CS671A: Introduction to NLP
Assignment #2: Document representation and classification.

Due on: 20-3-2018, 23.59 12-3-2018
MM: 500

1. In this assignment you have to use different document representations and classification algorithms to study how they perform in a sentiment classification task.

   The data set is the Stanford large movie review data set. It has binary sentiment labels. The data set is available at: https://ai.stanford.edu/ amaas/data/sentiment/. The data set is also cached on the course ftp site.

   You have to explore the following representations for documents:

   a) Binary bag of words.

   b) Normalized Term frequency (tf) representation.

   c) Tfidf representation.

   d) Average of the Word2vec word vectors in the document with and without tfidf weights for each word vector while averaging.

   e) Repeat the above with GLoVE vector representations for words.

   f) Averaged sentence vectors for sentences in the document.

   g) Paragraph vector - treat the whole document as a single paragraph.

   For classification algorithms try the following:

   - Naive Bayes.
   - Logistic regression.
   - Support vector machine (SVM).
   - A feed forward neural network.
   - A recurrent neural network (use an LSTM or GRU).

   Since the product of the number of representation and classification algorithms is quite large you don't have to try all combinations. Try as many combinations as you can (but at least 20) and try those that you think will allow you to compare the suitability of different representations for the sentiment identification task. [500]