MAT 345 - PROJECT #4
due Wednesday, November 28, 2017 at 10:00PM.

OBJECTIVE: In this project, you will perform k-means clustering. Implement the ID3 algorithm to create a decision tree for extra credit.

GRADING: The assignment is worth 5% of your course grade. The extra credit is worth 3% of the course grade.

INSTRUCTIONS: Students will work individually on this project, but they may ask questions and clarification from classmates and the instructor. Students must submit their projects on Moodle.

SUBMIT THE FOLLOWING: A copy of your code and a Project Report. Make sure your name is on all files submitted.

PROJECT:

Image coloring: You will run the k-means clustering algorithm to re-color an image with k colors only.

1. Choose a picture and read each pixel's RGB info. This will be your data set $\mathcal{D}$.
2. For $3 \leq k \leq 10$
   - cluster the data set $\mathcal{D}$ into k clusters.
   - use the centroids from each cluster to re-color the image.
   - run the algorithm a few times and observe the resulting image.
3. Print the resulting image for each k and include in the written report.
4. Decide which k would be better to use for your image. Explain your choice in the written report.

Extra credit: You will program the ID3 algorithm to produce a decision tree.

1. Choose a data set that is large and has meaning to you. In the Report, you must describe the data in detail, where you got it, how you got it, if it needed to be cleaned or manipulated in any way, and why you chose that data set. If you do not find a good data set, consider using the New York Times API to get articles to be categorized based on key words in the title; or creating a decision tree for tic-tac-toe. Include a copy of the data set with your submission.
2. Run the ID3 Algorithm to produce a decision tree for your data set.
3. Print the decision tree.

Project Report: in your report, you must include, but are not limited to:

- Your Name
- The programming language you used for the project
- **Image Coloring:** Printout of a resulting clustering for each k; best choice of k and explanation why; any relevant discussion of changes in output for specific choices of k, and when k varies.
- **Extra Credit:** Information about your data; printout of the resulting decision tree; any relevant discussion on how well the tree performs.