NLP Assignment 5 Report(MT19034)

Preprocessing Step

1) Specific symbol like (“.”, “,”, “!”, “/”) is substituted with space.
2) Whole document is converted to lower case.
3) nltk.word_tokenize() is used to generate words.

Assumption

1) Vector_size in doc2vec is taken as 20.
2) All the words come in question-queries are already present in document sentences of 1000. Otherwise it’s vector tag is assumed to be 0.

Refernce link of doc2vec

https://medium.com/@mishra.thedeepak/doc2vec-simple-implementation-example-df2afbbfbad5

Accuracy

1) Cosine Similarity is 0.24166666666666642.
2) Doc2vec is 0.25 (average- “actually it is comming in an range of .24-.26”).

Word1Vec versus Doc2Vec

1) In word2vec we calculate vectors of rare word and find cosine similarity between them. In doc2vec, we figure out rare occuring words as tag of given text and also calculate tag vectors. In simple, we have different documents from different categories and use rare occuring word as tags on those documents.
2) Word2vec calculate similarity between words where doc2vec calculate similarity between sentences or documents.

Return documents

Both model returning tags vectors of queries and documents. In cosine similarity it gives us an vector array of size 5843 and on the other hand doc2vec give us an array of size 20.

Similarity score between query and document

Similarity score is comming in range of (-0.2313127335821037 to 0.4129898126873801).

The type of option receiving high similarity score

Option C is getting highest similarity score where option A is getting lowest similarity score.