Министерство образования Республики Беларусь

Учреждение образования

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Факультет Компьютерных сетей и систем

Кафедра Информатики

ЛАБОРАТОРНАЯ РАБОТА №7

« Рекуррентные нейронные сети для анализа текста»

Магистрант: Проверил:

гр. 956241 Заливако С. С.

Шуба И.А.

Минск, 2020

**ХОД РАБОТЫ**

**Задание.**

**Данные:** Набор данных для предсказания оценок для отзывов, собранных с сайта imdb.com, который состоит из 50,000 отзывов в виде текстовых файлов. Отзывы разделены на положительные (25,000) и отрицательные (25,000). Данные предварительно токенизированы по принципу “мешка слов”, индексы слов можно взять из словаря (imdb.vocab). Обучающая выборка включает в себя 12,500 положительных и 12,500 отрицательных отзывов, контрольная выборка также содержит 12,500 положительных и 12,500 отрицательных отзывов, а также. Данные можно скачать по ссылке<https://ai.stanford.edu/~amaas/data/sentiment/>

**Задание 1.**

Загрузите данные. Преобразуйте текстовые файлы во внутренние структуры данных, которые используют индексы вместо слов.

**Задание 2.**

Реализуйте и обучите двунаправленную рекуррентную сеть (LSTM или GRU). Какого качества классификации удалось достичь?

**Задание 3.**

Используйте индексы слов и их различное внутреннее представление (word2vec, glove). Как влияет данное преобразование на качество классификации?

**Задание 4.**

Поэкспериментируйте со структурой сети (добавьте больше рекуррентных, полносвязных или сверточных слоев). Как это повлияло на качество классификации?

**Задание 5.**

Используйте предобученную рекуррентную нейронную сеть (например, DeepMoji или что-то подобное). Какой максимальный результат удалось получить на контрольной выборке?

**Результат выполнения:**

**Задание 1.** Загрузите данные. Преобразуйте текстовые файлы во внутренние структуры данных, которые используют индексы вместо слов.

Укажем путь и загрузим датасет.



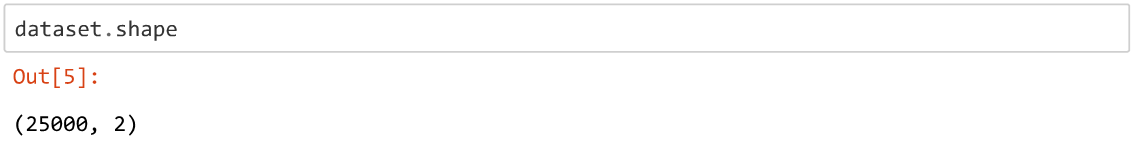


Рисунок 1 – Загрузка данных

Пример одного отзыва.

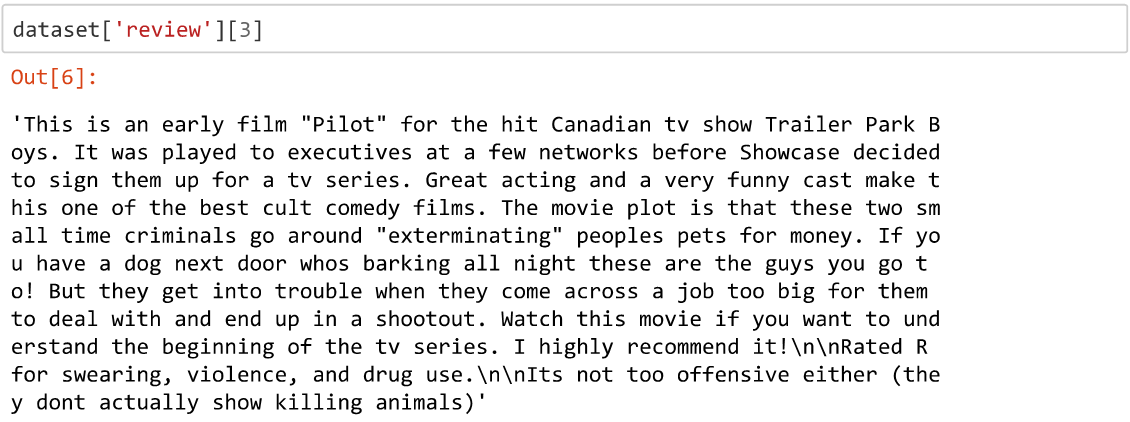


Рисунок 2 – Пример одного отзыва

Напишем функцию для обработки текста.

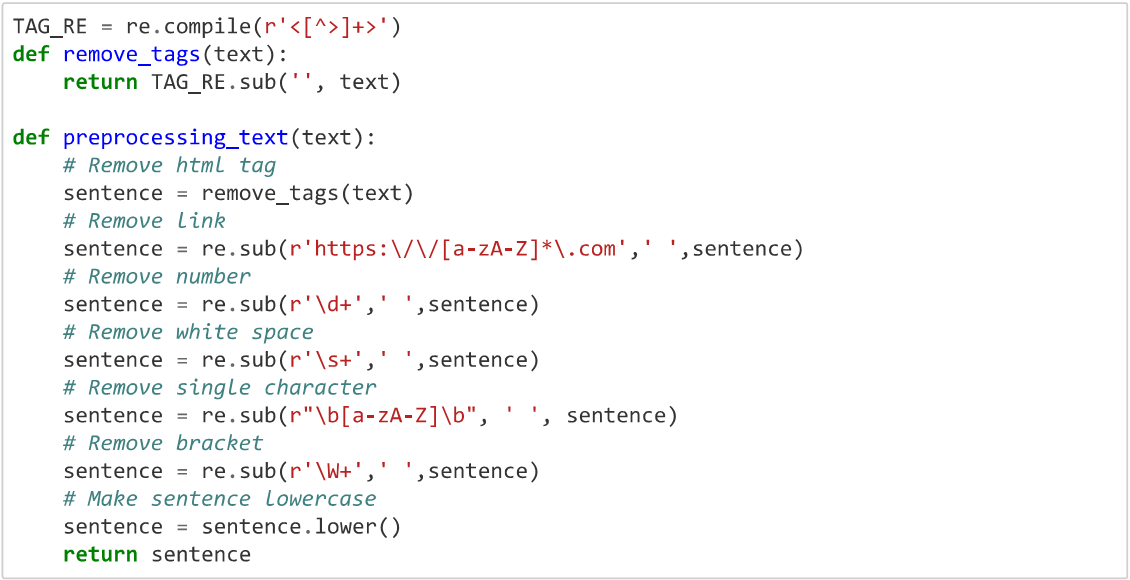


Рисунок 3 – Функция для обработки текста

Выполним препроцессинг и посмотрим на результат обработки.

