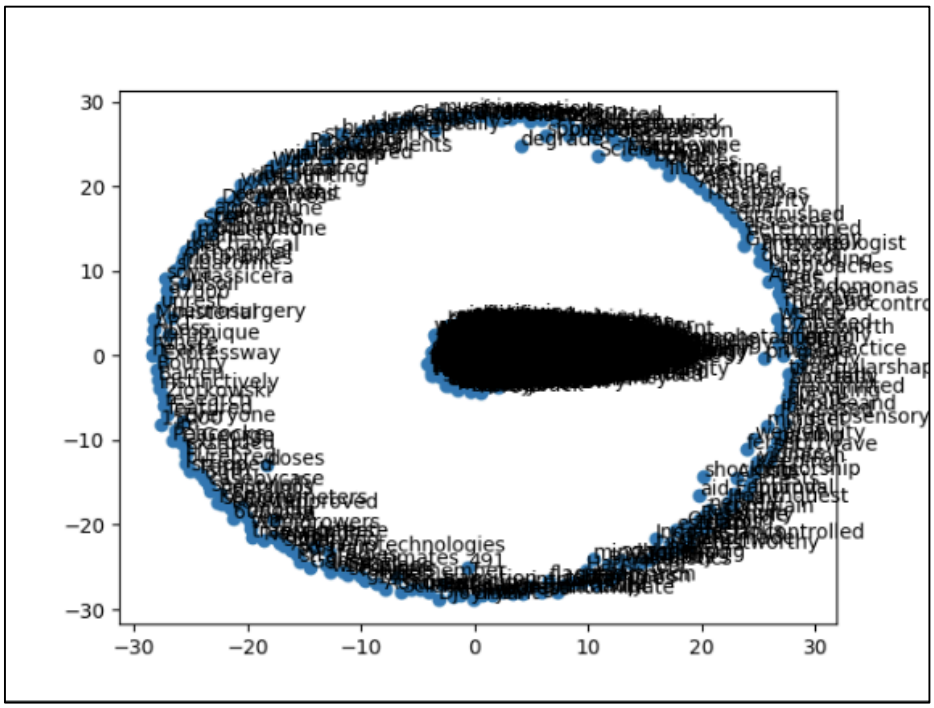
MCA ASSIGNMENT – 3

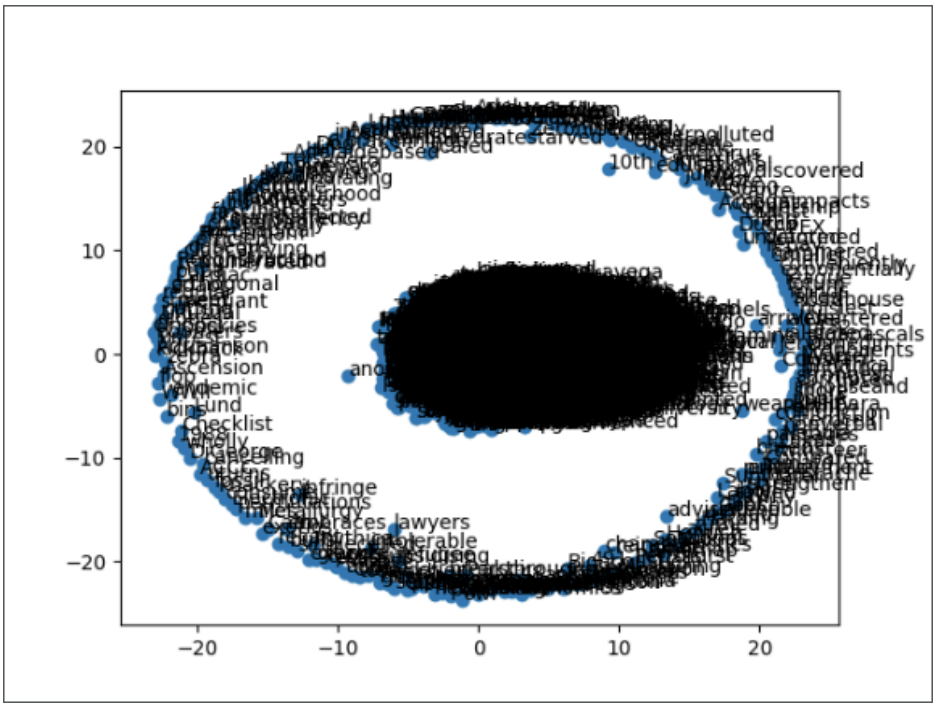
Q1.

Algorithm – We use the Skip Gram model with Negative Sampling to train the word vectors. Skip Gram model based on the input word, predicts the context of the word. To achieve this, we create pairs of words. Give those pairs an output of 1 that are in the context of each other else 0. A single Embedding layer is used to convert word to embedding vectors. The dimension of the embedding vector is 100. Both the vectors are multiplied and MSE Loss is used.

The Loss decrease from 9.0 in the first epoch to around 0.5 till the fifth epoch. From the TSNE graphs, it is clearly visible that the model is able to distinguish words into two sets. This is because the abc corpus contains words that are taken from two documents. Our Model after the first epoch itself is able to distinguish words from the two documents



After 1 epoch



After 4 epoch

Basically the Model is trying to separate words that come from two different documents

Q2

Baseline Retrieval
MAP: 0.5210112268116108

Retrieval with Relevance Feedback
MAP: 0.632

Retrieval with Relevance Feedback and query expansion
MAP: 0.634

Definitely there is an increase on using relevance feedback due to the fact the queries are further improved using the document representations. The queries are able to capture more information than before. In case of vector expansion, I have used a thesaurus that stores similarity score between all the word pairs. When the query is given as an input, the thesaurus is used to get 5 similar words for each word in the query. Thus this improves the searching power of the query.