# E6998 – HW#2

## Due 10/17/2017 at 11:59PM (23:59) - EST

1. Using the following code:

   https://github.com/phoulihan/columbia/blob/master/class/classEx/train.py

   https://github.com/phoulihan/columbia/blob/master/class/classEx/predict.py

   https://github.com/phoulihan/columbia/blob/master/class/classEx/streamTweet.py

   a. Train a model to classify the same topics from class: *fishing*, *hiking*, *machinelearning* and *mathematics* using the *train.py* script. **25-points**

   b. Embed the trained model from above into the twitter streaming engine class, *StreamListener*. **25-points**

   c. Using your twitter application credentials, perform real-time classification on tweets (~100) that have the following regex set in setTerms and save results to a mongo database/collection: **40-points**

      setTerms = ["fishing","hiking","machine learning","mathematics"]

      You need to persist the following fields:

      *body*, *topic* (the classifier topic prediction. i.e. fishing, hiking, machinelearning or mathematics), *followers*, *screen_name*, *friends_count*, *created_at*, *message_id*, *location*. I provided a sample JSON, sample.json, file that shows the exact JSON structure and fields you need to persist

   d. Export your mongo collection from above to a json file called *sample.json*. **10-points**

**Push all python code and the exported json file to your github accounts. Anyone with their own path name and twitter app tokens should be able to run steps a-c. Any steps that fail to run will receive 0 points. There will be no partial credit for code that fails to work.**