DSA4212 Assignment 3
Group 39 (Chang An Le Harry Jr, A0201825N)

Introduction

Word embeddings are a powerful technique used in natural language processing to represent words in a continuous vector space. They capture semantic and syntactic relationships between words and can be used as input features for various NLP tasks. In this report, we explore the Word2Vec model, a popular word embedding technique, using the 20 Newsgroups dataset from the scikit-learn package in Python.

Dataset Selection and Preprocessing

The 20 Newsgroups dataset contains a collection of newsgroup posts from 20 different categories. We preprocess the dataset by splitting the documents into sentences and tokenizing the sentences into words. This preprocessed dataset is used as input for training the Word2Vec model.

Word2Vec Model Training

We train a Word2Vec model on the preprocessed sentences from the 20 Newsgroups dataset. The model is trained with a vector size of 100, a window size of 5, and a minimum word count of 5. This trained model generates word embeddings that capture semantic relationships between words. To demonstrate the capabilities of this model, the input word ‘computer’ provides outputs of the words ‘workstation’ (0.77), ‘lab’ (0.76), ‘bulletin’ (0.75), ‘implementing’ (0.73) and packet (0.73), along with their respective similarity scores with the input word.

Model Evaluation

We evaluate the usefulness of the trained Word2Vec model by applying it to two different NLP tasks: text classification and sentiment analysis.

a. Text classification: We create document vectors by averaging the word embeddings for each document in the dataset. We then train a logistic regression classifier on these document vectors and their associated labels (the newsgroups categories). The classification accuracy is reported as a measure of the model's performance. For our model, the classification accuracy was reported as 46.0%.

b. Sentiment analysis: We use a simple rule-based approach to assign sentiment labels (positive or negative) to the documents based on the presence of specific positive and negative words. We create document vectors as in the text classification task and train a logistic regression classifier on these vectors and the assigned sentiment labels. The sentiment classification accuracy is reported as a measure of the model's performance. For our model, the classification accuracy was reported as 86.4%.

Conclusion

This report demonstrates the application of a Word2Vec model trained on the 20 Newsgroups dataset in various NLP tasks. The methods are relatively basic and serve as a starting point for understanding how word embeddings can be utilized in different tasks. For higher performance in these tasks, more advanced techniques and models, such as deep learning-based approaches or pre-trained language models like BERT and GPT, can be explored instead.

DSA4212 Assignment 3
Group 39 (Chang An Le Harry Jr, A0201825N)

References

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In Proceedings of the International Conference on Learning Representations (ICLR) 2013.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532-1543.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 5998-6008.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Agarwal, S. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems 33 (NeurIPS 2020), pp. 1-14.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI Blog. Retrieved from https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf