Machine learning for computational linguistics: Assignment 3

Deadline: August 1, 2016 (**changed!**)

Modified (July 1, 2016): smaller data set.

In this set of exercises you will work on sentiment analysis on a real-world data.

Please submit both your source code and **a single PDF file** containing the answers of the questions below via email. You can submit your homework either as email attachments, or (preferably) send the address of a git repository (GitHub, BitBucket, ...) containing both the source code and the answers.

Data

We will use a reference data set for sentiment analysis research, from http://www.cs.cornell.edu/people/pabo/movie-review-data/ by Pang et al. (2002). The data contains short documents (movie reviews). Each document is classified as positive or negative. Please use version 2.0 of the data.

Exercises

For all of the exercises below, use 10-fold cross validation for evaluation. You do not need to train the word vectors yourself.

1. Train an evaluate a logistic regression classifier, based on both word unigrams, and word bigrams. (You can limit the features to the words that appear at least N (e.g., 5) times in the corpus.)

   Report average accuracy of your models.

2. Repeat the first exercise, but this time use word vectors, with a multi-layer perceptron. You can represent each document as the sum (or average) of all the word vectors in the document (for bigrams you can concatenate the vectors).

   Report average accuracy.

3. Design and train a convolutional neural network (CNN) for the same task. You are free to choose the architecture, but make sure your convolutions cover at least bigrams, and the model learns multiple convolutions (features maps).

   Briefly explain your network architecture, and report the accuracy of the model.

4. Briefly (not more than half a page) discuss the results you have obtained. Include comparison of each model for their accuracy as well as computational complexity.

Notes

- Note that the model described in exercise 2 is a very poor model. It throws away a lot of information that is available to the model in the first exercise. You are encouraged to think about an alternative fully-connected multi-layer network that would be able to compete with the simpler model in the first exercise.
- The first version of this homework pointed to the data set by Andrew et al. (2011) http://ai.stanford.edu/~amaas/data/sentiment/. The current version requires you to use a smaller data set. Results with both data sets will be accepted.

References

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). *Learning Word Vectors for Sentiment Analysis.* The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, *Thumbs up? Sentiment Classification using Machine Learning Techniques*, Proceedings of EMNLP 2002.