Deep Learning for NLP
Assignment 1

Preprocessing:

After reading the training data in pandas dataframe, I have:

1. Removed everything except alphabet.
2. Converted everything to lowercase.
3. Removed stop words which are from nltk.

Initially I did stemming and lemmatization, but it did not give any improvement in performance and just increased time, so I finally did not use them.

In IDMB dataset,
Training set size: 25,000x2
Test set size: 12,500

In SNLI dataset,
Training set size: 550152x3
Development set size: 550152x3
Test set size: 19824x2

In AG's News topic classification dataset,
Training set size:120000x3
Test set size: 120000x2

After cleaning data, I have used TF-IDF for converting into vector. Earlier I have used simple bag of words, it gave low accuracy so, I have used TF-IDF in all questions.
In 2nd and 3rd question, the first two columns, I have converted into vectors separately. Then I concatenated both into one column.

Training:

As a classifier I have used naive bayes, which took hardly 30 seconds to train, in all cases. Only two columns are there, on which I have used sklearn’s multinomial naive bayes.

Results:

I have used 10-fold cross validation. And my accuracies are as follows:
1.80.139%
2.58.93.%
3.89.69%