CSI 5340: Introduction to Deep Learning and Reinforcement Learning



**Sentiment Analysis**

**Author: David Talson**

**Student Number: 8419286**

Table of Contents

# 1.0 Introduction

# 2.0 Methodology

In this section, the various techniques used to collect, pre-process, build, and test the model is discussed.

## 2.1 Data Collection

### 2.1.1 Movie Reviews Dataset

The IMDB dataset used for this task was gotten from the Stanford AI lab [1]. This dataset is a collection of movie reviews for binary sentiment classification with ‘1’ corresponding to a positive review and ‘0’ for a negative review. The dataset is highly polar as movie ratings <= 4 are considered negative and movie ratings >= 7 are considered positive [2]. No more than 30 reviews are allowed for each movie and the a disjoint set of movies are contained in the train and test data. There are a total of 25,000 train and test reviews each [3]. Each file in the train and test dataset is named using the pattern ‘id\_rating.txt’ where id is the unique identifier of the moves, and the rating is a numerical value between 0-10. The content of the text file is the review.

### 2.1.2 Word Representation

The Global Vectors for Word Representation (GloVe) is an unsupervised learning algorithm for obtaining a vector representation of words [4]. GloVe contains a pre-trained word vectors. The vector used for this project is ‘glove.840B.300d’ which contains 840 billion tokens, 2.2 million words represented using 300 dimension vectors.

## 2.2 Data Pre-Processing

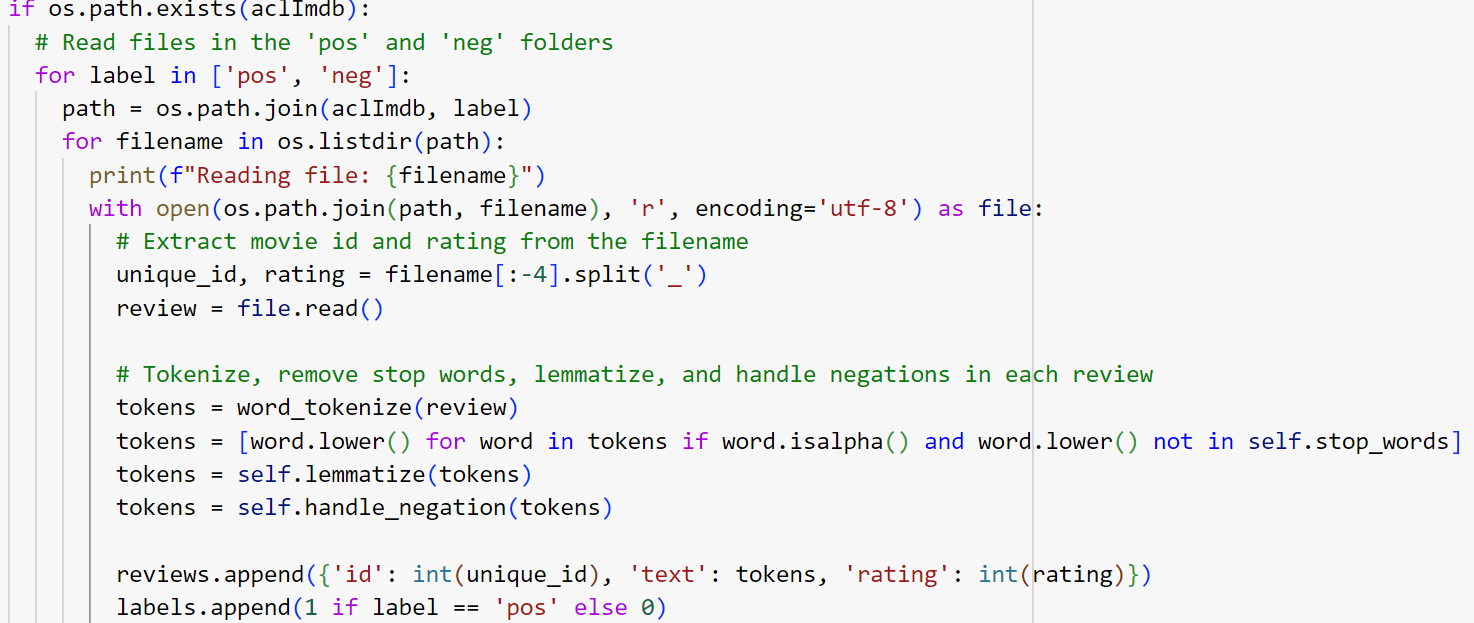
The data have to be pre-processed before being used in the model. The advantage of data pre-processing is to remove invalid data, prepare the data in a form ready to be accepted by the model and also to reduce dimensionality. Reducing dimensionality will help the model quickly learn, for example, words in a review that do not have an effect on the rating is removed.

### 2.2.1 IMDB dataset

A function to pre-process the dataset was created. The function performed the following:

* Create a dictionary of train and test data with the following keys: Review i {id:x, text: , rating:}
* Create a list of train and test labels for train and test data respectively
* Remove entries that are not alphabetic
* Remove stopwords which are words that have no significant effect on the text from the review

The figure below shows a snippet of the pre-processing code:

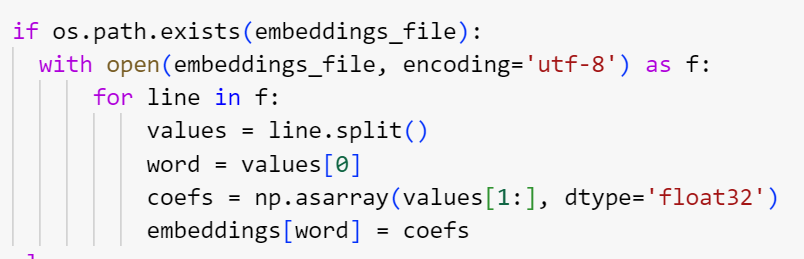


**Figure 1: Train and Test Data Preparation**

As can be seen from figure 1 above, the movie identifier and rating are extracted from the filename, the text file is then read and the review is tokenized. A list of tokens is created where non alphabetic characters and stop words are removed. The tokens list is then lemmatized and negations are handled. A reviews and labels are then combined to generate the dataset

### 2.2.2 GloVe Vectors

The GloVe embeddings was pre-processed to generate a dictionary of words and their corresponding embedding vectors. This is more of a parsing task as the vectors and words are in the text file. The figure below shows a snippet to generate the embeddings:



**Figure 2: GloVe Embeddings Pre-processing**

As can be seen from figure 2 above, the embedding dictionary is given a key value which is the word and the value is the coefficient.

### 2.2.3 Exploratory Data Analysis

Exploratory data analysis (EDA) helps to further understand the nature of the dataset. A summary of th

### 2.2.4 Vanilla Recurrent Neural Network

Vanilla RNN is a

# 3.0 Discussion

# 4.0 Conclusion

# 5.0 References

[1] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). [Learning Word Vectors for Sentiment Analysis.](https://ai.stanford.edu/~amaas/papers/wvSent_acl2011.pdf) The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)

[2] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). [Learning Word Vectors for Sentiment Analysis.](https://ai.stanford.edu/~amaas/papers/wvSent_acl2011.pdf) The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)

[3] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). [Learning Word Vectors for Sentiment Analysis.](https://ai.stanford.edu/~amaas/papers/wvSent_acl2011.pdf) The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)