NLP ASSIGNMENT REPORT

TASK 1:

Part 1: The preprocessing steps used are:

1.) Convert all text into lower-case
2.) Remove all punctuations.

Here the final output is the term-document matrix which is of size (total no. of documents * vocabulary size)

Part 2: The test.jsonl file contains 500 questions along with 4 options and the correct answer for each question.

In part 3 the queries are created by appending each question with the 4 options one-by-one and finding the predicted options.

Part 3: Here as output we get the total score obtained on running the model for the complete test.jsonl file. Also the accuracy is obtained using the formula specified.

The type of option receiving high similarity score:

The higher the no. of matching words in the (Qi + Oi) and one of the documents out of 1000 documents the higher will be the cosine similarity score. Then higher is the possibility of Oi having the highest similarity score.

The total score obtained is = 121.83333333333329

The accuracy obtained is = 24.366666666666656 %

TASK-2:

Word2vec model: It is a model which represents words as vectors. There are various ways to represent vectors like: one hot encodings , encoding each word with a unique number and word embeddings.

The word2vec model uses a neural network with a input layer, hidden layer and a output layer. Now skip gram model takes as input a word and assigns probabilities to other words for them to be a surrounding word of that word.

Now the doc2vec model is a kind of advanced version of word2vec model. Doc2vec model prepared feature vector for every document. While word2vec is based on the notion that word representation is enough to predict the surrounding words of a document but doc2vec is based on the intuition that the surrounding words of a document also depend upon the kind of document we are working on. Therefore, doc2vec also assigns a tag i.e a kind of hash value to each of the documents.

Preprocessing used:

1. The complete text is converted to lower case
2. All the punctuations are removed from the text

Here during the training of the model, document vectors are created using the genism library. Now, the model is saved for using during the testing phase.

In the testing part , queries are created by making (Qi + Oi) pairs. Now the vector are created using the infer_vector() function.

The maximum similarity value obtained corresponding on checking for all 1000 documents is obtained. The option(s) with the maximum value are obtained.

The final score and accuracy is obtained using the formula specified.

Parameters values used during training:

No. of episodes for which trained = 100

Alpha = 0.025

No decay is done in the alpha value

Vector size = 100

In the doc2vec model a different vector is generated on running every time for the (Qi + Oi) query even on using the same trained model. Therefore, the score and accuracy values also vary in range.

RESULTS:

Accuracy value ranges between 26 % and 28 %

INFERENCES:

To simply explain , the kind of option which gets high similarity score is:

The (Qi + Oi) vector for which the closest document vector is found in the 100 dimensional space as vector size taken by me is 100 gets high similarity score.

This means the document and the (Qi + Oi) talk about very similar contexts i.e have very similar semantics.