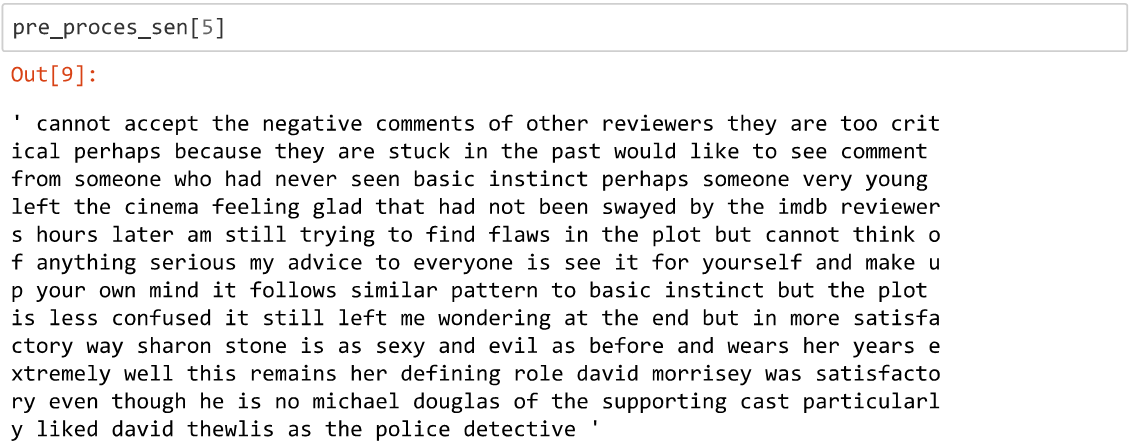
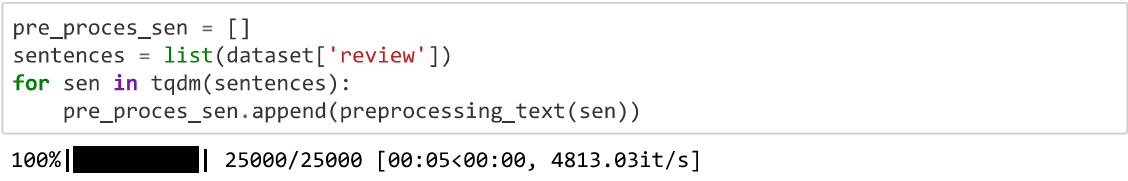


Рисунок 4 – Препроцессинг текста и отображение примера

после обработки

Импортируем необходимые библиотеки для удаления «stopwords».

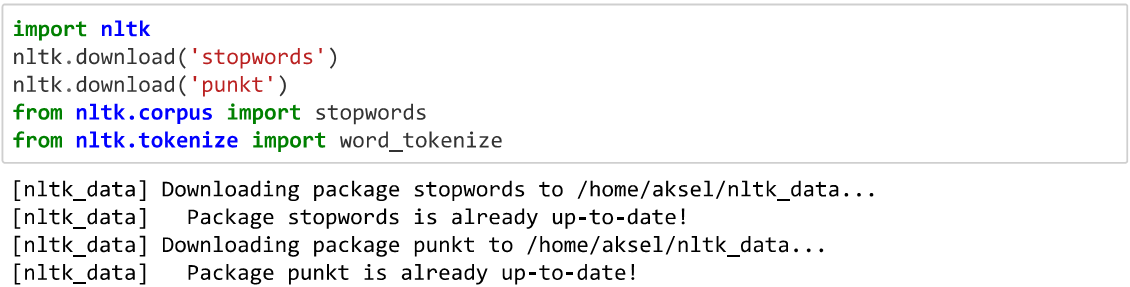




Рисунок 5 – Импорт необходимых библиотек для обработки текста

Из отзывов удалим «stopwords».

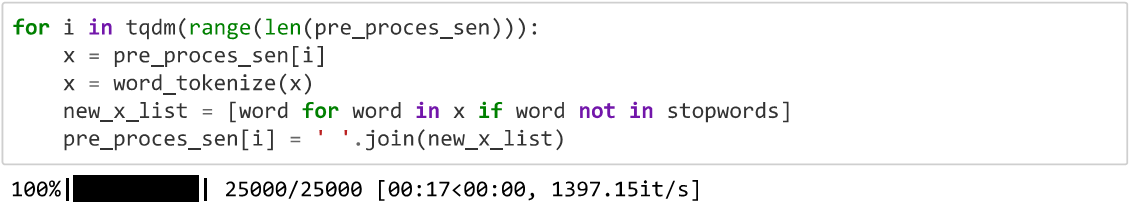


Рисунок 6 – Удаление «stopwords»

Пример отзыва после обработки.

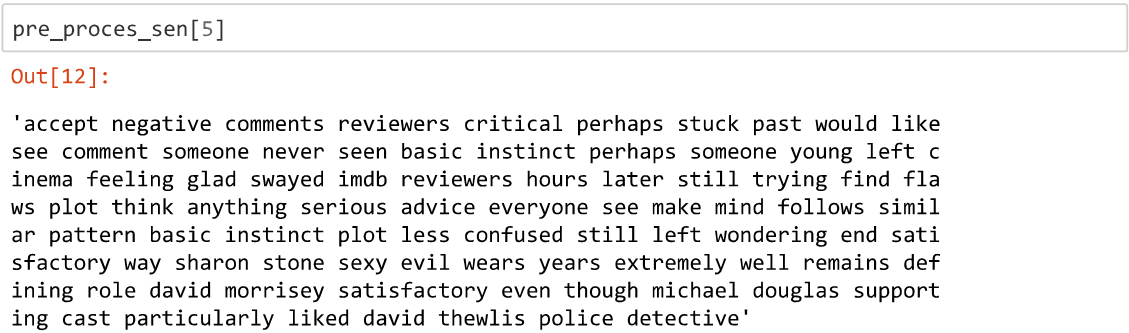


Рисунок 6 –Примера отзыва после обработки

Рисунок 7 – Замена исходного датасета на обработанный

Разделение подготовленного датасета на контрольную, валидационную и тренировочную выборки.

