Assignment 4 Report:

The data-set we used for this assignment was of assignment 1.

Further, new 15 queries are written for this data set.

Then trained a new WORD2VEC model on this data set. Further, calculated the vector embeddings for these documents and queries by using this trained word2vec model and further used mean-pooling for final representation.

Model Type	Training Data	Purpose	Advantages
Trained Word2Vec Model	Custom dataset (e.g., Assignment 1 data)	Generate domain-specific word embeddings	Tailored to specific vocabulary and context
Pre-trained Word2Vec Model	Large, general corpus (e.g., Google News)	General-purpose word embeddings	Broad vocabulary coverage, saves training time
Pre-trained DPR Model	Large-scale question-answering datasets	Dense passage retrieval for information retrieval	Optimized for query-document relevance scoring

Here is the thing, as mentioned in the table, the trained word2vec model will perform better as compared to the pre-trained one because it has learned the word-embeddings from the data set only which we are working on.

Pre-Trained model is good for having general knowledge but trained has the specific knowledge of the data set.

Considering the DPR model, it is better than both the word2vec models we have worked on till now. It is because of the reason, it is learning the context of the queries and documents by encoding them with the help of BERT encoders.

Why DPR is better than Word2vec Models:

- **Contextual Embeddings**: DPR uses transformers to capture word meaning within context, unlike Word2Vec, which provides static embeddings.
- **Better for Retrieval Tasks**: DPR is specifically designed for retrieval, making it more effective in matching queries to documents based on nuanced meanings.
- **Dual Encoder Architecture**: DPR has separate encoders for queries and documents, allowing for more accurate similarity scoring.
- **Improved Semantic Understanding**: DPR captures complex sentence structures and relationships, leading to more relevant document representations.
- **Higher Retrieval Accuracy**: Due to its deep learning architecture, DPR generally yields higher accuracy for information retrieval tasks than Word2Vec.