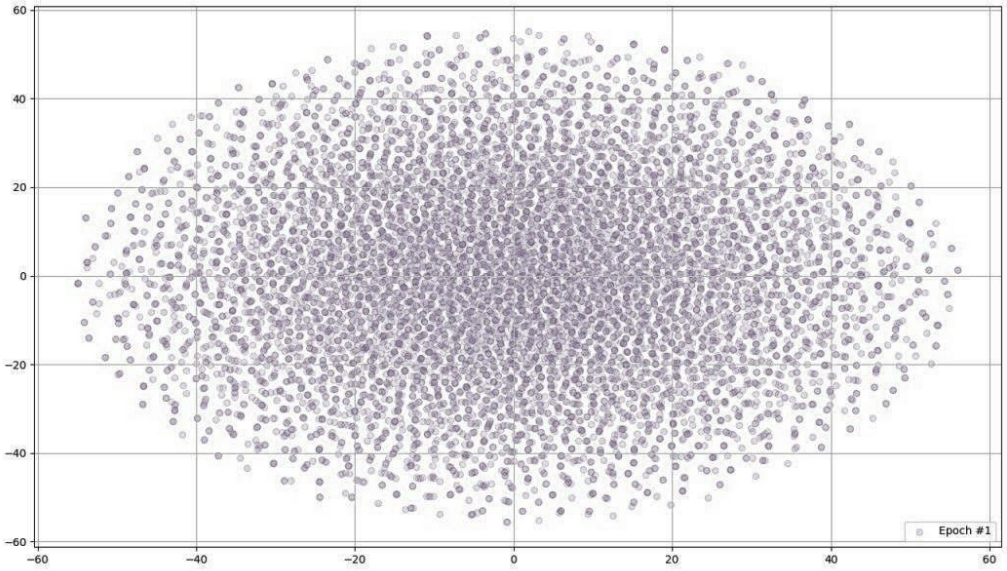
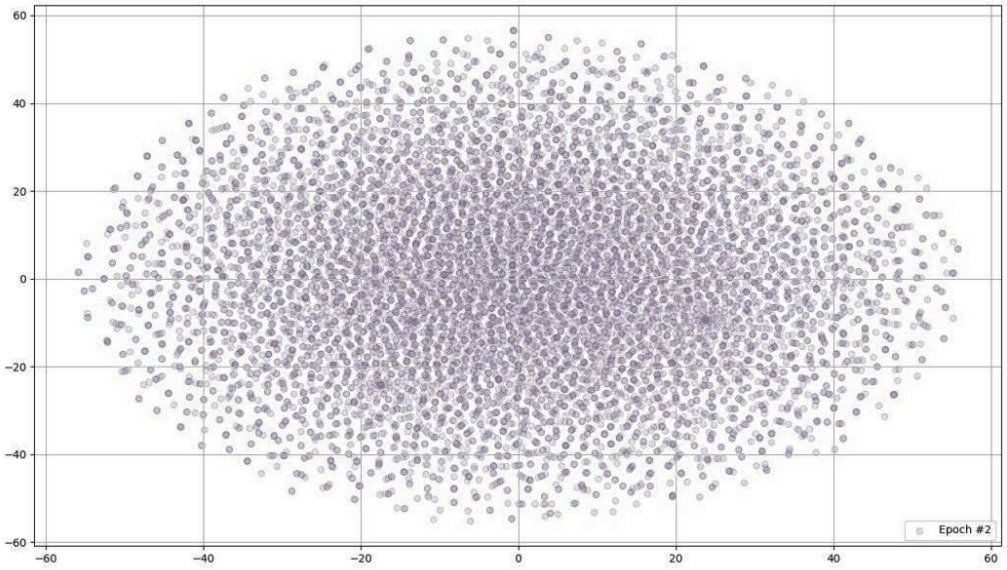
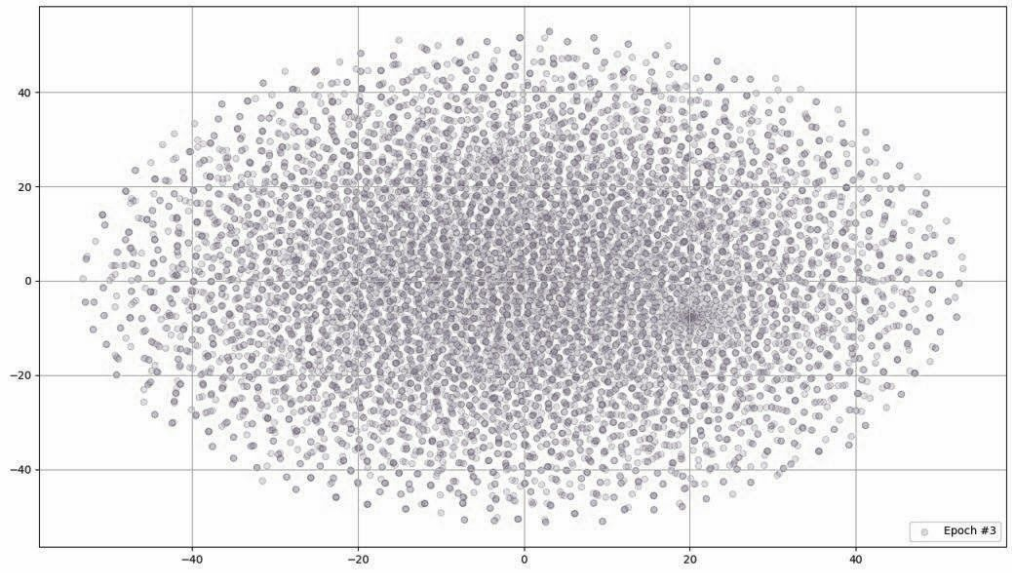
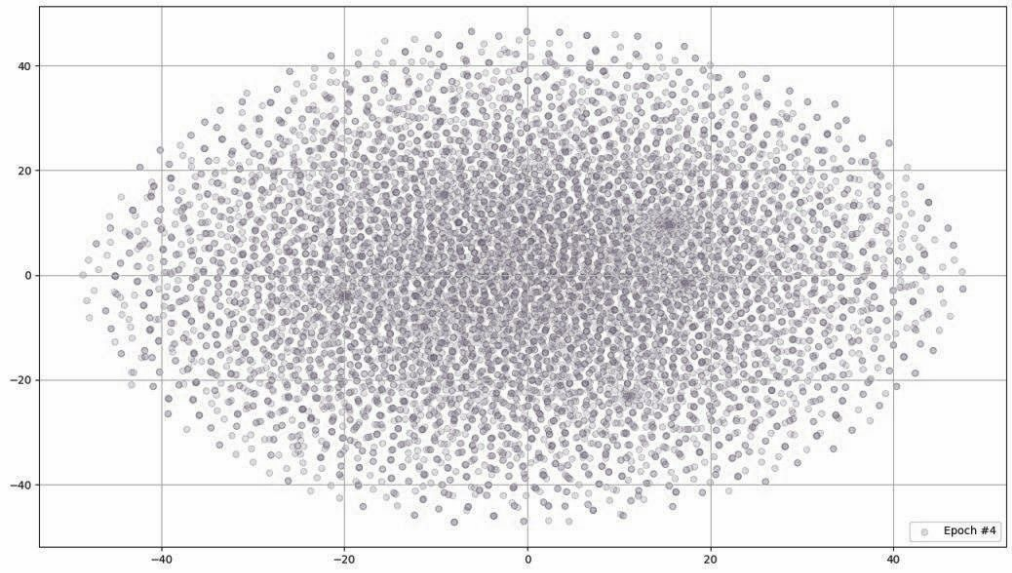
Homework-3

