Homework 5 – Sentiment Bakeoff!

A bakeoff is typically a supervised learning challenge in computer science.

In this bakeoff, the goal is to achieve the highest average accuracy in predicting labels for the NLTK movie_reviews corpus (see sentiment.py).

You are welcome to use the Python NLTK package (the starting script is provided in sentiment.py). You can also opt to write the movie_reviews corpus to a csv file and use another supervised learning package (such as RTextTools).

The Pang and Lee article demonstrates that different feature representations and algorithms produce different results. Your goal is to achieve the highest result relative to your peers. You can use any of the tools we've covered.

The homework that you submit should include:

1. Your annotated script (explaining what you are doing).
2. Average performance of your final algorithm across 4 different 75%/25% train test splits (please also report the results for each split)

Let me know if something is not clear!

Top 2 receive a box of fudge from [Bon Bon](Bon Bon)