Bazeley, Mikiko
Type: Capstone Project

Step 1: Start Planning Your Capstone

Bazeley, Mikiko

Requirement: Write at least 3 to 4 sentences explaining your idea and identifying the data you'd use to solve it.

All Candidate Projects: **[Springboard] [MLE] Project Ideas - Mikiko Bazeley**

Top 3 Project Ideas:

1. **Wardrobe Recommender**
    a. Serving recommendations for an individual based on selections of items; leverage computer vision and image processing as well as the Fashion MNIST & Deep Fashion datasets, augmented by additional scrapes of existing e-commerce fashion sites as needed. Goal will be to apply lessons from the Recommendation, Computer Vision, Image Processing modules (as well as additional sources like PyImage Search, FastAI's Deep Learning for Coders course, and Full Stack Deep Learning). This project has the most precedence and available labeled data of all three contenders.
        i. Datasets:
            1. Independent Dataset: DeepFashion Database
            2. Kaggle: iMaterialist (Fashion) 2020 at FGVC7
            3. Kaggle: iMaterialist (Fashion) 2019 at FGVC6
            4. Kaggle: iMaterialist (Fashion) 2020 at FGVC7
            5. Kaggle: iMaterialist Challenge at FGVC 2017
            6. Kaggle: Fashion MNIST || Explanation of Dataset: Fashion-MNIST || Further Documentation: [1708.07747] Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms

2. **Bias Checker**
    a. "Create a model which predicts a probability of each type of toxicity for each comment (toxic, severe_toxic, obscene, threat, insult, identity_hate). Each comment in Train has a toxicity label (target), and models should predict the target toxicity for the Test data. This attribute (and all others) are fractional values which represent the fraction of human raters who believed the attribute applied to the given comment. For evaluation, test set examples with target >= 0.5 will be considered to be in the positive class (toxic). " This project has the 2nd most precedence and available labeled data of all three contenders.

A useful project would be to detect the toxicity in a comment before tweeting - like a spell-checker, only for toxic comments.

i. Datasets:
   1. Kaggle: Jigsaw Unintended Bias in Toxicity Classification
   2. Kaggle: Toxic Comment Classification Challenge

3. Mining the Panama Papers

a. This ICIJ database contains information on more than 785,000 offshore entities that are part of the Panama Papers, the Offshore Leaks, the Bahamas Leaks and the Paradise Papers investigations. The data covers nearly 80 years up to 2016 and links to people and companies in more than 200 countries and territories. This project has some precedence but for the most part is unlabeled & unstructured data consisting of administrative paperwork and emails. There could be some interesting applications of NLP techniques and graph/network analysis but there's no central question that could be immediately answered.

i. Datasets:
   1. Independent Dataset: https://offshoreleaks.icij.org/pages/database