Grace Abanavas

12-18-14

LIS-664-01

Programing for Cultural Heritage

Final Project Abstract

Harry Potter Wiki Scrape

For my final project I data scraped the Harry Potter Wiki site. I initially intended to create a script where I scraped all the listed character’s names and birthdates on the site. Then I planned to transpose that information onto a timeline, using Knightlab’s TimelineJS program, where a person could view the characters of the Harry Potter series according to their birthdates. However, even through the poor construction of the site—some character pages are repeated twice or categorized under a general “catch-all” page that confuses the script—I found an API for all the students who attended Hogwarts School of Witchcraft and Wizardry. The students had individual pages that were easily scraped. Thus, I took that API link and—with the help of the professor—scraped the Hogwarts student pages to find each individual’s name and birthdate. I first scraped the names of each character. Then I extracted the “abstract” (where the birthdate information resided) from pages. From the abstract I created a regular expression to pull out the date information.

To insure that my requested information was compatible with my chosen timeline software, I added the model to my script that Kinightlab recommends for functionality. I found the model within Timeline’s Github account under their “File Formats” heading. Within the data scrap I created a readable JSON file that resulted in a text file. Unfortunately Timeline JS does not recognize a JSON file for easy upload to their recommended Google Drive spreadsheet. At a loss, the professor helped me format my timeline into a localhost http web page for my script to run properly. I uploaded all the files to my Github account to help create a future public domain for my resulting Harry Potter timeline.