Задача на DL

Обучить модель бинарного классификатора для определения подобия строк содержащих код KKS.

Дополнительные условия:

расстояние между:

10ABC и 20ABC,

...DDD.1.AAA... и ...DDD.2.AAA...,

...0001 и ...0002

должно быть максимальным

т.е. различия в строках где в коде есть отличие даже на одну цифру должно восприниматься классификатором как абсолютно разные строчки.

Напротив, если различие в коде нет, но есть различия в самой строчке не более 15% то данные строчки считаются идентичными.

Требований к визуализации нет.

Данные на вход – таблица в формате csv, столбцы (Исходные данные, референс).

Исходные данные – строки, содержащие код KKS на сравнение.

Референс – строки, содержащие код KKS для сравнения.

Данные выходные: та же таблица с дополнительным столбцом где указан результат работы классификатора (0 или 1).

Исходные данные сгенерировать по шаблону дополнительных условий:

Пример:

* «Система ЖД обвязки контура 11UDT станции «Первая»», строка для сравнения, например, «Обвязка контура станции «Первая» система 1UDT» считаются идентичными.
* «Лестничный пролет 11UDT15BQ005» строка для сравнения «Лестничный пролет 11UDT16BQ005» считаются не идентичными (различия помечены красным).
* Сгенерировать в количестве достаточном для решения задачи.

В ответе предоставить датасет, обученную модель, файл csv выходных данных.

\*сначала файл csv и датасет.

Подсказка, задача STS, использовать не большие модели.