Capstone Three: Project Proposal

Maria Meza

Capstone 3 will be based on a data set from Kaggle https://www.kaggle.com/competitions/tpu-getting-started/data. The data comes from 5 databases. The files for this project are in a TF Record format, each with an ID and a label. The data contains 7382 unique values. The purpose of this project is to correctly identify 104 species of flowers based on their images.

After understanding the size and format of the data to be used, the next step is to identify how to solve the classification problem. The images will be classified using Deep Learning techniques including Tensor Flow and Keras. The macro F1 score will be the evaluation method for this project. An ideal score will be close to a score of 1, in the range of 0.80 to 1. Nonetheless, the model with highest the F1 score will be chosen.

The constraints for this project come from the modeling method that will be applied. Since image files are being used there will be a large data requirement. Processing large amounts of data also comes with time constraints. Furthermore, to ease the time constraints I will be using Kaggle's TPU option. Nonetheless there is a limit of 20 hours per week for the TPU option on Kaggle. The run-time for this project is also capped at 3 hours. Moreover, the next constraint includes overfitting as Deep Learning models tend to overfit.