Team Information

Project Name: US Colleges & Universities Internet Database
Canvas Group: E3

Team Members:

- Harrison Berrier
    - EID: hlb962
    - Email: harrison.berrier@utexas.edu
    - Github username: harrisonberrier
    - Estimated completion time for each member: 12
    - Actual completion time for each member: 13
- Mohit Gupta
    - EID: mg58629
    - Email: mohit.gupta@utexas.edu
    - Github username: mohitg17
    - Estimated completion time for each member: 10
    - Actual completion time for each member: 12
- Nikhil Jalla
    - EID: nj5473
    - Email: nikhiljalla17@utexas.edu
    - Github username: nikhiljalla17
    - Estimated completion time for each member: 11
    - Actual completion time for each member: 12
- Silas Strawn
    - EID: scs3434
    - Email: strawnsc@gmail.com
    - Github username: StrawnSC
    - Estimated completion time: 11
    - Actual completion time: 10

Github Repo Link: https://github.com/UT-SWLab/TeamE3
Deployed Site: https://university-idb.uc.r.appspot.com/

Motivation and Users

We envision our site being used by prospective students to assess which college or university is right for them, as well as which major or field of study they should pursue. Our database lets students view universities in the US filtered by their city, or by the majors they offer. For example, if a student knows they are only interested in schools that offer a particular program, they can go to the major page for that field of study, and browse the colleges that offer that

major. On the other hand, if the student is interested in colleges only in a particular location, they can pull up all the colleges for a particular city. Once they're on a university or college's page, the prospective student will be able to see crucial information for making their choice. For instance, they will see facts on the cost of attendance, the acceptance rate, average SAT scores, the size of the school, etc. If they want to learn more, they can also follow a link to the institution's web page.

Requirements

User stories

Phase I:

- As a high school student, I want to be able to look through a list of universities in the United States so that I can decide where I want to go to college.
    - Estimate: 2.5
    - Actual: 2.5
- As a high school student that wants to move away from home, I want to be able to look at university education statistics for different cities so that I can decide where I want to move.
    - Estimate: 2
    - Actual: 3
- As a high school student that hasn't decided what I want to major in, I want to be able to look through a list of different majors and their associated statistics so that I can decide what I want to major in.
    - Estimate: 2.5
    - Actual: 3
- As a user, I want to read about the website that I am using so that I know where my information is coming from and what the creator's motivation is.
    - Estimate: 2
    - Actual: 3
- As a user that isn't very good with technology, I want the website's interface to be simple and navigation to be easy so that I can access the information that I want quickly and easily.
    - Estimate: 2.5
    - Actual: 3

Phase II:

- As a user with limited time and a general idea of what I'm looking for in a university, I want to be able to view some attributes of a university without having to navigate to that university's instance page.