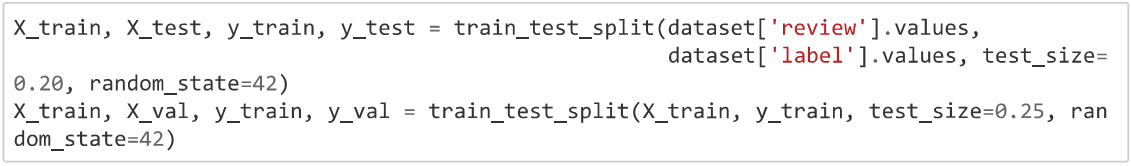
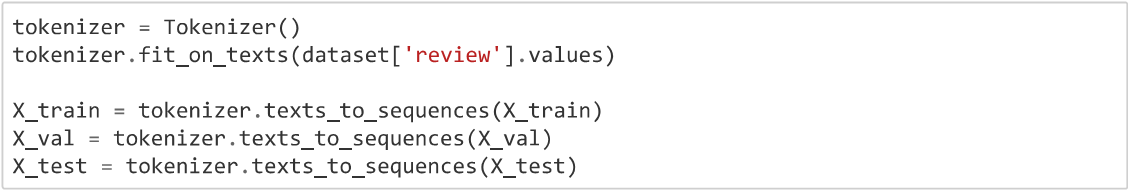


Рисунок 8 – Разделение датасета

Токенизация и создание последовательностей.



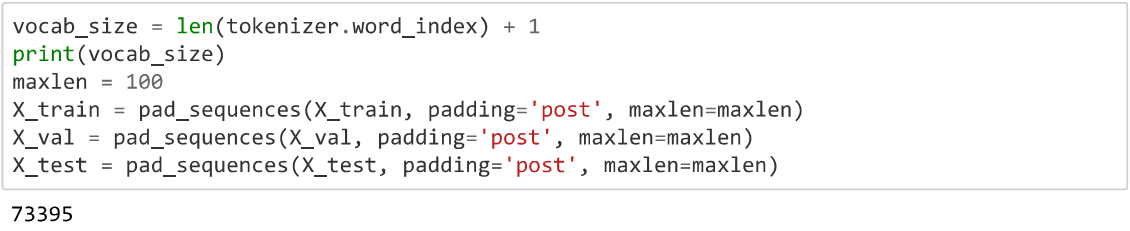


Рисунок 9 – Токенизация и создание последовательностей

**Задание 2.** Реализуйте и обучите двунаправленную рекуррентную сеть (LSTM или GRU). Какого качества классификации удалось достичь?

Реализуем двунаправленную рекуррентную сеть с помощью слоя Biderectional.