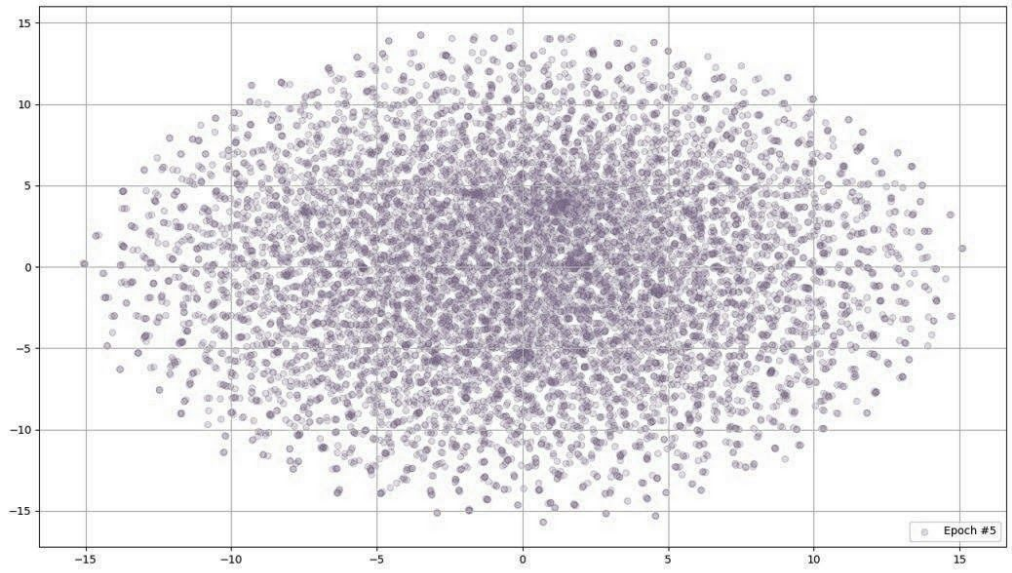
Following are the visualizations for each epoch