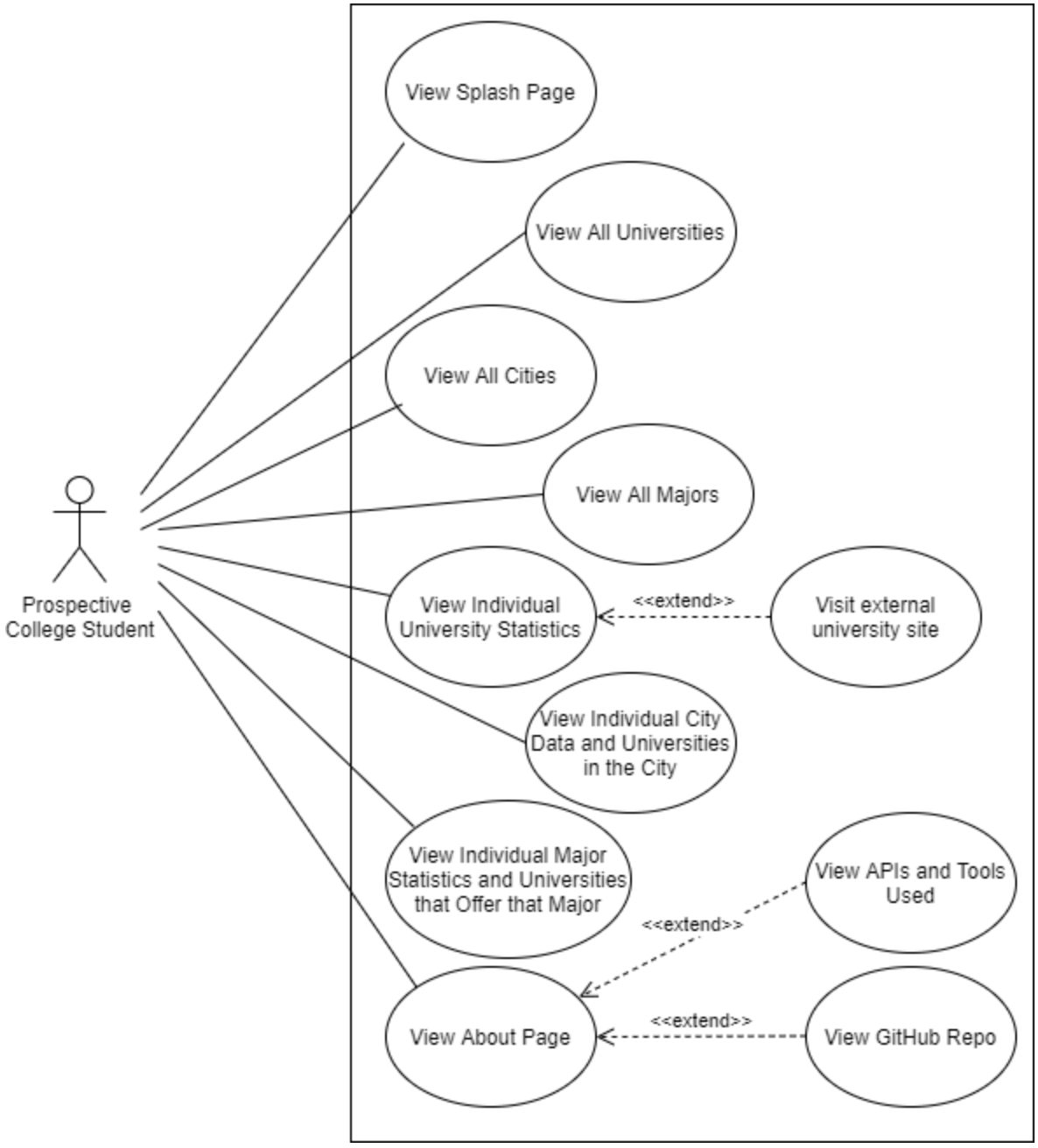
    - Estimate: 4
    - Actual: 4
- As a high-school student, I want to be able to view a list of all colleges offering a 4-year degree in the US so that I don't have to go elsewhere to find statistics on certain schools.
    - Estimate: 4
    - Actual: 3.5
- As a user with limited time, I want to be able to navigate by more than one page at a time so that I don't have to waste time clicking through pages of instances.
    - Estimate: 3
    - Actual: 2.5
- As a user looking through a long list of cities, universities, and majors, I want the list to load into pages rather than one large screen so that it makes it easier to navigate.
    - Estimate: 2.5
    - Actual: 3
- As a user that is scrolling quickly through a long list of universities, I want each university to be listed with a picture to help me sort quickly through them.
    - Estimate: 3
    - Actual: 5
- As a user that is scrolling quickly through a lot of cities, I want each city to be listed with an image to help me sort quickly through them.
    - Estimate: 3
    - Actual: 5
- As a user that is scrolling quickly through a long list of majors, I want each major to be listed with an image to help me sort through them quickly.
    - Estimate: 3
    - Actual: 2.5

Phase III:

- As a user looking at a city's page, I want to be able to see a list of the universities within that city so that I don’t have to navigate back to the universities page and filter by state.
    - Estimate: 2
    - Actual: 2.5
- As a user that is interested in exploring the universities of a city, I want to see a map on the city's instance page showing me where the universities are located.
    - Estimate: 3.5
    - Actual: 3.5
- As a user, I want to be able to sort universities by acceptance rate so that I can see which universities are considered to be the best.
    - Estimate: 3
    - Actual: 2.5
- As a prospective student seeking to compare different cities, I want to see how a particular city's median age and rent compares to those of other cities through some visual representation.
    - Estimate: 2.5

