Sentiment Classification Assignment
r koncel-kedziorski
kedzior@uw.edu

For this assignment, you are asked to create a Logistic Regression classifier which can analyze the language of a movie review and determine whether the author had a positive or negative opinion of the film.
The dataset we will use is the Cornell Polarity Dataset v2.0. Please download this from the LMS website, as it has been split into training and testing data.
Using the included skeleton code, create classifiers trained on the positive reviews in the ./pos directory, and the negative reviews in the ./neg directory.
You will need to create a vector representation of each review like we discussed in class.

Please create classifiers which use the 100, 1000, and 10,000 most frequent words (normalized for case, minus stopwords) as features.
Try using the one-hot representation, as well as the occurrence counts as feature values.

What are the differences between these classifiers in terms of performance and training time?

You may want to reference the documentation for the Logistic Regression class from scikit-learn (http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html).
Feel free to email me with any questions you have.