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

By: Dylan Pierce, Sean Stitzer, Adrienne Hembrick, and Sam Sunvold

- **What We Accomplished**
  - From the outset, our team's cumulative skills in this problem domain was sparse if not nonexistent. Because of this, the task ahead seemed daunting. We stayed determined and prepared to give our best effort.
  - For the first 1-2 months in this class we researched this topic space as much as possible. We came to the understanding that the longer we spent compiling research, the less research we had time for. That said, we put together a collection of papers to analyze this problem as well as others which were unrelated but still utilized similar solutions.
  - After our research phase, the algorithm we chose was Latent Dirichlet Allocation (LDA). The implementation of this algorithm was imported from SciKit Learn and has been tweaked to fit our application properly. We have a working prototype using this approach which is running accurately.
  - 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. Then a streams app that reads in the log data from the input topic transforms the data, and then posts them to an output topic. Lastly, a consumer client reads in the data from the output topic and uses our LDA model to make a prediction on each individual log.
  - Once an anomaly is found, we send the log via Slack to notify the user.
  - Our whole pipeline is containerized using Docker and also has an Ansible Playbook to automate the process setup.