Project Progress Report

Recalling from our project proposal, we plan to finish an application where users can enter a URL link and our system will identify faculty directory pages, identify faculty webpage URLs on the directory website, and format a structured faculty list with relevant information from faculty bios if the given URL is valid. Besides that, based on the feedback of the proposal, we decided to put on more effort on the faculty directory training model. By now, the progress we made includes one training model on faculty directory and negative dataset collection.

We have downloaded the positive datasets from https://docs.google.com/spreadsheets/d/198HqeztqhCHbCbcLeuOmoynnA3Z68cVxixU5vvMuUaM/edit#gid=0 and collected 300 negative data points by hand.

We have finished preprocessing html data and use text on the website as the input for our model. We removed the header and the redundant spaces in the text. And use Tfidftransformer from sklearn to do the feature vectorization and build vocabulary for us.

Currently we have only implemented the SVM model for the directory page classification task. The model yields a pretty optimistic accuracy data around 95% using 5-fold cross validation. Therefore we have no plan on adding a neural network or limiting the domains now.

Difficulties:

We found out that collecting negative datasets is time exhausting. Since we would like our negative data points to be as diverse as possible, we collected every one of the website URLs by hand. We collected 300 negative data points instead of 800 and the accuracy is pretty optimistic, therefore, we have no plan on adding more negative data for now.

We are still in the process of implementing other models and choosing the most optimal one in the near future.

We also have the second task which is identifying whether the page contains the profile links for the faculty and crawling the data. Currently we have not yet come up with a strategy to conquer it.