Problem Set 09
Data Mining and Machine Learning – Spring 2016
Due date: 2016-05-02 13:00 (monday)

All assignments must be uploaded to the assignments tab in ClassesV2 (notice that this is **not** the dropbox) by the date and time specified. Make sure that you follow the instructions exactly as described. You may discuss problem sets with others, but must write up your own solutions. This means that you should have no need to look at other's final written solutions.

You need to turn in all of your solutions as a zip compressed file, named `netid_pset09.zip`, with your actual netid filled in in all lower case letters. This archive should contain just the one file:

- `pset09.pdf`

The pdf file should contain your written response to the task outlined below.

Instructions

This final assignment is more open-ended than the previous 8. It is based on the data discussed in this paper:

> Blitzer, John, Mark Dredze, and Fernando Pereira. "Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification." In ACL, vol. 7, pp. 440-447. 2007.

You can download a slightly cleaned up version of dataset here:

> `http://www.stat.yale.edu/~tba3/class_data/mdsd.zip`

This directory consists of a two directories indicating whether a review is a positive or negative one, and within those directories, one file per review. This is very similar to the IMDB data we processed in the python notebook for Lecture 21. Here, however, I have not split the data into a training and testing sets; you should do this yourself in some reasonable way. There are 20k positive and 20k negative reviews.

Your task is to construct and evaluate a predictive model that predicts whether a review is negative or positive. You may use any combination of neural networks, support vector machines, generalized linear models, or any other techniques we have studied this semester. The only stipulations are that you must:

1. use some sort of word embedding; this may be either learned on the data or you can use a pre-trained GloVe or word2vec model

2. perform and display some experiments as to how you settled on your final model; for example, how did you choose the number of words in the embedding or the depth of a neural network model

You may also, but do not need to, use external tools such as `spacy` to do pre- and post-processing on the text data. **Note however that you will not be primarily judged on your overall classification rate (this is not a competition) but rather on your overall approach and how well it is described.**

For the assignment you should write up a description of your approach and final model *in prose*, clearly explaining the models you choose and how you addressed selecting and tuning parameters. You should also present results showing how well your model performs. This should include classification rate, but also other metrics, which may include things such as representative mistakes, a list of selected tokens if you use penalized estimation, and/or a confusion matrix. Make sure to properly cite any algorithms techniques you use. This write-up should not exceed three pages, inclusive of any references, tables, and plots. You will held accountable for writing clearly and accurately about your approach to the problem.

Please do not flood piazza or my e-mail looking for a magical checklist of what you need to do to get an A on this assignment! Instead, play around with the data, try out some new techniques, and fun with it. If the final product is reasonably presented and thought out, good grades will follow.