MCA-HW3 Report

Aarushii Agrawal

2017324

Problem 1:

The algorithm:

The sentences are passed as arrays of words into the model and a **training\_data**  is generated. This is basically an array of word vector along with its context vectors, vectors are one hot encoded words. Context vectors are vectors of words around it depending on the window size.

For training the model, a neural network of 3 layers (1 input, 1 hidden and 1 output) is used. Softmax function is used. For backpropagation, stochastic gradient descent is used.

Learning rate of 0.01 or 0.05 was used. Window size of 2 (2 words before and 2 words after) was used.

During visualization upon every 100 epochs, words similar to each other started coming closer. Same words overlapped (since they had the same vectors). Words used in similar contexts like - **“plenty”, “limited”, “few”** or **“producer”, ”manager”, “director”** came close to each other. Words like “company” and “community” came together.

Words like **“accident” and “unfortunate”** were very similar. Words most similar to “sports” were “sector”, “market” and “apparel. “Friday” matched with “Satuday” and “event”.

Words used in similar context (often together or interchangeably) showed up as similar words.

Problem 2:

Retrieval with Relevance Feedback

MAP: 0.7559349568268208

Retrieval with Relevance Feedback and query expansion

MAP: 0.7504322038185971

Performance with both kinds of relevance feedback improved significantly. Updating all the terms and updating only top 10 terms in a query showed negligible difference. This was expected since when we update top 10 terms, we are updating the most important terms. The other terms are of very less importance and have scores close to 0. Therefore, majority difference is made by the top few terms.