    - Actual: 2.5
- As a high school student who knows what school I'm interested in, I want to be able to search for a specific university so that I can view their statistics.
    - Estimate: 2
    - Actual: 3
- As a prospective student, I want to be able to search for cities that I am interested in so that I can easily access more details about those cities.
    - Estimate: 3
    - Actual: 3
- As a concerned parent of a high school student, I want to see which universities are in my home state so that I can keep my child close.
    - Estimate: 2.5
    - Actual: 3.5

Use Case Diagram



Testing

GUI Testing

Our GUI testing uses the Selenium IDE to reduce the potential points of error. We test all of the links on the navbar by clicking them and asserting the title of the page. The model page links are then tested in the same way.

Each page's pagination is tested as well. The test makes sure that both navigation arrows work as well as the early and late page navigation. To test the page navigation, the script checks the text of the active tag, confirming that the page changed. It also checks to make sure that the first page and last page elements exist, regardless of which page is active.

To test our searching functionality we test each model individually. For the universities we test that a school has been searched for correctly by selecting the search bar, entering the text for the school, clicking the search button, and asserting that the school's id is present on the page. We test this against a one-word search, a search from the resulting route (which is different from the base model route), and a multi-word search input. For the last search we navigate to the instance page and assert the header text to make sure that the page link is correct. Cities and majors are tested similarly, with a one-word search input, a search from the resulting page, and a multi-word input. We again navigate to the instance page of the last search result and assert the header text to make sure that the links are connecting to the correct instance pages. The search function is also tested for an incomplete search (for ex. searching "bos" and returning Boston, Massachusetts).

Unit Testing

We used python unittest to test our database queries and data processing. We tested create, read, update, delete, and search operations for each of our main models. To test the create operation, we create an instance of each model using dummy parameters and save it to the database. We then queried for that instance and checked if the instance was found. To test update operation, To test update operations, we loaded an instance from the database, modified one of its properties, and saved the object to the database. We then loaded that same object again and checked to confirm that the property was modified as expected. To test read operations, we queried the database for specific instances and compared the properties of the retrieved instance with the expected data. To test delete operations, we created a dummy object instance and saved it to the database, and then queried that object and deleted it. We then queried the database for the deleted object and confirmed that nothing was returned. Finally, to test searching, we queried the database with specific search parameters and compared the instances that were returned with the instances that were expected.

We tested database queries more extensively because we rely on the data that we pull from the database to be exactly as expected. We tested queries for the university, major, and city

objects, and compared the returned objects with the expected object. The objects are queried using varying combinations of their unique fields to ensure that all combinations work.

We also manipulate some of the data that we retrieve from the database before sending to our frontend. The data processing test queries the database and mimics the data manipulation that we execute in our main file. The modified sample instance data is compared to the expected data for a particular instance.

These tests together validate that the database is returning the expected objects when queried and that the expected data is being sent to the model and instance pages.

Design

Stack

We are using MongoDB (through MongoEngine), Flask, Jinja, and Bootstrap.

