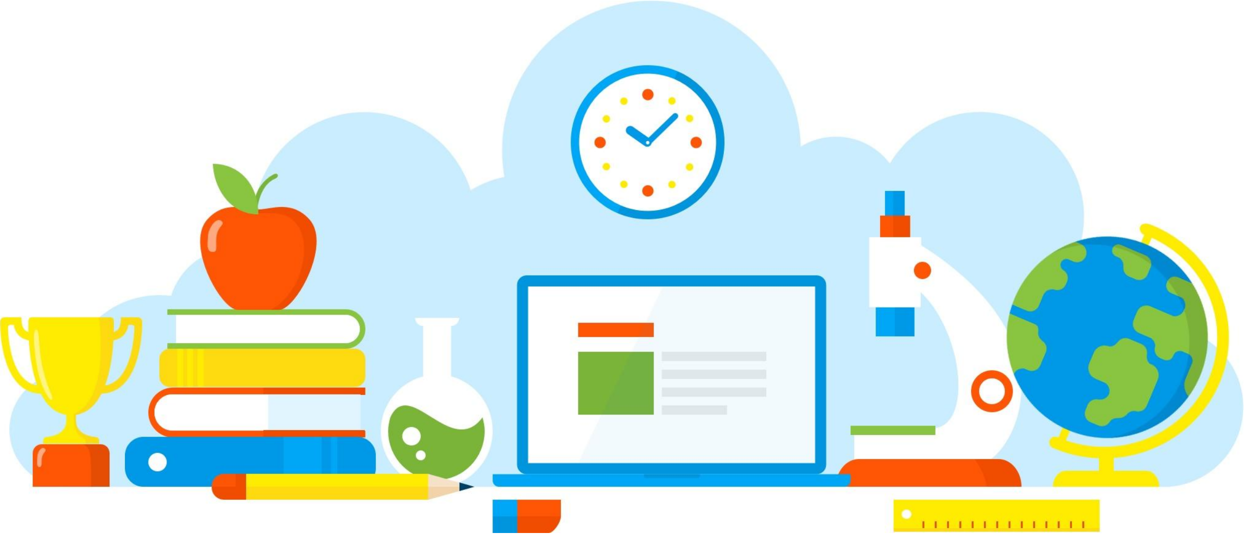


Deep Neural NLP

REPORT

Assignment 1: Word2Vec for Sentences

Nisarg Suthar -202003030
Manjal Shah - 202003037
Meet Kantesaria - 202211018
Jay Joshi - 202211079
Mohit Karelia - 202211080

Data

A part of the peS2o dataset was taken for training the word2vec model for sentences.

Research papers are collected only from the Computer science domain.

A random training data of 3000 documents is used.

The data was preprocessed into a DataFrame and saved as a CSV file.

Preprocessing

The following pre-processing cleanup was done on the input context matrix:

(i) stopwords

(ii) URLs

(iii) bullets

(iv) apostrophe

(v) hyphens

(vi) enumerations (e.g. how this list is being enumerated)

(vii) numerical-to-text conversion

(viii) punctuations.

Matrix Design

The matrix was designed to have the sentences as rows and the words as columns.

A sentence with a particular word would have entry one under the corresponding column, and all other columns would be 0. We kept a window size of two.

Hence, each sentence was represented as a **combination of different wordnesses**.

Training and Labels

Since we are trying to train a Skip-Gram model, we created a training and label vector.

The label vector contained the vectors of context sentences of a particular sentence.

So, there would be a (sentence vector, context vector) pair for every context sentence for training the model.

Embeddings

We decided to have 50-dimensional embeddings to compromise information capturing and training complexity.

Training

Since the large vocabulary size, we restricted our data to only 200 sentences.

The neural network had two layers: one hidden layer and a softmax activation layer.

We used the categorical cross-entropy loss function as the metric.

We trained the data in batches of size eight and kept stopping early to save time.

The first weight matrix had the dimensions of V x 50, while the second weight matrix had the dimensions of 50 x V.

It took 44 epochs to train with the final categorical cross-entropy loss of 0.0240.

Findings

We found our results not to be that accurate. However, there was some promise, as we could find some similarities between the input sentence and the sentences returned by our embedded vectors.

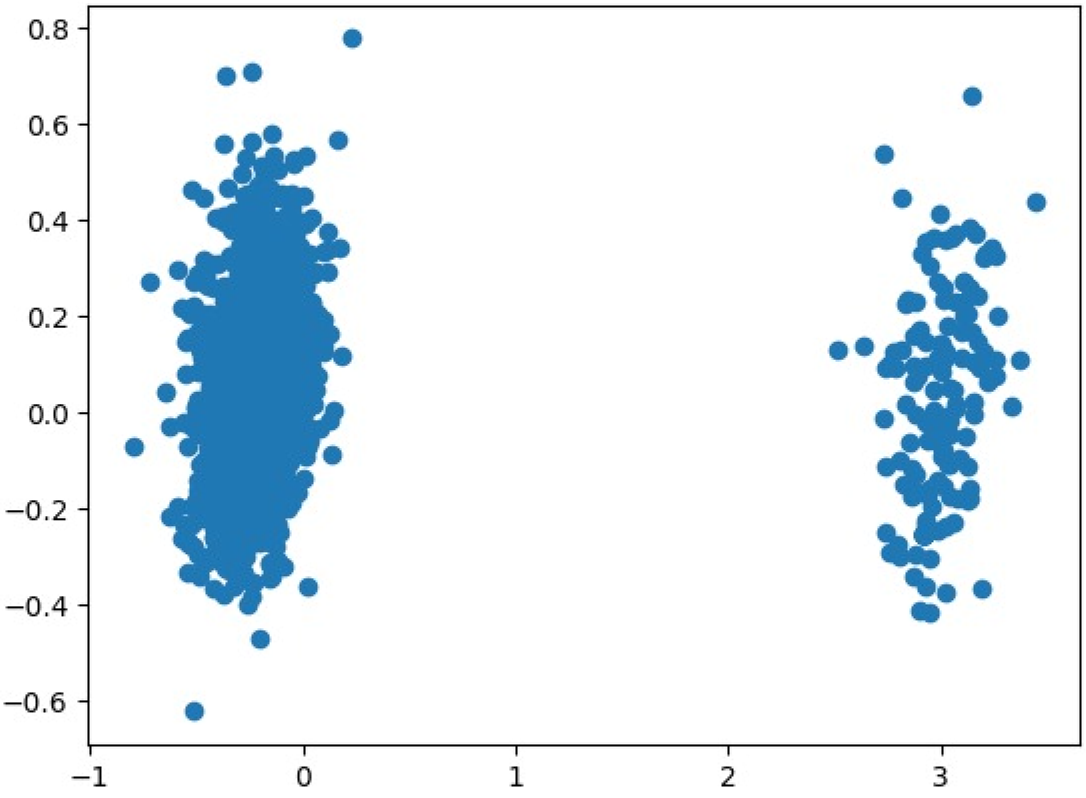
For example:

1. **Input**: in an attempt to induce glucose sensitive dna replication in vitro islets obtained from eighteen and twenty day old fetal pancreata were cultured in the presence of either triiodothyronine or human growth hormone

Output: however both high glucose and high amino acid concentrations increased the islet insulin secretion into the culture medium at all ages studied

Visualisation

We tried to visualise the word embeddings by doing PCA and found the following:



We hypothesise that there are 2 types of research papers in the dataset:

1. Related to biotech like glucose transportation, etc
2. Tech-related like voice and signal processing, etc.