Machine Learning Tool for Identifying Anomalies in Application Logs End of Semester Report

What we accomplished this semester:

- At the beginning of this semester, we were completely lost. None of us had lots of experience in the machine learning category and none of us knew how it worked. This is when we kicked it into overdrive. Our first 1-2 months in this class was spent doing nothing but research. We read lots of articles not only learning about what machine learning is, but also simultaneously scoping out algorithms for this project.
- After that research, it came down to 2 algorithms that we were going to experiment with. The 2 algorithms are LDA and the algorithm that uses 2 autoencoders with an isolation forest. The LDA approach is running and we have a working prototype. Yes, this needs to be tweaked before it is completely accurate, but as of right now, it is working.
- The 2 autoencoder and isolation forest algorithm is currently in development, and that is not running quite yet, this is set to finish next semester.
- We set up our Kafka pipeline which consists of a producer client that reads in a static log file and posts each log to an input topic, a streams app that reads in the log data from the input topic, does some data transformation, and then posts them to an output topic, and lastly a consumer client that reads in the data from the output topic and uses our LDA model to make a prediction on each individual log.

What we plan to accomplish next semester:

- Next semester, we hope to:
    - Containerize our pipeline that we have in order to make it easier to deploy in Capital One's system
    - Make some touch ups on the LDA model for accuracy to correctly work with our application
    - Also get a working prototype for the 2 autoencoder and isolation forest algorithm, then tweak that for accurate results
    - Lastly, come up with a viable alert system for our pipeline in order to display which logs are anomalies