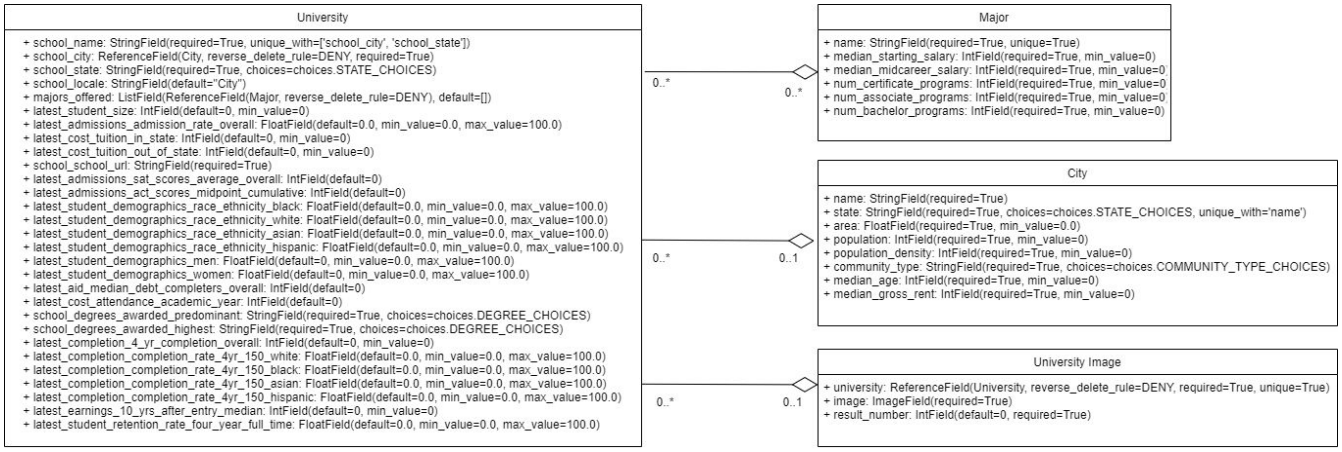
Pages

Every page inherits the base.html file, which includes the navbar. Each of the model base pages uses the model.html. Each instance uses the [model]_instance.html file. The landing page (index.html) is a central page where users can navigate to any of the model's base page. It contains a carousel and pictures that link to the models. The about page (about.html) contains information about the site and the developers.

Flask

The flask application queries the database to retrieve the requested data. It retrieves data based on the type of requested data and the pagination settings. We used Jinja templates to design our pages and we render the templates with the requested data from flask.
The home page is at the root path ('/'). Each model is at the path './[model]'. Each instance is at the path '/[model]/[instancename]'.

Database UML



Models

University Model

We collected data, including images, for over 1,900 universities, all of which is stored in our database and accessible via our website. Each university instance contains relevant statistics about the university along with its location and the majors that it offers. The page has an overview at the top that includes the school size, acceptance rate, and in-state tuition. Below that, there are sections for school info, majors, admissions, demographics, and tuition and aid. The left panel has links that navigate to each section. The right panel includes extra data about degrees awarded, completion rate, average earnings for graduates, and overall retention rate. All of the data is pulled from the CollegeScorecard API. The page showing all universities includes a picture of each university on its corresponding card, and each individual university page includes a picture of the city that it is located in.

We would like to present some of the data graphically to make the page more visually appealing. Demographics data can be easily represented in a pie chart. More data about admissions and completion rates can also be displayed in intuitive ways. We can also make the pages more interactive for the user by adding more responsive elements such as collapsibles and buttons. We also need to link the university instances to the city and major instances.

City Model

Given the number of university instances in our database, we ingested data and images for over one thousand cities. As with universities, all of these cities are viewable on our site. Each city instance is described by statistics that would be relevant to students who would like to attend a school in the respective city. Currently, the page lists the area of the city, population, population density, community type, and of course the schools that are in the city. Population data, city area, and median age were found using Wolfram Alpha's Full Response API. Rent data was a little harder to find, so it was scraped from a website called city-data.com. The schools in each city are listed based on the location data found in each university.

The city page also displays a satellite image of the city which was found by using Google Maps API. The page displays both a map view and satellite view.

Major Model

Our database holds data associated with 357 separate majors and images for every major. Like the other models, all of these majors can currently be viewed on our site. Each major instance page contains a list of universities that offer that major, as well as some aggregated statistics on the number of programs in the US for that field of study. We previously listed just the number of programs that offer a bachelor's degree, an associate's degree, and certificate (after less than one year of study) in the given major. However, since we decided to only list colleges and universities that offer a bachelor's degree as their "predominant" degree awarded, we now just display the number of schools that offer a bachelor's degree in a given major.

