# E1246 : Natural Language Understanding Assignment-1

**Gururaj K**
gururajk@iisc.ac.in

## 1 Introduction

In this assignment, the goal was to implement and experiment with prediction-based models for learning word embeddings. We use the Reuters corpus for this. NLTK provides an interface to access the dataset. Two Tasks were assigned in this Assignment.

**Task1** Implement a word2vec skipgram model. Use the SimLex-999 word similarity task(Hill et al., 2014) to evaluate models.Experiment with the hyper-parameters of your model such as batch size, number of negative samples, embedding size. Report your observations along with your explanations for the findings.

**Task2** Using the best model learnt in Task 1, verify the claims in the word2vec (Mikolov et al., 2013) paper about capturing relationships between words through the analogical reasoning task. Report your findings quantitatively as well as qualitatively.

## 2 Experimental Setup

Dataset used was the Reuters Corpus provided by NLTK. Pre-processing was done on the data in the beginning such as removing the punctuation, converting to lower-case,removing words which occur less than 10 times in the corpus etc.
Skip-Gram model with Negative Sampling model was implemented. Hyper-parameters such as context size, number of negative samples, embedding size were chosen based on validation performance. 80% -20% train - validation set split was used.
Evaluation of the model was done using Simlex-999 word similarity task. For each word pair in Simlex-999 dataset, the cosine similarity score between the embeddings learnt was calculated. These scores were compared with scores given in Simlex-999 using Spearmann Correlation.
Quantitative analysis was done using the analogical reasoning task introduced by Mikolov. The analogy answer was predicted as the closest word in the learned embedding space.
Qualitative analysis was done by choosing a subset of words from the corpus and visualizing the corresponding embeddings in the TSNE 2d plot.

## 3 Results

### 3.1 Task 1

Below are the results of Spearmann Correlation values for various values of Hyper-parameters

| context-size | Neg Sample | Embedding size | $\rho$ |
|---|---|---|---|
| 5 | 5 | 250 | 0.142 |
| 7 | 7 | 250 | 0.156 |
| 7 | 7 | 300 | 0.193 |
| **7** | **10** | **300** | **0.23** |

Table 1: Spearmann Correlation values ($\rho$)

**Observations :** We see that as we increase the context size, number of negative samples and embedding size, there is an increase in Spearmann Correlation values. This is because with increase in context size upto a limit, syntactic and semantic related words can be captured.
We found that optimal embedding size was 300 and further increase causes degradation in performance. We have highlighted the optimal values. This is because with increase in embedding size after a certain value, the efficient projection in high dimension decreases due to increase in noisy dimensions.

### 3.2 Task 2

**Quantitative Analysis** With the best trained model in Task 1, we performed analogical reasoning task on 4400 analogy pairs.
We got an accuracy of **56%** on this task.

**Qualitative Analysis** Below is the T-SNE 2d plots for visualizing the learned embeddings.

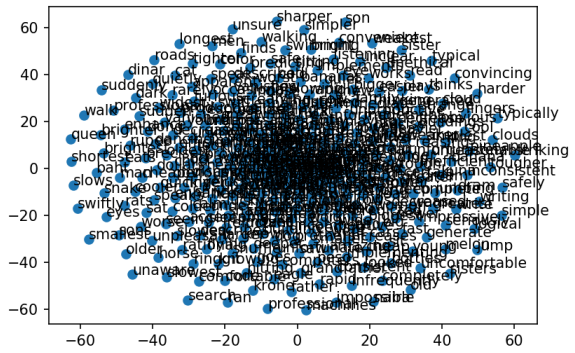


Figure 1: t-sne 2d plot of learned embeddings

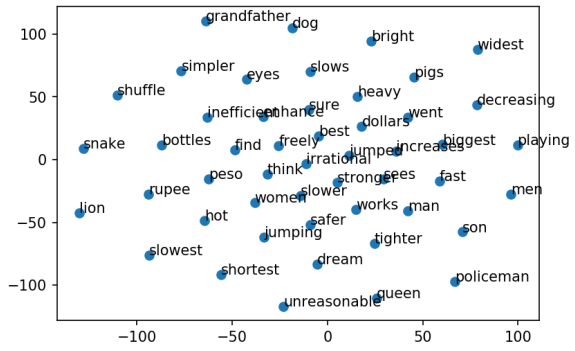


Figure 2: t-sne 2d plot of learned embeddings

**Biases** We observe biases such as ”slower” ”safer” ”hot” around ”women” and biases such as ”works” ”strong” ”fast” around ”men”.

## 4 Github

The code is hosted in the following Github repository. https://github.com/Gururajk/NLU_Assignment-1

## References

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *CoRR*, abs/1408.3456.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.