цифровой прорыв
сезон: ИИ

РОССИЯ – СТРАНА ВОЗМОЖНОСТЕЙ

Задача «Разработка модели для поиска музейных экспонатов»

Введение

Музеи – это настоящее хранилище данных, которое позволяет погрузиться в любую историческую эпоху и соприкоснуться с культурой различных народов. Сейчас в нашей стране для посетителей открыты более 5000 музеев. В их числе Пермский краеведческий музей, основанный в конце XIX века. Фонд этого культурного центра хранит порядка 625000 экспонатов, включая более 50 коллекций регионального, российского и мирового значения. Применение информационных технологий в музейном деле позволяет вести учет экспонатов, хранить фото, видео- и аудиофайлы, формировать каталоги коллекций, обеспечивать безопасность уникальных произведений и быстрый поиск необходимого экспоната.

Однако, структура базы данных, существующей в Пермском краеведческом музее, предполагает методы поиска только по заданным иерархиям. Например, найти все произведения из 1900 года, или все экспонаты со словом «чашка» в названии. При этом невозможно установить внутренние связи между близкими предметами. Более сложный поисковый запрос, например, все розовые чашки, относящиеся к периоду с 1895 по 1917 гг., с неразбитыми или отреставрированными блюдцами не может быть выполнен, так как чашка и блюдце – отдельные музейные предметы. Поиск с более свободными критериями запрашиваемого объекта пока невозможно выполнить. Поэтому при подготовке выставок часто приходится вручную пересматривать все выставочные образцы.

Условие задачи

Ваша задача – разработать модель, которая будет по текстовому описанию музейного экспоната определять, какая фотография из базы данных ему соответствует, что в дальнейшем позволит создавать выборки предметов по более гибким параметрам поиска.

Описание входных значений

- train.csv — файл, содержащий данные с описанием объекта и номером соответствующей ему фотографии;
- train/ — папка с музейными экспонатами для обучения;
- test.csv — файл с описаниями объектов для предсказания;
- test/ — папка с музейными экспонатами, с которыми необходимо сопоставить описание;
- submission.csv — пример файла для отправки.

Метрика

Хоть задача с виду похожа на классификацию, стоит учесть, что каждое описание и каждая фотография в наборе уникальны, а значит в ответе моделей не должно быть “дублей”. В качестве метрики задачи выступает такой показатель, как R^2

$$R^2 = 1 - SS_{res} / SS_{tot}$$

SS_{res} - сумма квадратов остаточных ошибок.
SS_{tot} - общая сумма ошибок.

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, в праве 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 конкурса, попадает в чёрный список с дальнейшим отстранением от участия в чемпионатах до конца сезона.