

Question1 :

The algorithm used is the skip gram model, applied to a neural network which maps each of the words of the corpus to the Embedding size vector.

Subsampling is done to remove rare words and get only relevant words to train on. Then negative subsampling or hierarchical softmax can be used to speed up conventional softmax, i.e, update the weights for the correct label, but only a small number of incorrect labels. The model trained is then used to plot word embeddings with tSNE. And from the plots it can be seen that with an increase in epochs there is some clustering seen.

Question 2:

Scores: [iterations 3 : alpha 0.75, beta 0.15, top 5 terms, top 10 documents]

Without relevance feedback : 0.493874823468742

With pseudo relevance feedback : 0.5737626732t273

With pseudo relevance feedback and query expansion : 0.5952922953803729

As we can see the MAP value increases in the case of pseudo relevance feedback and further in the case of relevance feedback with query expansion.

Precision is given by : $\frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$

The MAP was supposed to increase in the case of query expansion because, those documents which were more relevant than others were positively added to generate a new query which would now match with a higher similarity value to the relevant documents (the query is made more similar to the relevant and less similar to the non-relevant docs) hence the numerator of the Precision increases.

Further, when we do query expansion, in addition to before we add new terms to the queries which are the most important in the given retrieved relevant documents. Hence similarity with those of relevant documents further increases.