We also display the average starting salary for each major, which is based on the latest available earnings cohort of graduates who received a bachelor's degree in a given field. Since this data was collected at the university level, we had to aggregate it across all two thousand schools to get an average earnings figure by major. All of this data was acquired from the Department of Education's College Scorecard API. We also wanted to present a mid-career salary statistic. Since the Department of Education's API does not have this information, we approximate the mid-career salary as 1.693 times the starting salary, based on the Wall Street Journal's data on the average increase between early- and mid-career salaries. In future phases, we will seek a more accurate representation of mid-career salary, as we wouldn't expect the percent change to be exactly the same for every major.

Ideally, we would like to sort the universities in each major instance's list by the ranking for that particular major, but we have not found open datasets with such information. In future phases, we will explore the possibility of scraping data from the US News & World Report site, although it seems most of their data is behind a paywall. For now, we simply display an unsorted, paginated list of all the universities that offer a particular major on that major's instance page. This list is populated by a MongoEngine query on the University collection that selects only schools where the current major is in that school's list of offered majors.

Finally, we also intend to group majors by high-level categories in future phases. Given that we have several hundred majors, they tend to be narrow fields of study. It may be useful for the user to be able to look up schools that offer majors in"engineering" (a broad category) rather than "naval architecture and marine engineering" (one of our current majors).

Tools, Software, and Frameworks

Flask

Our backend framework is flask, a microservices framework for python. Flask is useful in that it is lightweight: it was trivially simple to set up the server and start running it locally. When we want to add new pages, we simply add a route for the new page. We also took advantage of Flask's template mechanics. Using flask templates, we dynamically generate instance pages for our major, university, and city models. This allows us to reuse the same HTML, instead of adding hundreds of nearly identical HTML files for each university.

Bootstrap

Our frontend framework is bootstrap, which we use alongside HTML and CSS to design our web pages. Bootstrap has been crucial for allowing our pages to be reactive. The pages in our site dynamically resize based on the size of the viewport, allowing our format to remain robust and look clean, despite the window size limitations of the user.

Google Cloud Platform

Our site is deployed to Google Cloud Platform. We used GCP because it is simple to set up and use with flask applications. All we have to do to update the deployment of our application is run the shell command "gcloud app deploy" in the project root directory, using the gcloud CLI.

MongoDB and MongoEngine

All of our data is dynamically pulled from MongoDB, including the images for every model instance. Our models are defined with MongoEngine, an open-source Python Object-Document Mapper. MongoEngine will allow us to define "schemas" (MongoDB does not technically have schemas since it's NoSQL) with python objects, so we can interact with our database purely in terms of University, Major, and City python objects that we design. MongoEngine ODM capability is useful for not just defining the schema for our model's fields, but also enforcing constraints on certain fields, such as uniqueness constraints. For instance, MongoEngine forces our model instances to have university names to be unique with respect to the university's city and state.

Google-Images-Search

We used the Google-Images-Search python module (available on PyPi) in order to automatically collect large amounts of images. Since we have about two thousand universities, one thousand cities, and several hundred majors in our database (and accessible via our site), manually collecting images for each model instance would have simply been intractable. The Google-Images-Search module allows the user to specify a query (and some other attributes like the kind of image and the image's licensing) and download one or more images using Google's image search. We wrote scripts to search for universities by their name, download and

resize the first image result, then upload the image to a new collection that maps the university instance in MongoDB to its corresponding image. We did the same to populate major images.

Python Google Maps Package

For city images, we decided to use satellite imagery, since there were a lot of open data sets available online (and we thought they would be visually interesting). We explored NASA and Google Earth APIs, but decided to use the googlemaps python package, (also available on PyPi). Although we did end up scraping latitude and longitude coordinates for each city, the google maps module allowed us to simply look up the location by city name and download a corresponding satellite image.

Sources

Flask documentation: https://flask.palletsprojects.com/en/1.1.x/quickstart/
We used the flask documentation to learn how to use routes and to help us format our URLs.

MongoEngine: http://docs.mongoengine.org/
This MongoEngine documentation helped us understand how to set up our model schemas, connect to the remote database, and query the database.

Bootstrap reference guide: https://www.w3schools.com/bootstrap4/default.asp
https://getbootstrap.com/docs/4.0/components/carousel/
We used the Bootstrap reference guide to help us understand how to use and format Bootstrap elements such as cards and the carousel.

