SI 507 Final Project Proposal

Tongyan Xu, 40433774

1. Data Sources:

Source Name	Data Description	Task Detail	Challenge Point
US News < HTML >	Obtain interesting info of National Universities which are listed on US News site. 13 features are considered: name, ranking, state, city, type, found year, endowment, tuitions and fees, and etc.	Need to parse 311 universities listed on 16 pages. For each university, 3 pages are required to obtain all interested features / info. Totally 949 pages of HTML are requested.	8
Google Places < API >	Obtain latitude - longitude data for each university for plotting them on the map.	1 request for each university. 311 requests are needed totally.	2
US BEA < CSV File >	Obtain GDP data for each state from 1997 to 2016. (past 20 years, unit: million dollars)	Directly extract data from CSV file. (61 rows * 22 columns)	0
Github < JSON File >	Obtain corresponding relations between two-letter abbreviations and full names of each state.	Directly extract data from CSV file. (59 records)	0
Total			10

2. Presentation Options:

Type	Description
Table	A list of National Universities of a given state or several states.
	A list of GDP data of a given state or several states. (still under consideration)
Plot	A bar plot showing how many National Universities belong to each state in US. X - axis: State, Y - axis: Number of National Universities
	A scatter plot attempting to show the relationship between number of National Universities and GDP value of states. (Endowment is also considered) X - axis: Number of National Universities, Y - axis: GDP of State
	A scatter plot showing difference between public and private National Universities. X - axis: Enrollment, Y - axis: Median Starting Salary
	A line plot showing GDP data (trend) of a given state or several states. X - axis: Year, Y - axis: GDP of State
	A map showing location of National Universities in a given state.

3. Presentation Tools

Plotly - a kind of API based data visualization tool used to make plots.

Django - a framework to create HTML to make interactive functions and show data tables.

* Planned data communication method: AJAX

* Planned table (visualization) framework: DataTables.js

* Planned plot (visualization) framework: None (redirecting to Plotly instead of using d3.js)

4. Current Progress

All data were parsed / requested and cached.

Database was built, and all data were uploaded / inserted.

Several important classes were defined (like class < NationalUniversity >).

All plotting functions were finished.

Some unittest were developed.

5. Future Tasks

Improve code structures.

Develop unittest.

Set up Django framework and write HTML and JavaScript.

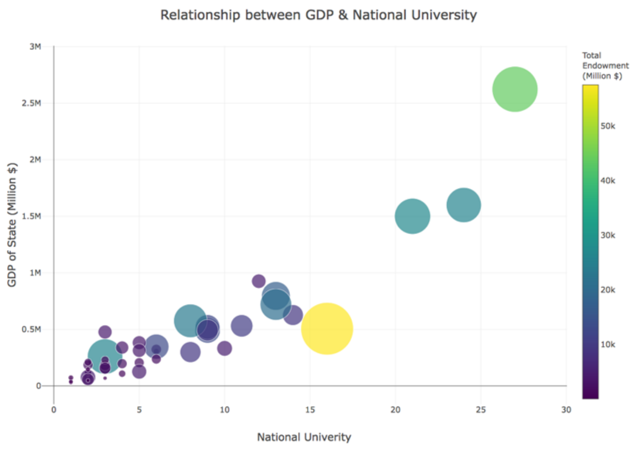
6. Current Results

Github link: https://github.com/TongyanX/SI507_Final_Project

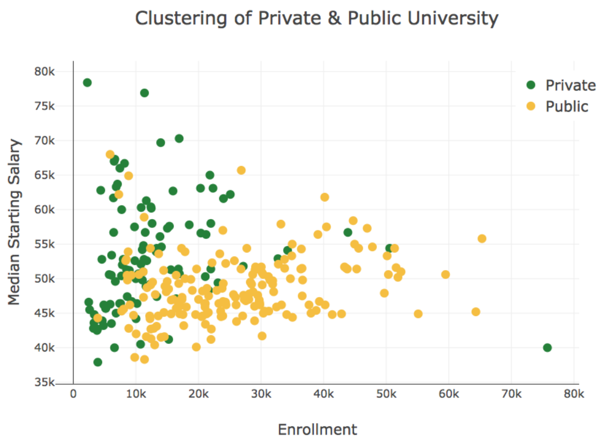
File list:

Script:	universityData.py	| commonFunc.py
	gdpData.py	| plotFunc.py
	stateAbbrData.py	| test_file.py
	classDef.py	| secret_file.py
Data:	GDP_by_state.csv	| state_abbr.json
DB:	National_University.db	
Cache:	national_university_gps.json	| national_university_info.json

Sample plots:



Relationship between GDP & National University



Clustering of Private & Public University