цифровой прорыв
сезон: ИИ

РОССИЯ – СТРАНА ВОЗМОЖНОСТЕЙ

Задача «Радар тенденций новостных статей»

Введение

Фраза “в нужный момент в нужном месте” хорошо описывает положение авторских текстов. Иногда качественно написанная статья проходит мимо своей потенциальной аудитории из-за более актуальных тем дня или неудачного заголовка.

Хорошо, что алгоритмы ИИ активно продвинулись в анализе текста и способны в автоматическом режиме анализировать и вычленять тенденции, а имея большой набор данных, можно научиться предсказывать их наперед.

Разумеется, что есть такие общемировые темы, которые невозможно предсказать, как, например, пандемия “коронавируса” или застрявший контейнеровоз, тем не менее исследования специалистов показывают, что в обществе есть тенденции, которые приходят и уходят в фиксированный временной период.

Условие задачи

У компании РБК довольно взрослая аудитория, которую она хочет расширить за счет добавления статей на актуальные темы. Для этого вам нужно проанализировать лучшие новости российских СМИ и научиться предсказывать их популярность. Ожидается, что для этого будут использованы NLP модели.

Описание входных значений

- train.csv — файл для обучения, содержит 7000 строчек, каждая из которых представляет из себя одну новостную статью
- test.csv — файл, содержащий 3000 строк, для предсказания
- sample_solution.csv — пример файла для отправки

В наборе данных присутствует уникальных 11 строк:

- document id - идентификатор
- title - заголовок статьи
- publish_date - время публикации
- session - номер сессии
- authors - код автора
- views - количество просмотров

- depth - объем прочитанного материала
- full_reads percent - процент читателей полностью прочитавших статью
- ctr - показатель кликабельности
- category - категория статьи
- tags - ключевые слова в статье

На что стоит обратить внимание

Разрешено использование предобученных моделей. Платные модели или "приватные" модели использовать не разрешается.

Метрика

Цель модели участников — предсказать 3 численные характеристики, которые в полной мере показывают популярность статьи: views, full reads percent, depth.

Для оценки качества решения используется метрика R2.

$$result = 0.4 * R2_{views} + 0.3 * R2_{full\ reads\ percent} + 0.3 * R2_{depth}$$