Тестовое задание для стажера на позицию «Разработчик ML»

**Задание:**

Разработать систему машинного обучения, которая по списку статически импортируемых библиотек exe файла предсказывает, является ли этот файл зловредным.

Для выполнения задания предоставляются три выборки: обучающая, валидационная и проверочная. Выборки представлены в виде tsv файлов с тремя колонками – is\_virus – является ли файл зловредным: 1=да, 0=нет; filename – имя файла для ознакомления; libs – через запятую перечисление библиотек, статически импортируемых этим файлом (мы использовали библиотеку LIEF для получения списка).

На обучающей выборке – train.tsv – следует обучать модель машинного обучения.

На валидационной выборке – val.tsv – требуется подсчитать, насколько хорошо модель справляется с файлами, которые она не видела при обучении. Характеристики требуется записать в текстовый файл validation.txt со следующим содержанием (изменив значения на ваши):

True positive: 2

False positive: 20

False negative: 18

True negative: 60

Accuracy: 0.6200

Precision: 0.0909

Recall: 0.1000

F1: 0.0952

Проверочная выборка – test.tsv – содержит только колонку libs. Для проверочной выборки требуется создать файл prediction.txt, в котором для каждой строки файла проверочной выборки будет содержаться один символ: либо 1 если модель предсказывает этот файл как зловредный, либо 0 иначе. Первая строка файла, соответствующая заголовку проверочной выборки должна быть “prediction”:

prediction

0

0

1

( … много строк пропущено … )

0

1

0

**Дополнительное задание:**

Для проверочной выборки создайте ещё один файл, explain.txt, где для каждой строки файла проверочной выборки будет содержаться причина (в свободном человеко-читаемом формате) по которой модель посчитала этот файл зловредным. Строки для не зловредных файлов должны быть пустыми.

**Требования:**

В качестве решения принимается: скрипты на Python 3.x, файлы validation.txt, prediction.txt, и, опционально explain.txt. Скрипты должен создавать такие же файлы в результате работы.

Должно присутствовать три скрипта: train.py, выполняющий обучение из обучающей выборки и записывающий модель в файл, validate.py, читающий модель из файла и создающий файл validation.txt, и predict.py, читающий модель из файла и создающий файлы prediction.txt и опционально, explain.txt. Каждый из трех должен запускаться без аргументов командной строки.

**Дополнительная информация:**

Предложенный вариант определения зловредных файлов очень далек от идеала и поэтому при оценивании в разумных пределах мы не будем обращать много внимания на то, насколько хорошо модель действительно работает.

При написании кода, обращайте особое внимание на то, чтобы его было легко читать. Желательно чтобы были комментарии для значимых вещей (“считаем суммарную вероятность N наиболее вероятных.”), и чтобы не было комментариев для очевидных (“эта строка увеличивает счетчик на 1”). Старайтесь писать простой и короткий код.

Использовать дополнительные файлы помимо перечисленных можно. Использовать сторонние библиотеки можно если эти библиотеки возможно установить через pip – в таком случае прикладывайте файл requirements.txt в стандартном формате pip.

На то, какую форму машинного обучения вы будете использовать, ограничения не накладываются.

Не присылайте двоичные исполняемые файлы и ваши виртуальные окружения с двоичными исполняемыми файлами в них.

Максимальное время на выполнение задания: 1 неделя