# **ECE219\_Project\_4\_Regression\_Analysis\_and\_Twitter\_Data.**

# **Instructions:**

* Questions 1-8 are based on ECE219\_Project04\_Regression\_Analysis.ipynb.
* Question 9-10 is based on ECE219\_Project04\_Twitter\_Data\_Analysis.ipynb.
* Open with jupyter notebook with the environment set up.
* Please run every cell one by one from beginning to end.
* Refer to ECE219\_Project04\_Regression\_Analysis\_and\_Twitter\_Data\_Report.pdf for the answers to the questions in the manual.
* For question 10, we have trained a model to predict the number of retweets, fan base, and relative time of a given tweet. Then, we applied it for data profiling.
* Note that the code is written in the Google Colabatory environment, and some installment or download instructions might differ depending on the local environment.
* GridSearchCV for neural networks takes about 1 hour. It can vary depending on the local machine's performance.
* Built-in google library (only available on the colab environment) was used to mount the google drive files for data reading. Change the directory name for local data reading.
* For Twitter data reading in question 10, a large RAM is preferred.
* Some model training and grid search processes take more than 30 minutes to complete. Some pickle files under the pickle\_files directory are provided to skip model training. To use the pickle files, uncomment the pickle load commands.
* There are some issues with the dot command in colab. To generate the random forest graph, run the corresponding code cell to generate the tree.dot file first, then run the dot -Tpng tree.dot -o tree.png in the terminal to get a .png graph. The code cell below will be able to display the graph. It is recommended to run the graph generation locally or download the tree.dot file and generate tree.png locally.

## **Files:**

* ECE219\_Project04\_Regression\_Analysis.ipynb.
* ECE219\_Project04\_Twitter\_Data\_Analysis.ipynb.
* Diamond dataset (retrieved from https://drive.google.com/file/d/1gofVOyha91l5ze5clqS2FJE\_MEXTdsEM/view)
* Twitter training files (retrieved from <https://ucla.app.box.com/s/24oxnhsoj6kpxhl6gyvuck25i3s4426d>.)
* Pickle files under pickle\_files (can be used to skip some training and gridsearch process)
* tree.dot (the random forest graph generated by export\_graphviz)