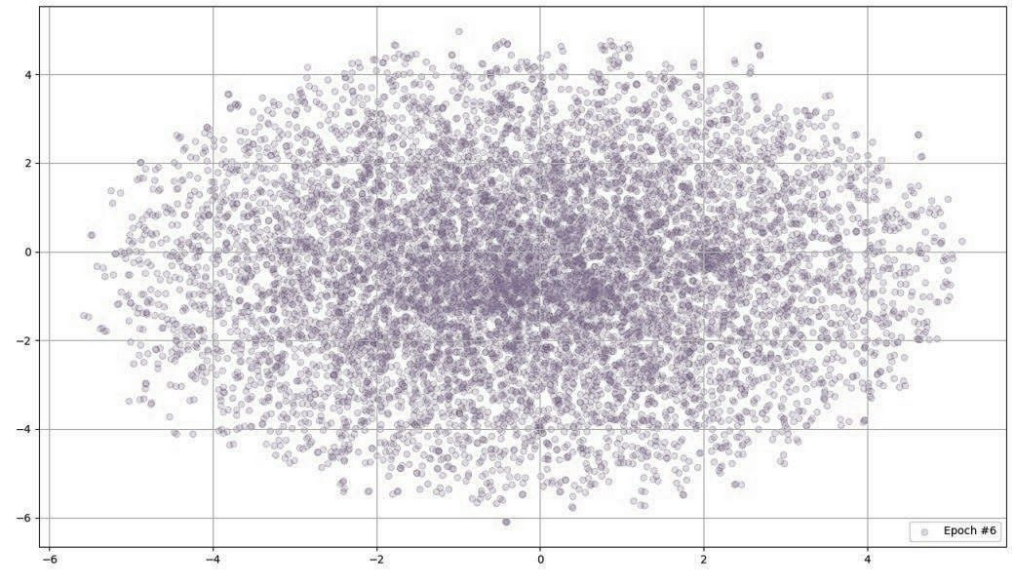


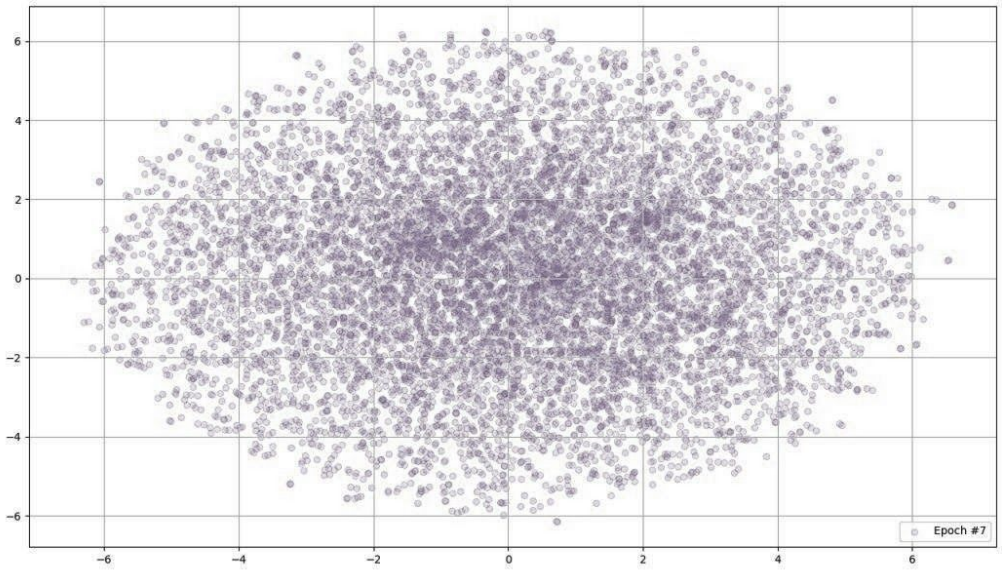



40
20
0
−20
−40
−40
−20
0
20
40
Epoch #3


40
20
0
−20
−40
−40
−20
0
20
40
Epoch #4


15
10
5
0
−5
−10
−15
−15
−10
−5
0
5
10
15
Epoch #5


4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
Epoch #6



Question 1:

1. 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.
    a. It takes a corpus and divides it into tokens.
    b. A probabilistic subsampling is done to remove rare words.
    c. A window is run across the corpus to generate positive and negative samples for training the network.
    d. Negative sampling is used to update the weights of only a few negative samples and one positive sample in one iteration.
    e. After training positive samples have more similarity than others.
2. After each epoch as the network gets better by training, the initial random spread tSNE plot starts to form small clusters, which indicates the fact that the similar words would have values near each other.

Question 2:

Scores: [iterations 3 : alpha 0.85, beta 0.15, top 5 terms, top 10 documents]

1. Without relevance feedback : 0.49
2. With pseudo relevance feedback : 0.57
3. With pseudo relevance feedback and query expansion : 0.59

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 : |{relevant documents} ∩ {retrieved documents}| / |{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.