CS 6320
Homework 3
Professor Moldovan
Due March 9, 2015 before class

Homework Prerequisites:

Install Python 2.7: https://www.python.org/download/releases/2.7/

Here is a good starting point for learning python: https://developers.google.com/edu/python/

If you prefer IDE, this is a good one: https://www.jetbrains.com/pycharm/download/

Install NLTK: http://www.nltk.org/

NLTK has:

- tokenization,
- POS tagging,
- corpora, including a data set in sentiment analysis by Bo Pang and Lillian Lee. http://www.cs.cornell.edu/People/pabo/movie-review-data/ It is included in the NLTK python package in nltk.corpus.movie_reviews.
- dictionaries, from stopwords to WordNet.
- stemmer (WordNet also has lemmatization)
- most used machine learning algorithms

Assignment:

1. Develop sentiment analysis based on maximum entropy classifier on movie reviews data
2. Identify discriminating features, experiment with lemmatization/no lemmatization (WordNet), filtering out punctuation, filtering out stopwords;
3. Experiment with unbalanced collection: change proportions of positive/negative examples in training data;
4. Submit: (1) source code for your experiments, (2) report, containing description for experiment setup (i.e. preprocessing/filtering, training data size, proportion), experiment results, and conclusion.