Отчёт о второй лабораторной работе по машинному обучению

Команда : broken GPT

Состав: Марьевичев Николай, Голованов Габриэль, Семин Никита, Малыгин Андрей, Андреева Арина

Первый пункт:

Для начала были определены языки, которые присутствуют в данном датасете, далее для данных языков были скачаны словари стоп слов для данных языков. Далее было создано два датасета один был очищен при помощи функции cleanTextFromStopwords8Punc от стоп слов и пунктуации, другой при помощи функции cleanTextFromStopwords8Punc8Dig был очищен от стоп слов, пунктуации и цифр. В дальнейшем использовался только второй датасет.

Второй пункт:

Во втором пункте ко второму датасету, полученному в первом пункте были применены лемматизация и стемминг, Для лемматизации использовался WordNetLemmatizer, для стемминга использовался SnowballStemmer для английского языка.

Третий пункт:

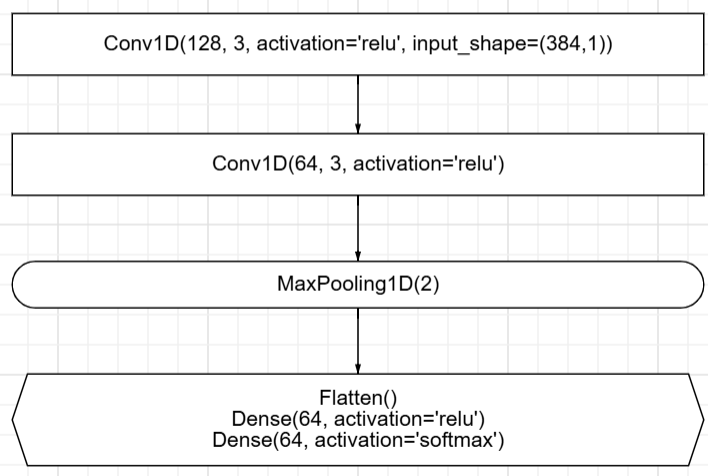
Векторизация проводилась на данных после лемматизации и стемминга, использовались tfidf vectorizer, word2vec и sentence transformers, векторизация проводилась только на description поскольку в этом поле находится больше информации по отношению к полю title, также в начале были удалены части description в которых кол-во слов меньше двух поскольку векторизовать данные из 1 слова sentence transformer не может.

Четвертый пункт:

CNN и RNN модели была реализованы при помощи библиотеки tensorflow

Для обучения использовались данные после векторизации sentence transformers на лемматизированных данных, в качестве оптимизатора использовался Adam,в качестве функции потерь для CNN была выбрана categorical crossentropy, а для RNN binary crossentropy, метриками качества в обеих моделях мы взяли accuracy и f1\_score. Далее будут приведены схемы данных моделей

CNN



RNN

