Natural Language Processing

Report A5

Pre-Processing

1. Removing Stop Words
2. Using nltk Porter Stemmer
3. Lowercase sentence
4. Removed all punctuation
5. Substituted anything which is not a word with space using regex
6. Removed digits from a sentence

Assumptions

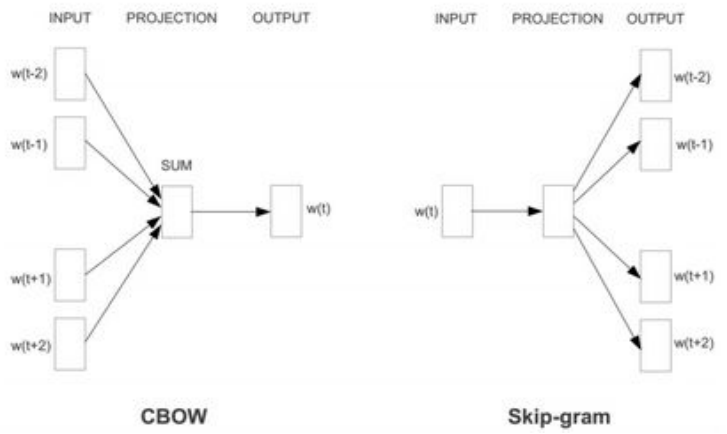
1. To calculate TFID vector of query we have to use IDF value for whole dataset.
2. Hyperparameter tweeks were not essential for this assignment. Therefore i have still tried with different parameters, but didn't mentioned in the report.

Observations

	Cosine Similarity	Doc2Vec
Accuracy	22.56 %	26.00 %

Difference between Word2Vec and Doc2Vec -

Word2Vec representation is created using 2 algorithms (CBOW and Skip-Gram) model. Continuous Bag Of Word creates uses a sliding window approach around current word, to predict using context information. Each word represents as a feature vector. Second algo is very opposite of CBOW, it uses current word to predict context. Its very slow as compared to CBOW but it's more accurate than it.



1. Doc2Vec uses Word2Vec representation and have also added one another vector (Document Tag). This vector is document unique vector. This vector is also used while training for a word W. It holds some numerical representation of the document. This vector acts as a topic of the document. Word vectors represent the concept of a word, the document vector represent the concept of a document. This algo is much faster and consumes less memory.

2. Word2Vec returns similarity score of the sentence with each document. Whereas Doc2Vec returns Document Tag and similarity score of a sentence. Parameter can be changed to return TopN sets of Tag and Similarity score.

3. Similarity Score between query and document -I observed that even the wrong options are being predicted with high similarity score. Either our metric is wrong or we don't have enough data to solve this problem.

4. Inferences type of option- I observed that there are equal no of wrong prediction in almost all the options. Which basically tells us that on increasing documents, much better results can be observed.

Comparison between both of these models can be explained based on less data. There might be some change observable on increasing documents.