SI330 Project Proposal

The proposed project is called a Systematic Exploration and Analysis of US top Universities Academic Data. The goal of the project is to provide a quantified and comprehensive way to explore US top universities through the reliable documented data sets provided by both US government and prestigious university ranking organizations. It aims to provide an insight of this topic for perspective student applicants, academic professionals, company HRs and anyone else that are interested in knowing more about the classic higher education system of the country.

Two data sources that are going to be manipulated are:

1) Name: College Scorecard Data
   Size: records in hundreds
   Location: https://collegescorecard.ed.gov/data/documentation/
   Format: JSON
   Access: GET API

2) Name: university ranking organization websites
   Size: around several hundred KB
   Location: https://www.usnews.com/rankings,
   https://www.topuniversities.com/university-rankings,
   https://www.timeshighereducation.com/world-university-rankings,
   http://www.shanghairanking.com/
   Format: HTML
   Access: Website Scraping

Initial processing could include JSON response decoding, missing values handling, encoding unification and preliminary filtering of fields and records. The datasets are to be first converted into CSV files and then parsed and combined by csv.dictReader. The result would be new CSV files containing the statistics of the source files.

A perspective visualization of the project: a visualization system that could uniformly display the statistics of all the top universities and support user interactions such as recommendations based on user customization of the weight for fields.