МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра МО ЭВМ

ОТЧЕТ

по лабораторной работе №6

по дисциплине «Искусственные нейронные сети»

Тема: Прогноз успеха фильмов по обзорам

Студент гр. 8383 _____ Ларин А.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург

2021

Цель работы

Прогноз успеха фильмов по обзорам (Predict Sentiment From Movie Reviews)

Задание

- Ознакомиться с задачей классификации
- Изучить способы представления текста для передачи в ИНС
- Достигнуть точность прогноза не менее 95%

Требования

1. Построить и обучить нейронную сеть для обработки текста
2. Исследовать результаты при различном размере вектора представления текста
3. Написать функцию, которая позволяет ввести пользовательский текст (в отчете привести пример работы сети на пользовательском тексте)

Выполнение

В работе требуется провести анализ настроений.
С помощью анализа настроений можно определить отношение (например, настроение) человека к тексту, взаимодействию или событию. Поэтому сентимент-анализ относится к области обработки естественного языка, в которой смысл текста должен быть расшифрован для извлечения из него тональности и настроений.
Настроения подразделяются на положительные и отрицательные. Таким образом задача сводится к задаче бинарной классификации.

Использован датасет IMDb . Он состоит из 50 000 обзоров фильмов от пользователей, помеченных как положительные (1) и отрицательные (0).

- Рецензии предварительно обрабатываются, и каждая из них кодируется последовательностью индексов слов в виде целых чисел.
- Слова в обзорах индексируются по их общей частоте появления в датасете. Например, целое число «2» кодирует второе наиболее частое используемое слово.
- 50 000 обзоров разделены на два набора: 25 000 для обучения и 25 000 для тестирования.

Датасет был создан исследователями Стэнфордского университета и представлен в статье 2011 года, в котором достигнутая точность предсказаний была равна 88,89%.

Был импортирован датасет. Информация была векторизирована с размерностью вектора 10000 и разделена на обучающую и тестировочную выборку.

Затем была создана модель следующего вида:

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_4 (Dense)              (None, 50)                500050
_________________________________________________________________
dropout_2 (Dropout)          (None, 50)                0
_________________________________________________________________
dense_5 (Dense)              (None, 50)                2550
_________________________________________________________________
dropout_3 (Dropout)          (None, 50)                0
_________________________________________________________________
dense_6 (Dense)              (None, 50)                2550
_________________________________________________________________
dense_7 (Dense)              (None, 1)                 51
=================================================================
Total params: 505,201
Trainable params: 505,201
Non-trainable params: 0
```

Затем она скомпилирована с оптимизатором adam, функцией потерь binary_crossentropy (т. к. решается задача бинарной классификации) и метрикой точности(accuracy)

Модель натренирована на 40000 обзоров и проверена на 10000
По результатам точность составила 0.893

Далее модель была усложнена

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_8 (Dense)              (None, 5000)              50005000
_________________________________________________________________
dropout_6 (Dropout)          (None, 5000)              0
_________________________________________________________________
dense_9 (Dense)              (None, 5000)              25005000
_________________________________________________________________
dropout_7 (Dropout)          (None, 5000)              0
_________________________________________________________________
dense_10 (Dense)             (None, 1000)              5001000
_________________________________________________________________
dropout_8 (Dropout)          (None, 1000)              0
_________________________________________________________________
dense_11 (Dense)             (None, 200)               200200
_________________________________________________________________
dropout_9 (Dropout)          (None, 200)               0
_________________________________________________________________
dense_12 (Dense)             (None, 40)                8040
_________________________________________________________________
dropout_10 (Dropout)         (None, 40)                0
_________________________________________________________________
dense_13 (Dense)             (None, 8)                 328
_________________________________________________________________
dense_14 (Dense)             (None, 1)                 9
=================================================================
Total params: 80,219,577
Trainable params: 80,219,577
Non-trainable params: 0
```

По результатам обучения точность составила 0.899

В конце обучения модель сохраняется в файл model.h5

Затем был написан модуль для использования моделей на пользовательских данных. Он принимает на вход(в аргументах командной строки) путь к файлу, читает из него отзыв, прогоняет через модель и печатает результат на экран

Результаты тестирования:
1.
Входные данные:
Good stuff. Love it!
Ответ модели:
0.74286515
2.
Входные данные:
Very bad film. Didn't like it
Ответ модели:
0.39905626

Выводы.

Были изучены способы представления естественного языка для дальнейшей обработки. На основе датасета imdb обучена модель для классификации отзывов на фильмы. По результатам получена точность 0.899
Был написан модуль для прогона модели по пользовательским данвным. С его помощью модель была протестирована. Для положительного отзыва получен ответ 0.743, для отрицательного 0.399. Результаты тестов удовлетворительны