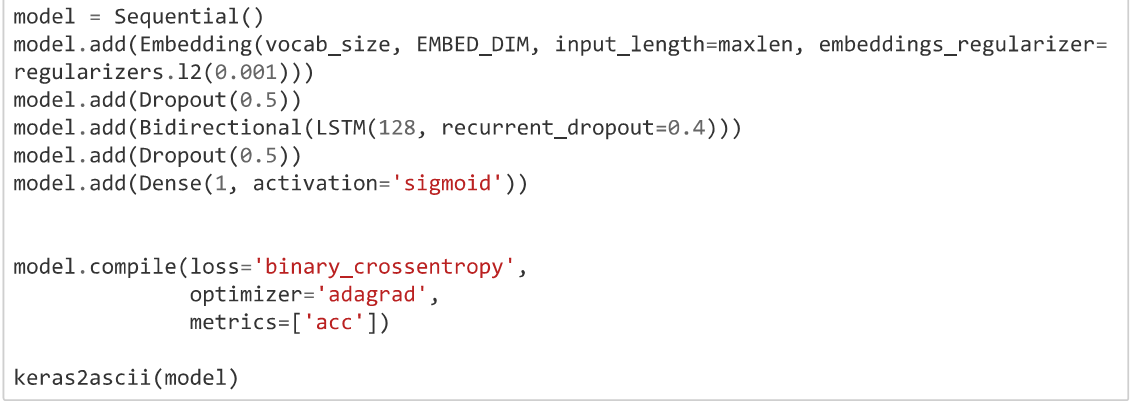


Рисунок 10 – Реализация двунаправленной рекуррентной сети

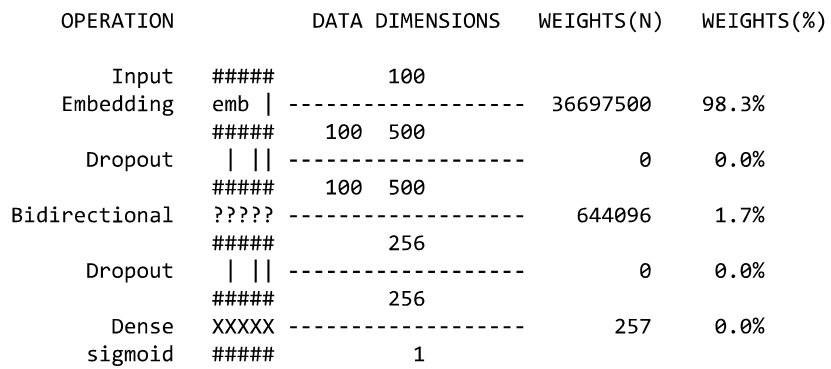
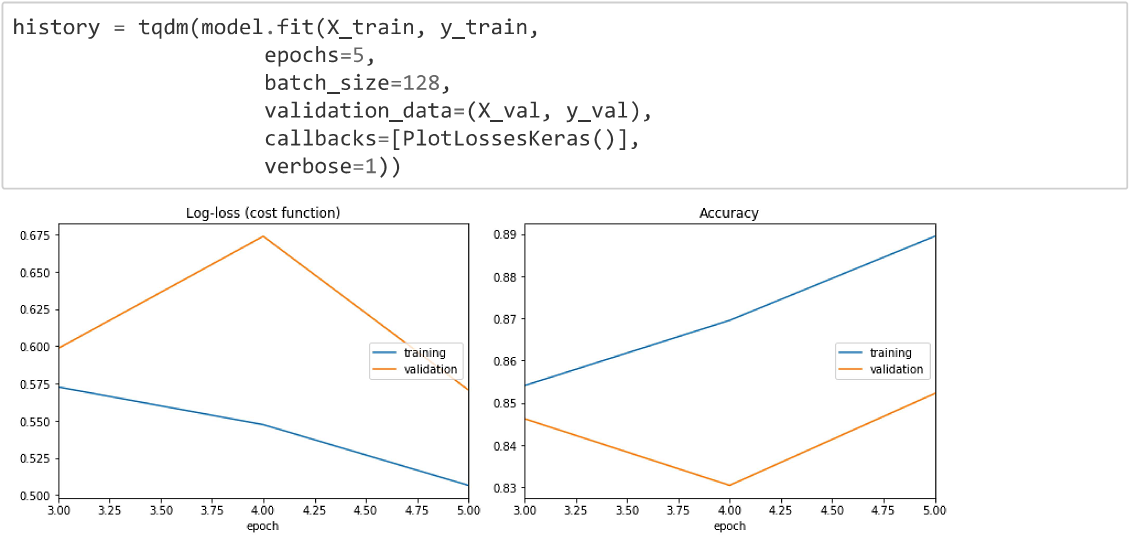


Рисунок 11 – Схема двунаправленной рекуррентной сети

Обучение будет осуществляться с небольшим числом эпох, из-за специфики рекуррентых нейронных сетей.



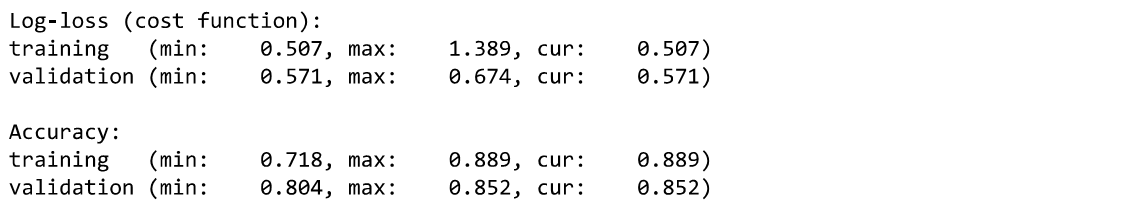


Рисунок 12 – Результаты обучения

Применив всего лишь 5 эпох, были получены достаточно неоднозначные графики. Максимальная точность на валидационных данных составила 0,852, а минимальный log-loss – 0,571

.

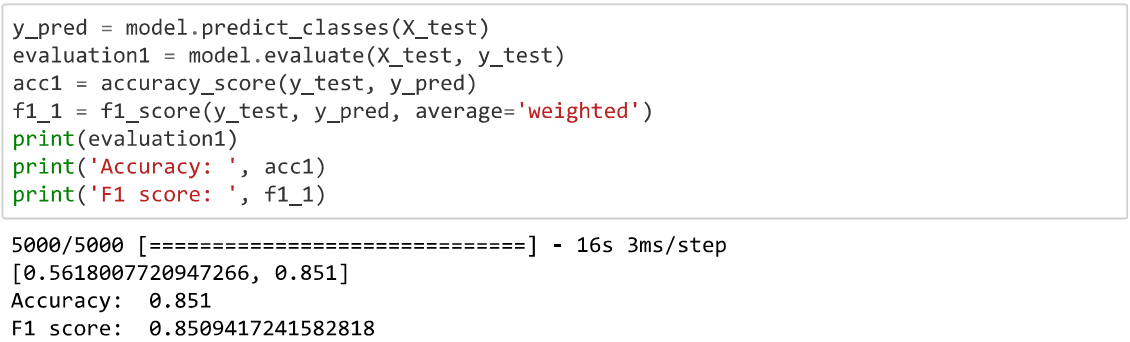


Рисунок 13 – Результаты модели на контрольной выборке

Таким образом, точность модели на контрольной выборке составила 0,851, а log-loss 0,561.

**Задание 3.** Используйте индексы слов и их различное внутреннее представление (word2vec, glove). Как влияет данное преобразование на качество классификации?

Используем word2vec в качестве другого внутреннего представления отзывов.

