Привет кандидат на вакансию!

Основная задача новостного мониторинга - обрабатывать входящий поток новостей, находя в них интересные пользователям события. В задании предлагается построить модель для выявления в новости информации о задержке ввода некоторого объекта в эксплуатацию.

Рекомендуем для начала выбрать какой-нибудь простой интерпретируемый бейзлайн, провести EDA, поискать зависимости в данных. Возможно, в описываемой задаче заработает unsupervised-подход (но это не точно).

В задаче есть следующие **входные данные**:

- Обучающая выборка (train_data.csv)
- Новостной поток за несколько дней (test_data.csv)

Формулировка задачи:

Обучить модель, которая будет искать в потоке новости с информацией о событии. Выбрать метрики, которые с вашей точки зрения будут наиболее релевантны для решаемой бизнес-задачи. Обосновать выбор метрик и модели.

Формат результатов:

- Файл с кодом и описанием алгоритма поиска релевантных новостей. Можно присылать в виде архива(zip формат) или постить на git. В любом случае, нужен readme файл, прочитав который, можно будет воспроизвести процесс обучения модели;
- Развёрнутое описание результатов, метрик, возможно интерпретация предсказаний модели (если модель интерпретируемая). Презентация результатов имеет тоже большой вес. К примеру, если получится построить относительно простую модель, которая будет выбивать чуть худшие метрики, чем более сложная модель, но при этом будет интерпретируемой/работать быстрее/потреблять меньше ресурсов/запускаться на CPU, возможно стоит остановить свой выбор на ней и развернуто описать как вы пришли к такому выбору (опять же, зависит от конкретных результатов, сильно гнаться за скоростью инференса не стоит);
- Файл test_data.csv с добавленным полем, содержащим вероятность принадлежности новости к положительному классу.

Пожелания:

Не забываем, что сейчас 2024 год с соответствующими технологиями.