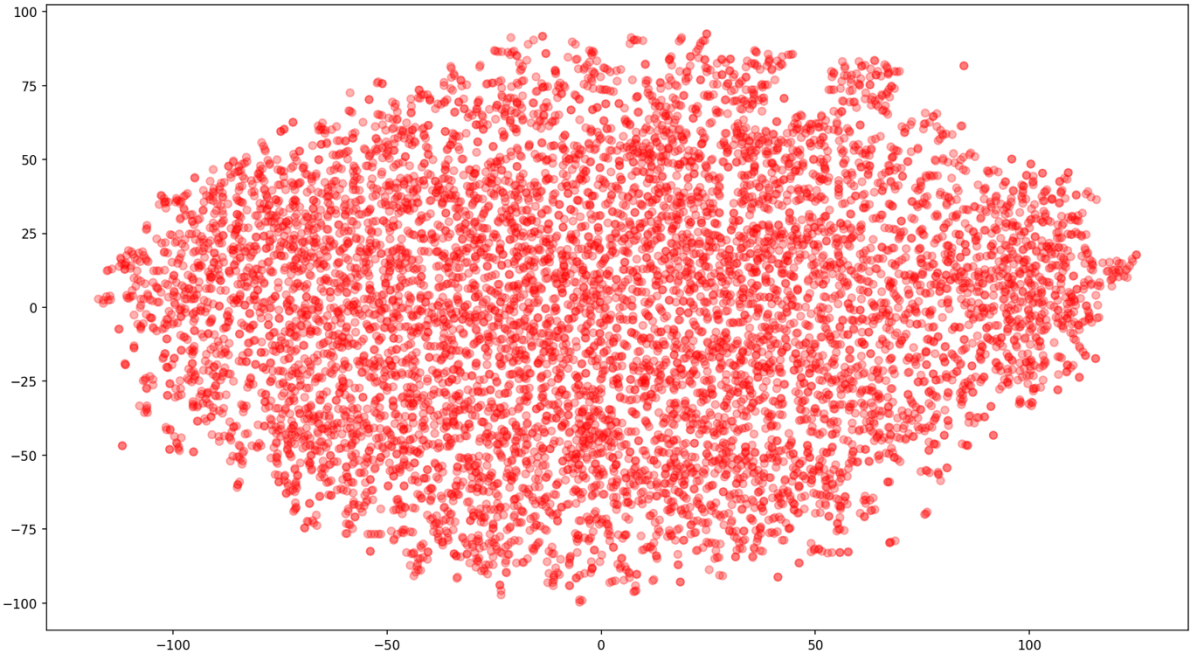
KSHITIJ GULATI
2017162

MCA Assignment 3 – Analysis

Question1

Word2vec has 2 possible types of implementations, namely CBOW and Skip-Gram. Here we implement the Skip-gram version of the algorithm.

1) The corpus is first pre-processed for any special characters and numbers
2) We then divide the words of the corpus into tokens
3) Next, we create training samples for our model. We iterate through the entire corpus with a window in which the central word is called target and the other words are context words. We take such (target, context) pairs for the entire dataset and then encode them using one-hot encoding
4) We then use these training samples to train a neural network which generates a better vectorized representation of the words after each successive epoch. Each word can be represented in terms of a vector of *n* dimensions where n is the number of neurons in the hidden layer of the neural network
5) We then use TSNE to plot the representations of all words in the corpus in 2D
6) We train this model for 25 epochs and the plot after the 25th epoch is as follows:



After training our model sufficiently, we begin to notice the formation of small clusters as we can see in the above diagram with the large dense red spots which shows that similar words have similar representations. The png images for the 25 epochs have been attached in the zip file.

KSHITIJ GULATI
2017162

Question 2

MAP (Baseline Retrieval) = 0.49
MAP (with relevance feedback) = 0.58;
iter =3, alpha=0.8 ,beta=0.2
MAP (relevance feedback with query expansion) = 0.59;
iter =3, alpha=0.8 ,beta=0.2, r= top 5 terms

It is evident that the MAP value increases in the case of retrieval with relevance feedback and further in the case of retrieval with relevance feedback with query expansion.

MAP increases when we use query expansion with relevance feedback because the documents which are more relevant to a particular query in comparison with others are added to the query after each successive iteration making the query more similar to the highly relevant documents and dissimilar to the non-relevant documents. This way the query is enhanced after each successive iteration which allows it to retrieve more similar documents with greater ease.
In case of query expansion, we enhance these queries by adding the top 5 or n most relevant features to the query for each document. Hence, retrieving similar documents becomes far easier.