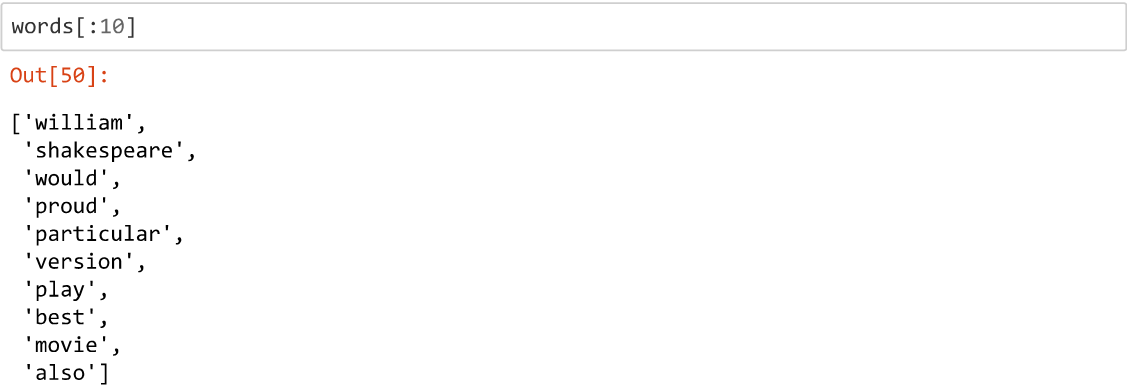
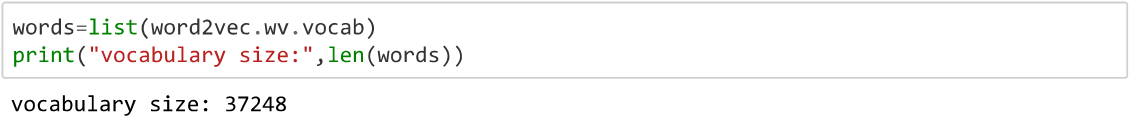
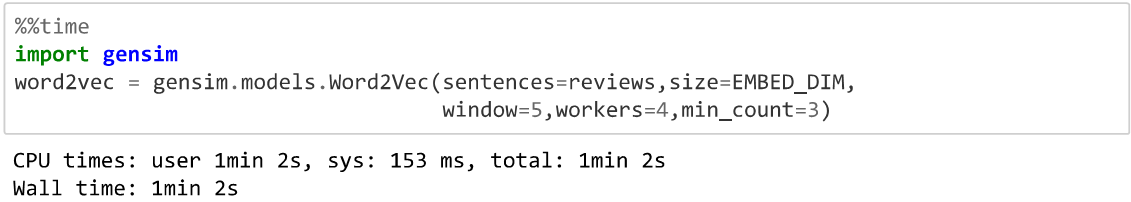
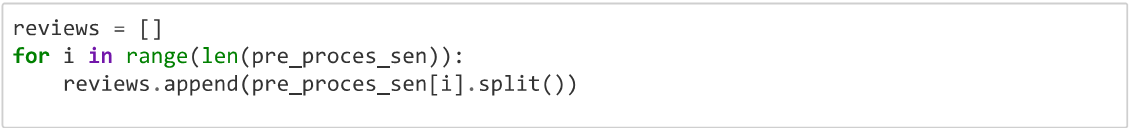


Рисунок 14 – Использование word2vec

Заполним веса, полученные с word2vec.

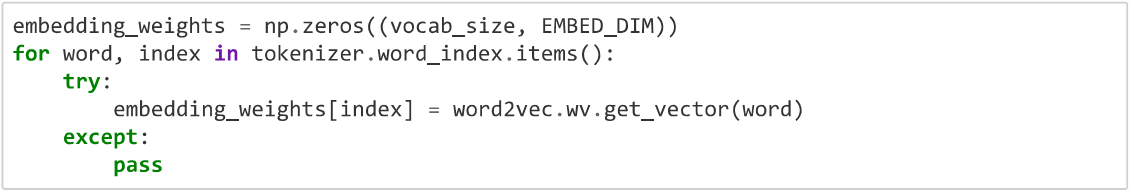


Рисунок 15 – Заполнение весов

В той же самой сети дополним поле weights в слое Embedding полученными весами с word2vec.

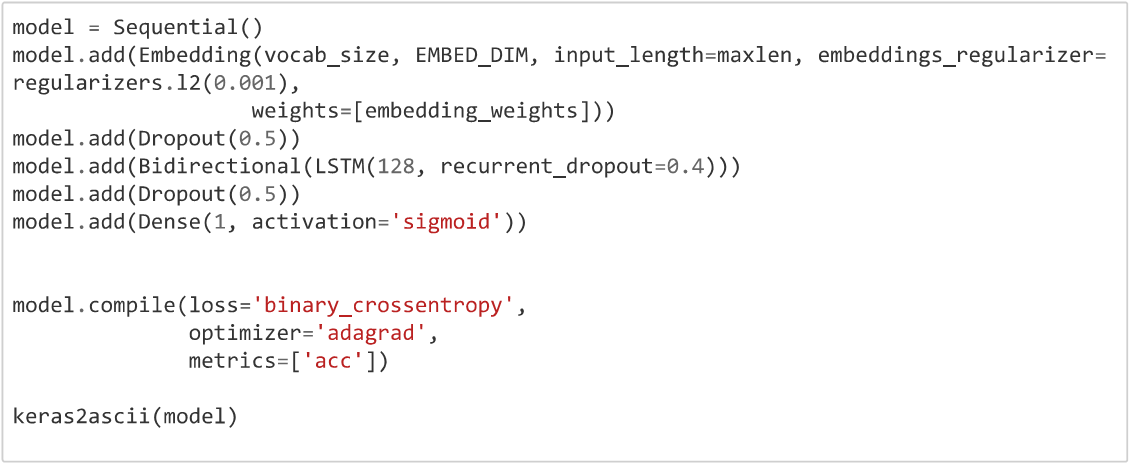


Рисунок 16 – Реализация двунаправленной рекуррентной

сети с добавлением полученных весов

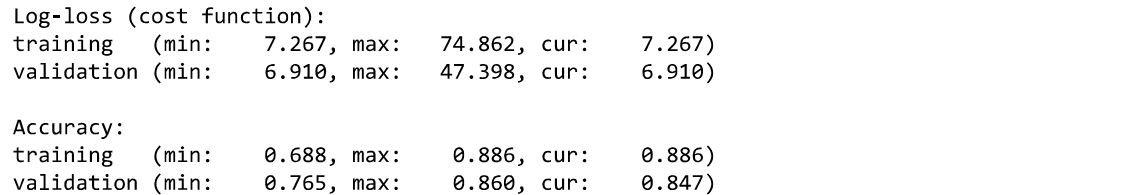
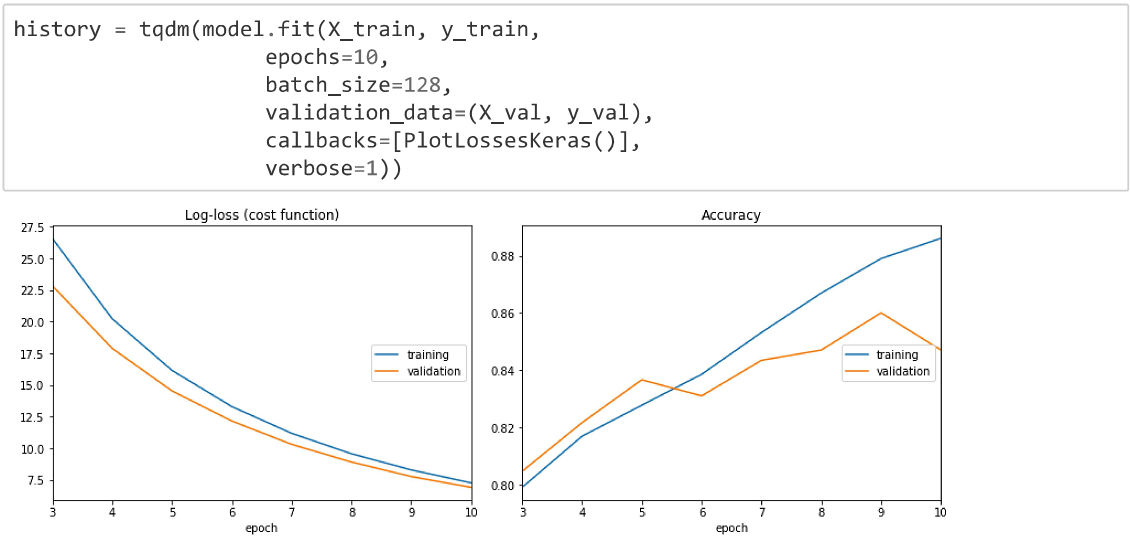


