QSS Data Science Challenge at HackDartmouth IV

We would like to challenge students with the Yelp Dataset. We've randomly split the Yelp dataset into a development and evaluation set. The development set has records for 749,575 Yelp reviews. The evaluation set has records for 250,425 Yelp reviews. Both files are identically formatted JSON files (with the stars variable removed from the evaluation set). The data files are approximately 1GB uncompressed.

THE CHALLENGE: Predict review star ratings from other available data (especially review text). Students might attempt to use machine learning, regression or clustering techniques to make a prediction.

Files:

https://www.dropbox.com/s/u54a8ahlswncwvm/QSSHackathonData.zip?dl=1

- yelp.json: development set
- yeldHeld.json: evaluation set
- starter.py: starting point for the challenge (loads the data and imports some libraries that might be useful)
- stars.csv: submission template

Helpful starting points:

- Yelp provides many useful examples (including conversion from JSON to CSV) on their github dataset project: https://github.com/Yelp/dataset-examples
- Yelp offers a schema for the dataset (reviews.json for this task): https://www.yelp.com/dataset/documentation/json

Evaluation:

- The winner will be determined by the largest *proportion* of correct star ratings for the evaluation dataset.
- Students must submit a CSV file with their predicted ratings for each record in the evaluation dataset (example attached).
- Although we will assess a winner by computing accuracy, we encourage students to consider precision and recall in developing their approach (overfitting a model will reduce accuracy in the evaluation set). We also encourage students to use cross-validation in the development of their model.