Automating Data Collection for Cure Alzheimer's Fund

Srinivas Kaza, Matthew Pfeiffer, Hunter Gatewood

`https://github.com/Spferical/cure-alzheimers-fund-tracker`

Cure Alzheimer's Fund

The Cure Alzheimer's Fund (`curealz.org`) is a registered public charity whose mission is to "fund research with the highest probability of preventing, slowing, or reversing Alzheimer's disease. . . " The CureAlz organization has an annual data-collection-a-thon where a small team spend an entire day compiling information about the research their donations have funded.

Because this process is time consuming, the data is compiled only once a year. Thus, our point of contact, Maddie, described the organization's desire to automate this data collection process to allow for both a reduction in required man-hours, as well as higher resolution in funding data throughout the year.

Our Project

Our goals for the project, then, were to

1. collect the desired data from Google Scholar and NIH RePORTER, and
2. allow for easy, constant access to the data.

Data Collection

Google Scholar

Google Scholar (most unfortunately) has no official API, which caused a decent problem initially. Our workaround for this was to extend `scholarly.py`, a Python module to scrape Google Scholar, adding the ability to use `scholarly` with an exact phrase.

The next issue we ran into was Google Scholar throwing CAPTCHA's at us after a decent number of queries (thanks, Google), despite `scholarly` implementing a number of features attempting to work around the CAPTCHA's. Our solution for this was keeping a database of scraped data and scraping progress and setting a cron job to periodically resume scraping (Google Scholar only marks connections as a bot for some number of hours).

NIH

The NIH made things a bit easier on us with ExPORTER, a site offering CSV downloads of their RePORTER database by year. However, around 1% of the CSV lines were malformed, which was unfortunate. To work around this, some of the lines contain non-decoded characters.

Data Distribution

After considering a number of ideas, we decided on writing a short Django app hosted on `scripts.mit.edu` which allows users to download both the Google Scholar and NIH data in CSV format.

Summary

The Cure Alzheimer's Fund nonprofit organization desired a way to automate a time-consuming data-collection project. We created a Django app hosted on `scripts.mit.edu` to distribute the data we collected from Google Scholar with `scholarly` and from the NIH RePORTER with ExPORTER.