цифровой прорыв
сезон: ИИ

РОССИЯ – СТРАНА ВОЗМОЖНОСТЕЙ

Задача «Поиск и классификация товаров народного потребления»

Введение:

Формирование тенденций потребления товаров у населения — важный аспект для анализа состояния страны. Такая информация может многое рассказать о средней заработной плате на выбранной территории, потребительской способности населения, количестве покупок конкретных товаров.

Вот только у поставщиков не существует общего ГОСТа для заведения позиций товара. Знакомый нам кочан белой капусты может быть записан как: «белокачанная капуста», «капуста белокочанная», «капуст. бел. коч.». Это может создавать ошибки при формировании отчетов и усложнять анализ данных.

Участникам предлагается найти и классифицировать конкретные категории товаров в огромном наборе различных наименований.

Условие задачи:

Цель этой задачи — поиск и классификация конкретных категорий товаров в огромном наборе различных наименований. Необходимо разметить следующие категории:

- Хлебные изделия
- Напитки газированные
- Вода
- Молоко 2.5-3.2 %
- Молочные продукты
- Макароны
- Фрукты
- Продукция общепитов
- Товары без категории

Описание входных значений:

train.csv — тренировочный датасет, содержащий 600 000 строк с наименованиями товаров и их категориями
categories.xlsx — содержит название и общую характеристику категории
test.csv — датасет для предсказания, содержит только наименования товаров
sample_submision.csv — пример решения для отправки

На что стоит обратить внимание:

В наборе данных существует сильный перевес в сторону товаров без категории и именно поэтому от участника ожидается использование NLP-моделей.

Также стоит обратить внимание, что условия принадлежности товаров к определенной категории описаны в файле categories.xlsx.

Метрика:

Так как задача ставит перед собой цель — точное соотношения категорий, в качестве метрики выступает recall.

Recall считается как:

$$Recall = \frac{TP}{TP + FN}$$

где: TP (True Positive) - количество верно угаданных значений одного класса
FN (False Negative) - количество не правильно угаданных значений одного класса

Правила чемпионата:

1. С момента открытия датасета до момента завершения приема решений репозиторий участника, в котором он ведет разработку по задаче текущего чемпионата, должен оставаться закрытым.

2. Участник обязан открыть доступ к репозиторию на чтение по ссылке (которая была прикреплена в ЛК в поле «Ссылка на код (гитхаб)») не позднее чем в течение 12 часов с момента окончания дедлайна отправки решений на региональном чемпионате.

3. Согласно п. 5.8 Положения в процессе верификации решений организаторы и технические эксперты, проверяющие решения участников, вправе назначить интервью с участниками чемпионата. Участник получит приглашение и ссылку на интервью не позднее чем за 12 часов до публикации итогового лидерборда. Пропуск интервью участником является поводом для дисквалификации.

4. Организаторы вправе исключить участника из призовых позиций лидерборда за непредоставление одного из артефактов решения задачи: тизера, скринкаста, презентации, ссылки на репозиторий.

5. Организаторы вправе дисквалифицировать участника в случае выявление плагиата кода или несоблюдение Положения конкурса.

6. Участник, получивший 2 дисквалификации за сезон проекта, попадает в чёрный список с дальнейшим отстранением от участия в чемпионатах до конца сезона.