HACETTEPE UNIVERSITY DEPARTMENT OF COMPUTER ENGINEERING - BBM497

A PREPRINT

HATİCE ACAR

21526599

May 17, 2019

1 Introduction

We have two task in this assignment. We work on work2vec model in task1 and doc2vec in task2.

2 Task 1

We have four word for a one line in the given data and we tried to find last word by using others . To make this firstly I load the Google News vectors data and after read the given data line by line . I find the vectors of first three word . we can call them v1 v2 and v3 respectively . After calculate the result of v2-v1+v3 , I give the result vector as a parameter for calculating cosine similarity with the other all words in the model and I hold the cosine result in a dictionary . Finally I find the max value from dictionary and compare fourth word in the line . Accuracy will increase if they are same. I limited the Google News Vectors model with 5000 word because the code worked very long time and I didn't get a result.So accuracy came 0.2552 .

3 Task 2

In this task we have a csv data that has three column.Second column is include a text about the film and third one include the type of this films. I use these types as tag. I separate first 2000 line as train data and 448 line as test data.I separate also second column buy using nltk.sent-tokenize and nltk.word-tokenize and put them the train and test lists. I get the feature vector from doc2vec model and train it by model-doc2vec.train.I had four vector by using this doc2vec model and vector-for-learning method. These are y-train, x-train, y-test and x-test. Finally I use LogisticRegression classifier and find the accuracy as 0.449888. While using doc2vec model I take vector size as 300 and while training ,epoch as 10. At the begining I take epoch as 50 but code run slow very much also accuracy is smaller than 10 epoch.It is still very slow but faster than other trials with different parameters.