CS591 Final Project

ScrapeAlchemy

Peter Gilbert and Alex Thomas

12/18/19

For our project we created a library that allows the simple fusion of a spider from Scrapy and inserts it into a database through SQLAlchemy. We started by making a library that takes the data we’ve scraped through scrapy and figures out what type it is based on the file that we’ve generated or an item that is inserted. Inside our scrapealchemy library we have a Library class that allows for the creation and insertion into a SQLAlchemy table. In our library we use calls to the SQLAlchemy library. For our create table function we create a temporary class that uses base in order to make a temporary table, where it auto generates the id with the table name as what the user passed in. When inserting an item into the table it checks whether it is a scrapy item object or if it is in the form of a dictionary. Also in our library we have a method that infers the type of the objects using a hierarchy where we check for different types and if it’s not that type, it goes on and checks if it’s the next type and so on. Once it finds the type it will then use it as a SQLAlchemy type. That’s pretty much the overview of how we solved the problem.

In our Fusion file we have an example of our test that takes a search on Amazon for “smartphone” and then scrapes the results. The results that we are getting is the name, the url of the image for the smartphone, and the price of that specific smartphone. In fusion we have our test being called which sees if it should generate the json file if it doesn’t exist, then we create our tables and insert the data that we have gotten from Amazon. We then have a short query that outputs and reads the data we inserted.

We faced a couple challenges while doing this project, but in the end I think we were able to make a pretty good functioning library. At first, we didn’t exactly know where we were going to go with the project idea, since it was pretty broad, but we narrowed it down to a simple library that is able to convert the data that we scrape with scrapy and place it into a database. Our original idea had to do with just comparing prices on different ecommerce websites, but we are able to do that with what we have now, as well as do even more. Another challenge that we faced was the inability to make relationships because of just reading the data, but in the future maybe we can edit the tables and create a new function for our library that allows to add relationships. We weren’t too familiar with SQLAlchemy before this project, since we had only used it before in the third project, and we had never used scrapy before, so this project was a challenge but we were able to make a complete project. Thanks for a great semester!

-Peter and Alex