CSCI-B555: Machine Learning - Programming Project 4

To run the program, use python3 pp4.py

No command line arguments are necessary.

Ensure that the dataset files reside in the pp4data folder, are in the format as given in the assignment and the index file has .csv extension.

The code in pp4.py has implementation of both Collapsed Gibbs sampler for LDA and the logistic regression implementation on LDA and Bag of words model.

The run time for this program is comparatively more and takes approximately 20 mins on the entire dataset.

The submission consists of:

1) pp4.py
2) topicwords.csv : CSV files of 5 most frequent words for the topics
3) LDAvsBoW.png : Graph plot
4) Report results in PDF