Sentiment Analysis of Movie Reviews

By Subham Anand – z5151878

and Vivek Arunagiri – z5141013

Pre-processing

Following are the steps involved in pre-processing of data:

- The input string is converted to lowercase.
- Pre-defined stop words and few extra words added by us to the list of stop words are removed.
- Words containing any special characters other than ‘ and “ are removed.
- Punctuations and words length less than three are removed because they are mostly meaningless and act as noise in our data.
- A list of tokens(words) is returned from this method.

Design Decisions

Following are the steps involved in our define graph to model our network:

- The average review length is 125, so the MAX_WORDS_IN_REVIEW is set to 200 to include most of the reviews.
- The BATCH_SIZE and EMBEDDING_SIZE are both set to 50, the learning rate is set to 0.0001 and drop out probability is 0.6.
- Two placeholder is defined for the input data and labels with size [Batch size, max words in review, embedding size] and [batch size. 2] respectively.
- The model consists of a two-layer LSTM, with each layer having a size of 50 cells.
- Then the outputs are passed to a fully connected layer with a soft max activation function.
- Loss is calculated by using soft max cross entropy with logits version 2.
- We get the optimizer by Adam Optimizer and passing the learning rate to it.
- And at last accuracy is calculated using reduce mean.