Aiden Moosa
am2497
Sidney Sussex

Assignment 2 Report

Introduction

The simple bag-of-words (BoW) model produces desirable results in many NLP applications, however Le and Mikolov (2014) have shown that their predict model doc2vec usually outperforms BoW, provided that the hyperparameters chosen are near-optimal. I am looking to test their findings in the domain of movie review sentiment classification, where I will test a simple BoW support vector machine (SVM) against a doc2vec SVM.

Datasets

The dataset provided is of 100,000 movie reviews divided equally into training and testing sets, which are further subdivided into negative, positive, and unlabelled sets. I want to be able to compare my results from this assignment with the previous one, therefore I have combined the data from the training and testing sets such that I can perform 9:1 split stratified cross-validation on them as prior. To save time, I pre-processed the reviews by tokenizing and stemming them, converting them to a format recognisable to doc2vec, and then saving them to disk.

Methodology

Results

The increased length of time taken to train the BoW SVM on this much larger dataset has meant that I was unable to gather results for comparison with the doc2vec SVM. The accuracy score from the doc2vec SVM was 80.1%, which is much lower than any other score found in the previous assignment, though this is to be expected as Le and Mikolov stress in their paper the importance of tuning hyperparameters in the performance of the doc2vec model. As I did not have time to perform such tuning, this is to be expected.

Word count: 250

Aiden Moosa
am2497
Sidney Sussex

Bibliography

Maas, A. et al 2011. Learning Word Vectors for Sentiment Analysis. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. http://www.aclweb.org/anthology/P11-1015

The source code for implementation can be found at
https://github.com/AidenMoosa/Natural-Language-Processin-Task-2