Classify the movies based on reviews

Project Report

Step 1. Data Collection:

Use crawler to get reviews of the firth 5th most popular movies from IMDB (the code is in file *imdb_reviews_Crawler.py*). The movies' name and id are in the form below.

Movie Name	Movie Id
Black Mirror: Bandersnatch	9495224
Sharknado	2724064
The Normal Heart	1684226
The Sunset Limited	1510938
Temple Grandin	1278469

Step 2. Data Preprocess:

Use nltk, sklearn (CountVectorizer) and keras (to_categorical) to make one-hot encoding for movie reviews and movie id (the code is in file *classification.py*)

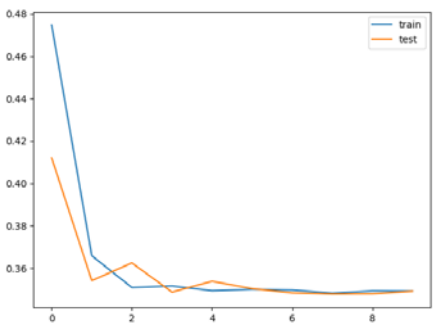
Step 3. Build the Model:

Use keras to build the model. The structure of the model is shown in the figure below:

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_2 (Embedding)      (None, None, 32)          200160
_________________________________________________________________
lstm_3 (LSTM)                (None, 32)                8320
_________________________________________________________________
dense_3 (Dense)              (None, 5)                 165
=================================================================
Total params: 208,645
Trainable params: 208,645
Non-trainable params: 0
```

Step 4. Train the Model:

Split the data into two parts: train (80%) and test (20%). Put the train data in the model and then draw the curve of loss:



We can find that the accuracy of our model is approaching to 0.7.

Step 5. Evaluation:
Put the test data into the model and we can get the Test RMSE, which is 0.336, and proves that the model has relatively high accuracy.