Homework 10: Due Mon 02-08-2021 Doris Chen
(20 pts)

Problem (Movie Review Sentiment Analysis): Negative movie reviews are contained in the file `rt-polarity.neg` and positive reviews are contained in `rt-polarity.pos`. Train a Naive Bayes classifier to predict if a movie review is negative or positive.

- The Jupyter notebook `rt-polarity-preprocessing.ipynb` completes preprocessing of the data for you.
- Use tf-idf vectorization to extract features (attributes) from the movie reviews.
- Complete an explicit grid search of the range of n-grams to use in tf-idf vectorization by filling out the table below.
- Use grid search cross-validation to optimize the pseudo-count hyper-parameter (alpha) of the Naive Bayes classifier and compute the validation error.
- Use the optimal value of alpha and all of the data to compute the training error.
- Note: The number of attributes equals the size of the vocabulary.

Naive Bayes Classifier Using TF-IDF Vectorization:

n-grams	# attributes	alpha	validation error	training error	baseline error
(1,1)	18067	1.78	0.235	0.09	0.5
(1,2)	101368	0.91	0.234	0.007	0.5
(1,3)	183646	0.69	0.233	0.002	0.5