Name: Anubhav Jaiswal
Roll Number: 2016014

Q1.

I have used SkipGram as the model for word2vec.
The dataset is formed with pairs, one being the center and other being the context. The context is the word that is inside the window of the center word. I have used 4 as the window size, and words that have 10 or higher frequency in the abc corpus.
The model encodes the original embedding vector, of the center word, into a 100-dimensional vector and decodes it back. The output of this model is used to obtain the loss against the context word.
The embedding generated by the model is the vector of that particular word.

The words that are similar to each other are clubbed together. However, words like chimpanzee and sperm are separated from each other.

Changes during training:

1. Per Pair loss reduced to about - 0.0085
2. Some words were chosen to be displayed on the graph.
3. The similar words in the list, moved relatively closer to each other.

Reference:

1. TSNE plot - https://medium.com/@aneesha/using-tsne-to-plot-a-subset-of-similar-words-from-word2vec-bb8eeaea6229

Q2.

I implemented Rocchio's feedback method. Alpha = 0.8, beta = 0.2

The cosine_similarity method gave a score of 0.52101
Relevance_feedback method gave a score of 0.63478
Relvance_feedback_exp method gave a score of 0.63743

However after 3 epochs:
Relvance_feedback_exp method gave a score of 0.62084

For query expansion, I used the standard auto thesaurus methods. It constructs a matrix of n x n, (n is the number of words). This matrix is $C = A \cdot A^T$, where A is a matrix of count of words in each document.

The query expansion method didn't see much of an improvement over the normal relevance feedback method.