Footer guide: https://www.w3schools.com/howto/howto_css_fixed_footer.asp
We used the footer guide to learn how to format a page footer.

Forms guide:
https://developer.mozilla.org/en-US/docs/Learn/Forms/Sending_and_retrieving_form_data
We used this as a guide for developing the search functionality

CSS variables: https://www.w3schools.com/css/css3_variables.asp
We used this CSS variables reference to help us format our elements and make them more visually appealing by changing font, color, size, etc.

Major salary data: https://www.visualcapitalist.com/visualizing-salaries-college-degrees/;
http://online.wsj.com/public/resources/documents/info-Degrees_that_Pay_you_Back-sort.html

Google Maps API documentation:
https://developers.google.com/maps/documentation/maps-static/usage-and-billing
We used the Google Maps API documentation to figure out how to grab satellite images and automate the process for all of the cities.

College Scorecard: https://collegescorecard.ed.gov/data/documentation/
We used the College Scorecard API and data dictionary to collect data on universities.

Flask Pagination: https://gist.github.com/mozillazg/69fb40067ae6d80386e10e105e6803c9
We used this pagination example to implement pagination on our website using the flask_paginate module.

Google-Images-Search: https://pypi.org/project/Google-Images-Search/
We used this python package and it's documentation to ingest images for each university instance and major instance.

Google Maps Package: https://pypi.org/project/googlemaps/
The documentation for this google maps package allowed us to collect satellite images for each city in our database.

Beautiful Soup Documentation: https://www.crummy.com/software/BeautifulSoup/
Beautiful Soup was used to pull relevant data from websites if it could not be easily obtained through APIs.

Wolfram Alpha Full Results API: https://products.wolframalpha.com/api/documentation/
We used the Wolfram Alpha Full Results API to collect data on cities.

Reflection - Phase I

What we learned:

We've learned how to work together and divide work evenly among us. We were able to distribute tasks early on and each person completed their parts when convenient for them. We also learned how to build a website from ground up using html, css, bootstrap, and flask. Through designing the site, we've learned more about bootstrap and how to style components. We've also learned how to use the APIs selected to acquire the data that we need.

Five things that went well:

1. We quickly moved to the use of templates so that multiple pages can be changed quickly.
2. We have been able to make decisions quickly and then adapt as problems arise, given our heavy usage of slack to coordinate and communicate.
3. Our API's are well documented, making them easier to interact with. In particular, the Department of Education's College Scorecard API has extensive higher education data,

and provides a spreadsheet explaining all the possible fields one can request from the API.
4. Since flask is a lightweight backend, we were able to get started putting pages together and making routes for them easily.
5. Flask also allows us to pass in parameters to the html, and we used this to dynamically generate different instance pages off of the same base html. Although the data is currently hard-coded in the python flask app, we will be able to easily migrate this over to database queries without having to change the frontend code.

Five things that went poorly:

1. We haven't been able to add a stylesheet other than bootstrap so we are stuck writing styling into each file at the moment.
2. Formatting pages to fit requirements has been very difficult versus formatting a page without specifications because of the way that html elements and they're styling interact with each other.
3. At first, the college scorecard API was overwhelming because of the sheer amount of data and the complexity of crafting queries to get the needed information.
4. Often, the styling of elements with HTML, CSS, and Bootstrap does not work as we would expect. For example, when we were making the about page, we had to put in elements with our photos and bios together. For some reason, adding more text to the paragraph element next to an image would increase the image's size. We ended up using different bootstrap elements to get around this.
5. When we were first setting up the deployment on Google Cloud Platform, the deploy would seem to work in the CLI, but when we tried to visit the site, we would get a 500 error. Apparently, if you don't have the exact right directory structure and naming of certain modules, GCP doesn't know how to deploy your flask app.

Reflection - Phase II

What we learned:

We learned how to set up, access, update, and query a MongoDB database. To set up the database, we learned how to create collections and set up data models. We learned how to use a python object-document mapper, MongoEngine, to manipulate data in the database. To populate the database, we learned how to write ingestion scripts that call APIs to collect data and store them in the database. We also learned how to test both our frontend and backend code using Selenium and unittest respectively. To collect images, we learned how to use existing APIs to automatically retrieve relevant images. We learned how to use pagination to limit the number of instances that are loaded at a time.

