**Лабораторная работа №4**

**Генерация текстов на основе LSTM**

**Задание 1.** Загрузите текст из произведений Ницше ('nietzsche.txt', origin='https://s3.amazonaws.com/text-datasets/nietzsche.txt').

Выведете следующее:

А) длину всего корпуса;

Б) количество предложений;

В) сколько всего символов используется?

**Задание 2.** Сократите текст наполовину избыточными последовательностями символов maxlen

maxlen = 40

step = 3

sentences = []

next\_chars = []

for i in range(0, len(text) - maxlen, step):

    sentences.append(text[i: i + maxlen])

    next\_chars.append(text[i + maxlen])

print('nb sequences:', len(sentences))

print('Vectorization...')

x = np.zeros((len(sentences), maxlen, len(chars)), dtype=np.bool)

y = np.zeros((len(sentences), len(chars)), dtype=np.bool)

for i, sentence in enumerate(sentences):

    for t, char in enumerate(sentence):

        x[i, t, char\_indices[char]] = 1

    y[i, char\_indices[next\_chars[i]]] = 1

**Задание 3.** Создайте модель LSTM для генерации текста

А) Напишите вспомогательную функцию для выборки индекса из массива вероятностей

Б) Напишите функцию, которая будет вызываться в конце каждой эпохи и печатать сгенерированный текст

В) Запустите модель на обучение

Имейте ввиду, что требуется не менее 20 эпох, прежде чем сгенерированный текст начнет звучать связно. Рекомендуется запускать этот скрипт на графическом процессоре, так как рекуррентные сети требуют довольно больших вычислительных затрат.

Г) Проверьте работу модели в онлайн режиме.

Для примера смотрите файл lstm\_генерация.ipynb

**Задание 4.** Создайте самостоятельно генерацию текста для **РУССКОЯЗЫЧНОГО НАБОРА** глав Wikibooks.

Полный текст Wikibooks содержит более 270000 глав на 12 языках https://www.kaggle.com/datasets/dhruvildave/wikibooks-dataset/data