Name- Ali Hassan

Git Hub-https://github.com/Aliahassan/NLP_Assignment3_AliHassan.git

Question 1

Detailed Analysis

Firstly we are reading TSV file then we are extracting question id and corresponding similar question id then we are returning dictionary of similar question and list of all text. This concept is based on natural language processing, which turns words into vectors. Following that, embedding is being created for each sentence. The quick text model is then trained. We are reading a tsv file and posting xml in the main function. We are retrieving the question id from the post xml file using the TSV file we received. Next, for each question ID, we create an embedding and determine how similar the questions are to one another.

Question 2

Detailed Analysis

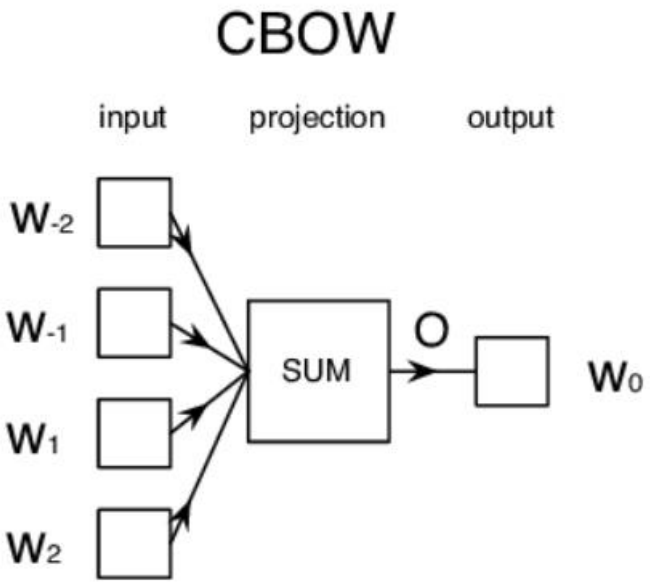
This model is based on feed-forward neural network to check the similarity between two similar questions. The data was obtained from the prior experiment, and a feed-forward neural network was used to calculate the probability and determine if the questions were comparable. After that, we showed the model's accuracy and loss curves. We had accuracy of about 63%, and by adding more epochs, we could raise accuracy to 90%. We used keras to create the neural network and libraries like matplot.lib to plot the graph.

Question 3-What are the differences between Skipgram and Continuous bag of words approaches? Describe the advantages and disadvantages of each.

Continuous bag of words

The CBOW model utilises all of the terms in the target word's neighbourhood to learn to predict it. The target word is predicted using the sum of the context vectors. A predetermined window size surrounding the target word determines which nearby words are taken into account.

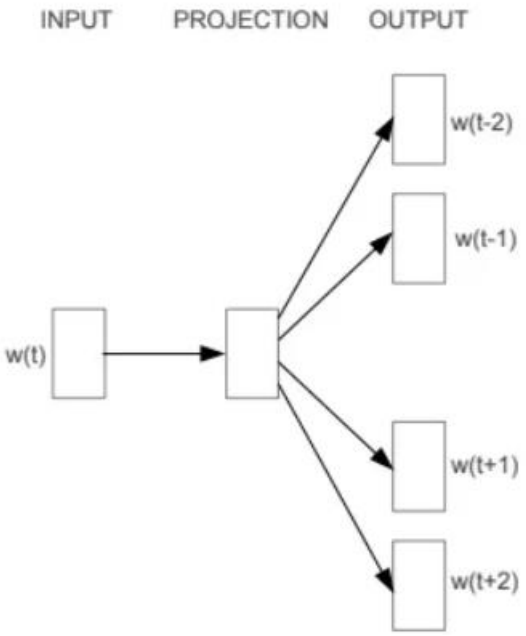
The word in the middle is predicted by the CBOW model by combining the dispersed representations of the context (or surrounding words). While the dispersed representation of the input word is utilised to forecast the context in the Skip-gram model.



Skipgram model

We need unsupervised learning algorithms that can figure out the context of any word on its own because the vocabulary of any language is enormous and cannot be classified by humans. Skip-gram is one of the unsupervised learning techniques used to find the most related words for a given word.

The context word for a particular target word can be predicted using skip-gram. The algorithm is the opposite of CBOW. Here, the goal word is input, and the output are the words in the context. This problem is challenging since more than one context word needs to be predicted.



Skip-gram

Skip-gram: performs well with little training data and accurately depicts even uncommon words or phrases.

CBOW: more quickly trained than skip-gram, somewhat more accurate for frequent words.