COM 307 Machine Learning/Data Mining FINAL PROJECT (due by April 28 before class start)

In this project you will be using real-life data analyzed using machine learning and/or data mining methodologies. You will select the data and the algorithm that you wish to use to analyze it, implement the algorithm, and present the results to the class in approximately 10-minute presentations in the class time leading up to finals week. You will be graded based on your (on time) attendance during the final presentations (5% of your final grade), your code/methodology (18% of your final grade), and your presentation itself (7% of your final grade). For full credit on your code you will need the following:

1. A properly functioning implementation of a machine learning algorithm (5%).

2. Well-documented and readable code (1.5%).

3. Validation. This can be built into your primary code or be a separate piece of software. Either way, 1 & 2 should be true for this code as well (3%).

4. A summary of both your strategy/the data you used and what you found. These can be 2 separate files/essays/write-ups or combined into one, but make sure that you include both (1.5%)!

5. A minimum of two "extras" as outlined below (7%).

For your presentation, you will need to:

1. Explain your data source and why it was selected (any importance to science or society, "just interesting," "to be funny," or what have you)

2. Explain the methodology that you used to analyze it. What machine learning algorithm(s), what feature(s) you used, did you try multiple things and how did each perform, etc?

3. Explain any "extras" that made the analysis complicated.

Extras:

1. Use a machine learning algorithm that wasn't used in projects 1, 2, or 3
2. Use data that is hard to access, parse, etc.
3. Use different machine learning algorithms and compare/contrast the results
4. Extensively test more than just standard parameters. Add or remove features, change underlying distributions, etc. and compare results
5. Any other ideas for "complications?" Ask and I'll see if they're sufficiently complicated to count!

[Submit]

Submit the following files (zipped or not):

1) Your code (in whatever language you used)

2) Your training data & where you got it from

3) Results. Both validation results and predictions on a test set

4) The write-ups as described above

5) Optional: Your presentation (PowerPoint or PDF, most likely, but anything that I can open is fine)