Рисунок 17 – Результаты обучения

Как видно из графиков, максимальная точность на валидационных данных составила 0,86, а минимальный log-loss – 6,910. При данном представлении слов, достаточно хорошо себя ведет log-loss. Точность модели на валидационной и тренировочной выборке ведет себя немного неоднозначно.Уже после 5 эпохи начинается заметное расхождение с точностью на тренировочных данных, однако расхождение не такое значительное, какое было в предыдущем задании.

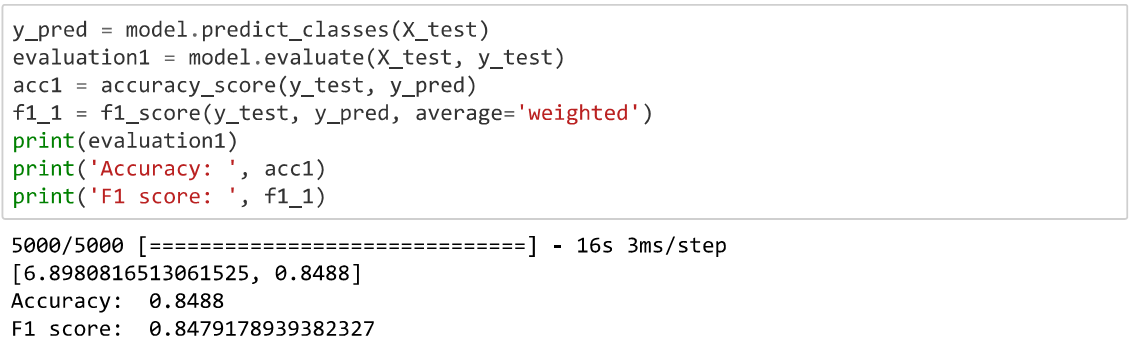


Рисунок 18 – Результаты модели на контрольной выборке

Таким образом, точность модели на контрольной выборке составила 0,848, а log-loss 6,89, хоть эти результаты и хуже, чем на предыдущей модели, однако линии обучения ведут себя более стабильней, что является более важным при оценке адекватности модели.

**Задание 4.** Поэкспериментируйте со структурой сети (добавьте больше рекуррентных, полносвязных или сверточных слоев). Как это повлияло на качество классификации?

Добавим сверточный слой, и слой пулинга.

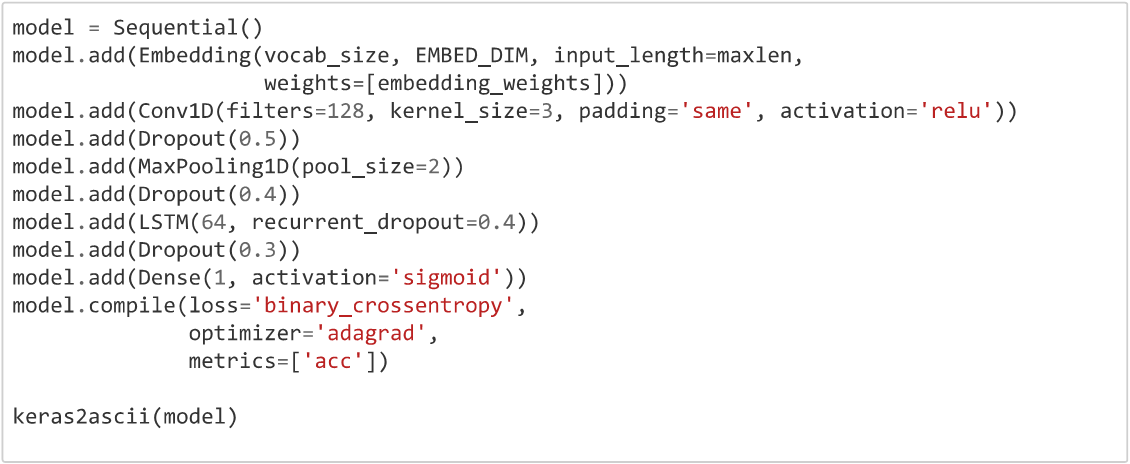


Рисунок 19 – Реализация рекуррентной нейронной сети с добавлением сверточного слоя и слоя пуллинга

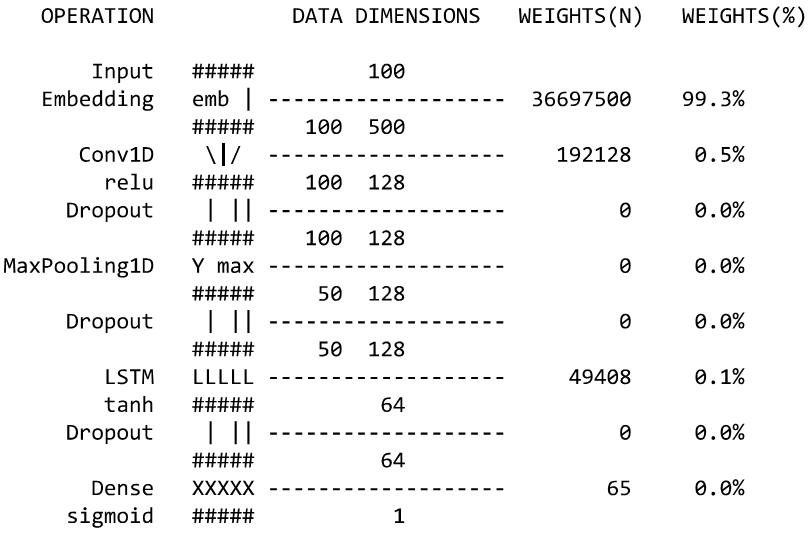


Рисунок 20 – Схема рекуррентной нейронной сети с добавлением сверточного слоя и слоя пуллинга

