

# Word2Vec skip-gram with negative sampling

**Prateek Sachan**

M.Tech (CSA)

SR: 15754

prateeksachan@iisc

## Abstract

Word2Vec skip-gram model implementation with negative sampling optimization on nltk's reuters data. Complete implementation is done in numpy from scratch.

## 1 Experimental Setup

Reuters training data contain around 12 lakh tokens but around 3 lakhs are either numbers or symbols like etc. Before preprocessing, the 12 lakh tokens forms the vocabulary of 26432, which includes all the tokens appearing atleast once. This data is also used for experimentation.

After preprocessing the data, around 9 lakh tokens were left which were used for training the word2vec model. These token form a vocabulary of 24935 tokens, whose selection is similar to unprocessed part.

Due to resource limitation major experimentation was done for 100 dimension vectors on both raw data and preprocessed data and each model is executed for 10 epochs.

## 2 Negative Sampling

Negative sampling is done the way exactly defined in the paper using the same probability distribution. For every model, no. of negative samples per context word is fixed to 10.

## 3 Results

W2v model is experimented with different choices of context window. Table 1 shows correlation coefficient( $\rho$ ) on simlex999 dataset on these different contexts windows. Out of 999, 375 pairs were not present on the reuters data, therefore table shows results on remaining pairs. Table also shows the coefficient of glove embedding on same vocabulary. Glove embedding's high score was

Model	Context Window	Rho
Glove-100d	-	0.26
w2v-before	3	0.0952
w2v-before	5	0.0911
w2v-before	10	0.0631
w2v-after	3	0.0870
w2v-after	5	0.0341

Table 1: Correlation Coefficient on 100d word vectors

Section	Total	Not found	top-5
capital-common-countries	506	506	0
capital-world	4524	4524	0
currency	866	866	0
city-in-state	2467	2467	0
family	506	494	0
gram1-adjective-to-adverb	992	572	3
gram2-opposite	812	572	4
gram3-comparative	1332	276	8
gram4-superlative	1122	742	3
gram5-present-participle	1056	550	5
gram6-nationality-adjective	1599	1599	0
gram7-past-tense	1560	568	4
gram8-plural	1332	1122	0
gram9-plural-verbs	870	490	3

Table 2: Analogy task scores

expected because of its large training data and training time.

In Table 1, w2v-before refers to models trained on raw data and w2v-after on preprocessed data.

## 4 Analogy Task

Performance on Analogy Task is lower than expected, this maybe due to the less training data. Table 2 shows results of top 5 nearest neighbours for the vectors from the best model.