Five things that went well:

6. We started testing both our frontend and backend early on, which helped us resolve bugs much quicker than before. We used Selenium to confirm our frontend functionality and we used unittest to validate our database queries and data manipulation.
7. We were able to easily communicate with the database using python objects through MongoEngine. The object-based model allowed for logical data access and manipulation. It also allowed for greater flexibility in the structure of the data.
8. Since we used templates for phase I, we were able to transition from hard-coded data to dynamically accessed data very easily. Our data requires very little backend processing once it is received from the database, so large amounts of data can be aesthetically displayed in the model and instance pages. Data access is also very fast.
9. We wrote ingestion scripts to consume data from our APIs and populate our database very efficiently. After running our ingestion script initially, we were able to easily adjust the data and fields as needed.
10. We were able to divide responsibilities efficiently so that each person's work was not entirely dependent on another person's work. This allowed us to work on our own schedules and make good progress on all components of the project simultaneously.

Five things that went poorly:

6. Currently, all of the instances are loaded upfront when a model page is accessed, so there is a short loading time on the page. We will need to dynamically load instances for a given page in order to decrease loading time. We should also add indexes to our MongoDB collections to make queries faster.
7. We used a python package that uses the google images API to get images for our instances. While it mostly worked very well, we need to go through and replace images that are not accurate.
8. We needed to adjust our data models slightly after we had already run our data ingestion scripts, so we had to re-run some of those scripts with the updated data model. This was time consuming since we had to query an API for some of the new data.
9. After integrating the database code into the main app script, we weren't able to deploy our app to Google in its existing form. We had to make changes to the deployment process to make it possible to deploy to Google App Engine.
10. Some of the APIs we had planned to use were hard to work with so we had to find alternative APIs with more easily accessible data.

Reflection - Phase III

What we learned:

We learned the value of refactoring code. Up until this phase we had a significant amount of repeated functionality because of differences between our three models. For example, rendering each model page has a lot of similar code, but there are differences between the models that kept us from consolidating the functionality. During this phase, when implementing searching, we had to render a slightly different version of those model pages so we took the time to break the common functionality into separate methods which has improved the readability of our code and reduced the number of lines of code. We have also started to utilize comments more because, until now, the code had grown in an organic way while the whole team worked on it, so we all understood it. Our refactor moved significant portions of the functionality and comments helped to keep us all on the same page.

Five things that went well:

11. Using mongoengine made it much easier to query our database and helped with searching and sorting. Where we might've had to write our own searching and sorting functionality for a large data structure, we were able to rely on a well-tested framework.
12. The refactoring from this phase should improve our ability to make significant changes in the structure of our code. Changes that previously would've taken too much time to make will now be more manageable because of increased modularity.
13. We were able to ingest additional data for our city and major models, which allowed us to make the instance pages more rich with information and offer more insight to the user.
14. Since most of the data had already been ingested in prior phases, we could focus on the coding the site's searching, sorting, and filtering.
15. Even though we have thousands of university model instances, the searching, sorting, and filtering runs fast.

Five things that went poorly:

11. Because of the way that we render instances on the model base pages and because we're not using a react or angular framework, we have to redirect to load search results rather than being able to dynamically load them into the current page.
12. To split up the work, searching and sorting were assigned to different people so they have not been integrated with each other and will likely need to be refactored in the future.
13. Some of the APIs and data sources that we used did not have consistently formatted data, which made it difficult for us to write ingestion scripts to add additional properties to our model.
14. A true search engine in our site would be able to do fuzzy string matching. To be fair, our search is much better than requiring exact matches; it allows the user to match by a case-insensitive substring with whitespace trimmed. For instance searching for "harv "

matches “Harvard University” as well as “Harvey Mudd College.” Our search can’t do approximate string matches, like matching “Harvard University” from a slight misspelling, such as “Harverd.”

15. Over time, our main flask file has grown quite large, and is occasionally repetitive. We should have put more thought into how we could have broken up the file into separate modules.