фактической разметки. PRAUC(Precision = 75%) — это площадь под графиком Precision-Recall Curve в зоне, где Precision > 0.75.

В первые дни соревнования на лидерборде будет отражаться лучший скор из всех отправленных засылок по 40% теста (public leaderboard). На финале соревнования будут открыты результаты на оставшихся 60% (private leaderboard). В день будет 5 попыток заслать предсказания.

Экспертное жюри проверит на воспроизводимость решения топ-10 команд с лучшими скорами на прайват лидерборде. При этом претенденты на победу должны обязательно набрать скор выше чем даёт разметка из submission_example.csv. Если команда дисквалифицирована (т.е. если кодом решения не воспроизводятся скоры модели в сабмите), то жюри проверит решение следующих команд в топе лидерборда. На финальном этапе участникам предстоит защитить свои решения перед экспертами, после чего будут оглашены 3 команды, занявшие призовые места.

Доступ к платформе DS Works будет предоставлен 21 мая для зарегистрированных пользователей. Поэтому каждому члену команды необходимо предварительно зарегистрироваться на сайте https://dsworks.ru с той же почты, с которой он регистрировался на конкурс ЛЦТ.

Как пользоваться платформой можно узнать из митапа от экспертов DS Works, который появится в вашем личном кабинете на платформе ЛЦТ.