

Х-MAS HACK

15-18 декабря 2022

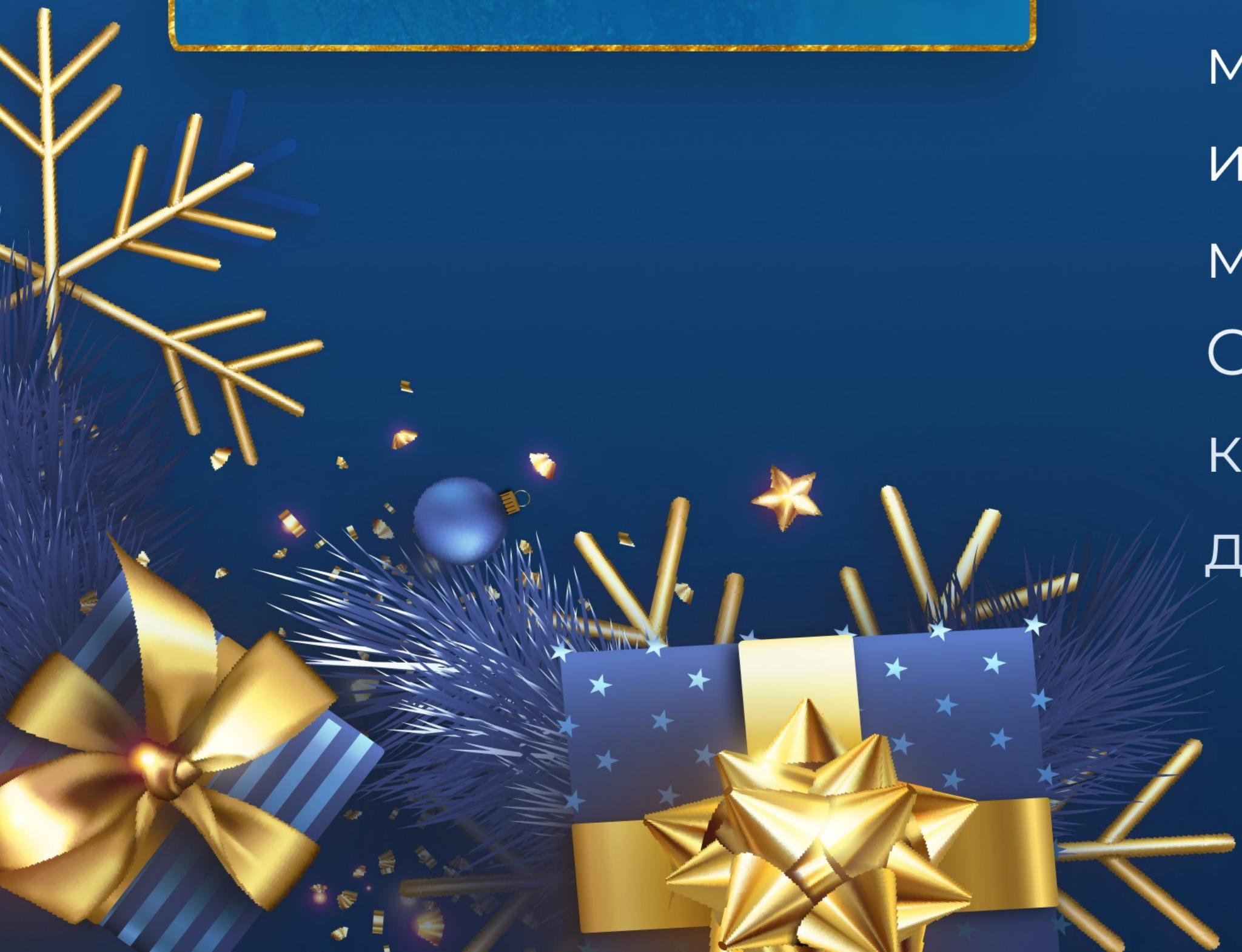
КЕЙС №1

Определение вида договора
с интерпретацией результатов

embedika

Цель

**Разработать решение
для автоматического
определения вида
договора**



Проблема

Большой бизнес — это всегда много договоров. При этом высококвалифицированные специалисты нередко задействованы в рутинных задачах по регистрации и анализу документов.

Один из первых этапов работы с документами — маршрутизация. В зависимости от вида договора и других параметров выбирается дальнейший маршрут согласования и регламент проверки. Сроки от регистрации документа до того, как он попадает к нужному сотруднику, могут достигать 14 дней.

Вводная часть

Ваша задача

Разработать решение для автоматического определения вида договора. Решение должно принимать на вход документ в форматах doc, docx, pdf с текстовым слоем и выдавать вид договора, а также интерпретировать результаты. Интерпретация результатов предполагает наличие признаков и критериев, по которым был выбран вид договора. Успех решения в значительной степени будет определяться по качеству интерпретации результатов.

Классификация

На вход вы получаете документы в форматах doc, docx, pdf. Нужно извлечь из них тексты и по текстам сделать предсказание, к какой категории относится каждый договор.

Для оценки качества работы классификации вашего алгоритма важно сделать отчет.

Отчет можно предоставить в качестве ipython-notebook с оценкой качества либо на отложенной выборке, либо с помощью 5-fold-подхода. Дополнительным плюсом будет возможность воспроизвести решение на компьютере эксперта.

Дополнительным плюсом также будет веб-интерфейс к вашему решению с возможностью загрузки нового договора в форматах doc, docx, pdf. Без претензий на красивый дизайн. :)

Для этого можно:

- собрать docker-образ со всеми необходимыми установленными библиотеками и вашим запускаемым ipython-кодом;
- предоставить ipython-notebook со всем кодом классификации с версиями необходимых библиотек, указанных в заголовке.

Кроме этого, для воспроизводимого решения рекомендуется составить инструкцию:

- как правильно запустить решение;
- как передать файлы в решение;
- (повторно) какие версии библиотек использовать для успешного запуска.

Визуализация

Кроме возможности классификации входящих документов алгоритм должен уметь визуализировать причины принятия решения.

Визуализация допустима в виде:

- набора ключевых фраз, которые найдены в тексте документа, присущие этому классу договоров;
- отрывков документа, однозначно свидетельствующих о классе документа;
- ваши предложения в лицу экспертам принимаются. :)

Результаты визуализации будут оценены выше, если они будут содержать только правильно сопряженные фразы и предложения, без разрезов предложений или других «артефактов» обработки, мешающих понимать смысл визуализации.



Представить результат алгоритма в виде JSON-файла с полями:

- docID
 - имя документа (имя файла документа)
- class
 - предсказанный класс
- confidence_level
 - уровень уверенности модели при предсказании класса выше
- interpretation
 - текстовые отрывки в произвольном виде

Кроме этого, нужно описать выбранный вами подход и принципы работы алгоритма визуализации.

Участникам требуется:



Предлагаемые технологии

Python, Transformers, BERT, NLP, NLU, text classification, explainable ML decisions.

Данные для обучения

Участникам будут предоставлены 120 договоров с указанием их видов.

Мы специально стерли заголовки текстов и названия файлов, чтобы задача не была слишком легкой. ;)

Некоторые тексты могут относиться к нескольким классам, это нормально. Оценивайте их по доминирующей части смысла договора.

Классы текстов могут быть следующие:

- Аренда
- Поставка
- Подряд
- Услуги
- Купля-продажа

ЖДЁМ ВАС НА ОТКРЫТИИ ХАКАТОНА

Х-MAS HACK

15 ДЕКАБРЯ В 19:00