COM6513 Lab 7: Continuous Word Representations
Registration Number: 160122440

1. INTRODUCTION

The aim of the lab sessions is to explore continuous word representations and see how they can be used in applications such as sentiment analysis.
The algorithm was implemented with four functions: a function to create the summation of word vectors, a function to create the vector representation of the documents (bag of words), a function for training the perceptron, and finally a function for testing.
Two-thousand (2000) documents were provided, 1000 for positive reviews and another 1000 for negative reviews. The first 800 documents of each were used as the training set (total 1600) and the last 200 were used for testing (total 400). A matrix was created for both the training and testing datasets that included the weights (counts) for each feature. The summation of the word vectors consisted of an array with 300 elements. These elements were added to the matrices of training and testing respectively. An additional vector was added to the last column of the matrices to manually classify documents as positive and negative (positive = +1, negative = -1). These signs were used to compare the predicted classification with the actual classification of the documents, during the training procedure. If prediction was wrong and the original document classification was positive (1), then the positive weights were benefited and the negative weights were penalised. If prediction was wrong and the original document classification was negative (-1), then the negative weights were benefited and the positive weights were penalised.

2. IMPROVEMENTS

Shuffling, multiple passes and averaging improvements were applied to the algorithm, since these improvements greatly improve the accuracy. Shuffling is a generally a very important improvement since the results provided are considered to be more reliable. This is because data partitions normally come from different sources and shuffling them improves the training procedure. Moving on, multiple passes benefits the accuracy because the training procedure is repeated several times, and at each iteration the accuracy is slightly improved. Finally, averaging provides a better overall representation of the weights.

3. EXPLORE REPRESENTATIONS USING GENSIM PACKAGE

The top-5 most similar words/phrases according to the model for 10 words/phrases were first found. These are shown in table 1 below:

Woman:	she	her	Certified_Nurse_Midwife	Ms	silicone_gel_implant
Leo_Messi:	Lionel_Messi	Messi	Diego_Forlan	Xavi	Andreas_Iniesta
France:	soccer	football	Basketball	Villebon_Sur_Yvette	PARIS_AFX_Gaz_de
Football:	Soccer	her	football	Footbal	Basketball
very:	extremely	incredibly	quite	pretty	extraordinarily
Studying:	studying	Examining	Analyzing	studied	Studied
sweet:	sweetness	sweetest	caramelly	syrupy_sweet	yummy
hello:	hi	goodbye	howdy	goodnight	greeting
the:	this	in	that	ofthe	another
tree:	trees	pine_tree	oak_tree	evergreen_tree	fir_tree

Table 1: 5 most similar words/phrases for 10 words.

4. IMPROVE SENTIMENT ANALYSIS BY WORD_VECTORS:

4.1 Continuous Word Representations Improve Performance

Three approaches were tested when constructing the feature representation for each document, which are: using only the normal bag of words (one-hot encodings), replacing the one-hot encodings with the sum of vectors representation, and finally combining both.

(The python file provided (*lab7.py*) can be executed with 3 methods, each representing the approaches described above: *python lab7.py -1* for using the normal bag of words, *python lab7.py -2* for using only the summation of vectors representation, *python lab7.py -3* for using the combination of the previous two.)

- Using the first approach the accuracy provided is 0.83.
- When replacing the one-hot encodings with the sum of vectors representation, the accuracy falls slightly to 0.78.
- Finally, when combining together the two approaches, the accuracy rises to 0.845.

Therefore, it is proved that the continuous word representations improve the performance of the classifier. The improvement is not major, but it is certainly sufficient to be noticed.

4.2 Erros When Using Only Vector Representations

By using only the summation of the word vectors, no significant errors were observed. The only drawback found was that some words were not in the vocabulary of the gensim package. Therefore the vector representations of these words could not be used in the summation procedure.

4.3 Utility of Word Embeddings Affected By Training Data Size

When using only 10% of the training data, the accuracy falls to 0.75 (from 0.845). Therefore, it can be said that the utility of word embeddings is indeed affected by the size of the training data. In order to confirm this, the procedure was repeated with 20% and 30% of the data being used. The accuracies provided were 0.785 and 0.7975 respectively, therefore confirming that utility of word embeddings is affected by the size of the training data.