**Capstone Project Ideas**

1. **Quroa Question Pairs challenge**: Millions of people visit Quora every day to search for answers to their questions, and many ask similarly worded questions. The goal of this project is to identify which questions asked on Quora are duplicates of the same intentional question. The training dataset contains more than 400,000 question pairs, and each pair is labeled as duplicates or not, with a total of 37% of duplicate pairs. Being able to predict duplicate pairs is useful because it helps users to find the best-quality answer more quickly without searching through all similarly worded questions.

Dataset: <https://www.kaggle.com/c/quora-question-pairs>

1. **Yelp Dataset 2017**: This updated dataset includes information about local businesses in 11 cities across 4 countries (US, UK, Germany and Canada). It also contains 200,000 pictures with captions (e.g. food, drink, menu) from 89000 businesses. There are many questions one can explore with this dataset. I am interesting to see how much of a business’ success depends on its location and whether reviewers rate their experience differently if they had to travel far to find the place. Finding the answer to this question is useful for urban planning.

Dataset: <https://www.yelp.com/dataset_challenge>

1. **Google Cloud & YouTube Video**: The goal is to develop classification algorithms to accurately assign labels to over 7 million YouTube videos. The dataset includes video labels from a vocabulary of 4716 classes, with 3.4 labels for each video on average. It comes with pre-extracted audio and video features from every second of video. We are given access to use the Google Cloud ML platform to analyze the data.

Dataset: <https://www.kaggle.com/c/youtube8m>