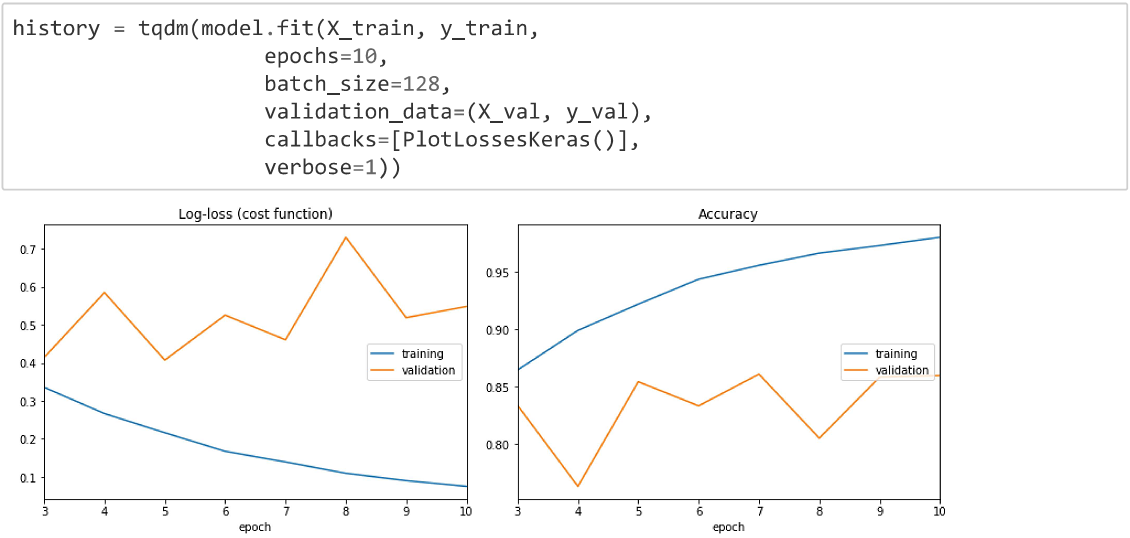
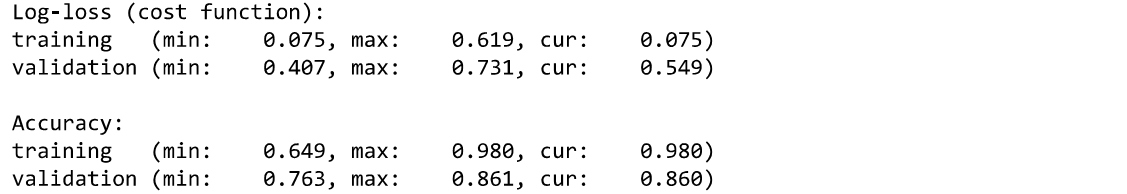
 

Рисунок 21 – Обучение рекуррентной нейронной сети

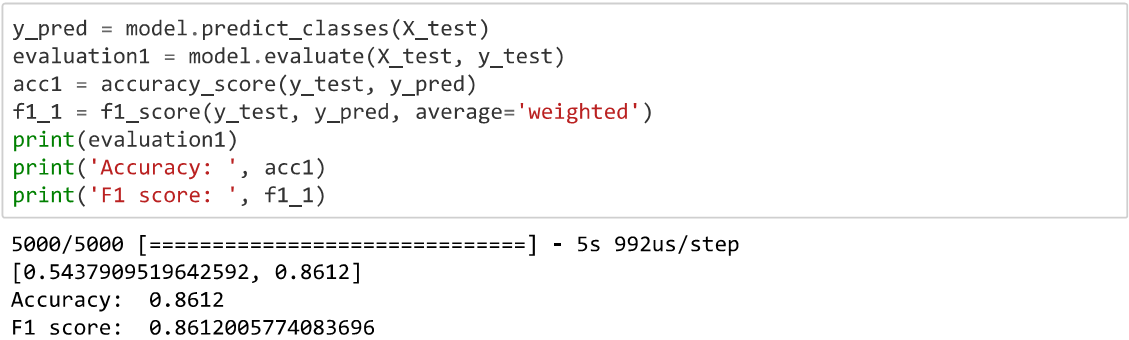


Рисунок 22 – Результаты модели на контрольной выборке

Таким образом, точность модели на контрольной выборке составила 0,86, а log-loss 0,54, хоть эти результаты и лучше, чем на предыдущей модели, однако линии обучения ведут себя нестабильно.

**Задание 5.** Используйте предобученную рекуррентную нейронную сеть (например, DeepMoji или что-то подобное). Какой максимальный результат удалось получить на контрольной выборке?

Используем предобученную рекуррентную нейронную сеть –DeepMoji.

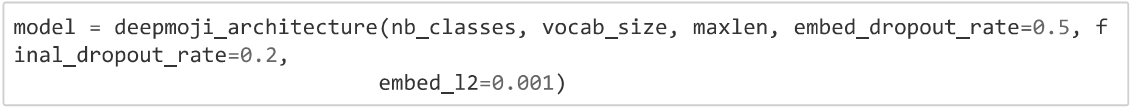




Рисунок 23 – Импорт и реализация предобученной модели DeepMoji

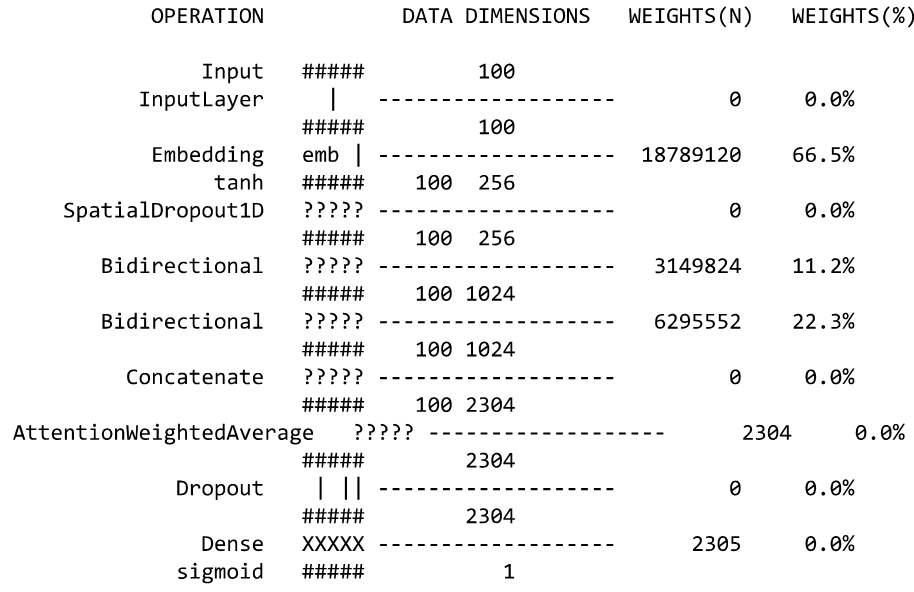


Рисунок 24 – Схема предобученной нейронной сети DeepMoji

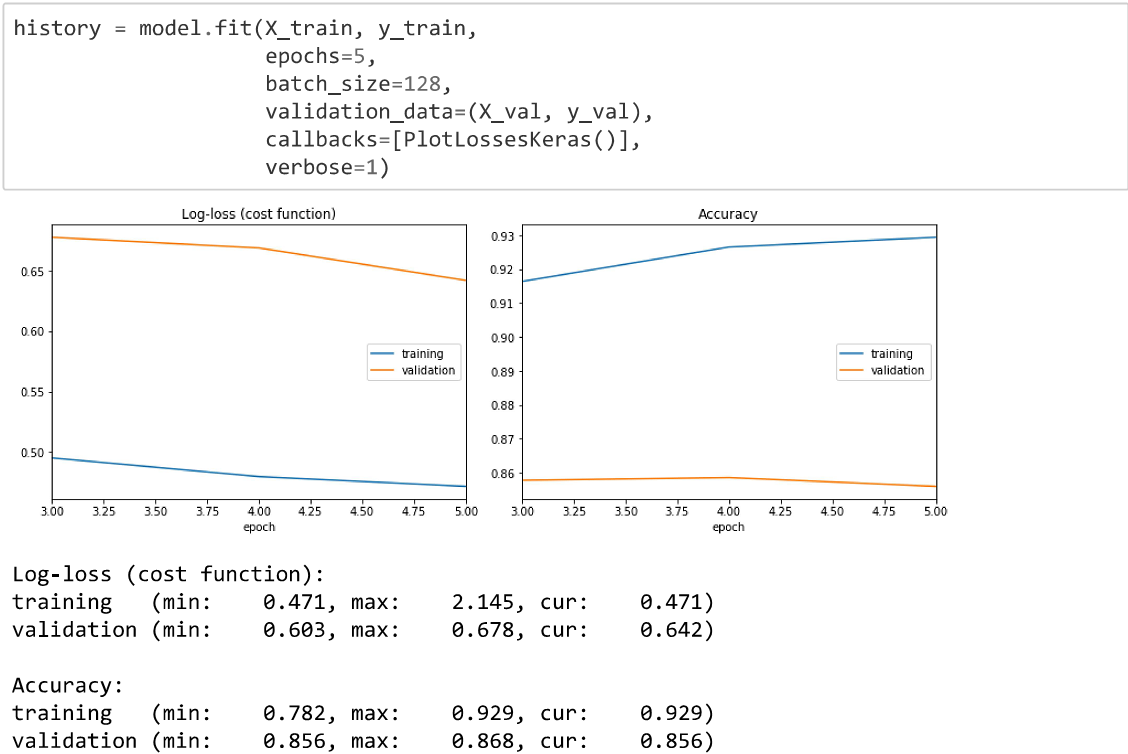


Рисунок 25 – Обучение рекуррентной нейронной сети

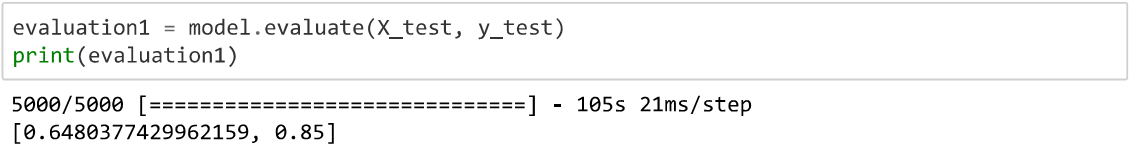


Рисунок 26 – Результаты модели на контрольной выборке

Таким образом, точность модели на контрольной выборке составила 0,85, а log-loss 0,64, хоть эти результаты и хуже, чем на предыдущей модели, однако линии обучения ведут себя более стабильней, что является более важным при оценке адекватности модели. Также отметим, что линии точности и log-loss не сходятся, а напротив идут параллельно друг другу. Можно предположить, что сети не хватает какого-то толчка выбраться из локального минимума функции, поэтому точность модели никак не изменяется с увеличением эпох.

**Вывод:**

В ходе выполнения лабораторной работы были изучены подходы для работы с текстом. Были реализованы различные представления текста: токенизация, создание «мешка слов», а также word2vec. Были опробованы рекуррентные архитектуры нейронных сетей, а также их комбинации с о сверточными сетями, а также со слоями, выполняющими функцию пуллинга. Так же была опробована предобученная нейронная сеть DeepMoji, которая применяется, для разделения настроения. В результате наилучшая точность модели была достигнута для контрольной выборки – 0,86 в реализации с комбинацией различных слоев.