Cure Alzheimer's Fund Data Collection (CAFDC)

README (technical)

Description

Data collection and distribution from Google Scholar and NIH RePORTER for the Cure Alzheimer's Fund written in Python 2, distributed with Django, and hosted on `scripts.mit.edu`. The data is downloadable as a CSV file.

Google Scholar

CAFDC uses `scholarly.py` to scrape Google scholar for entries with the exact phrase "Cure Alzheimer's Fund" (false positives are later removed manually by CAF). Entries are then converted to Django objects and stored in a local mySQL database.

Every so often, Google Scholar will throw a CAPTCHA. In the event of this or any error, `scholar_data.py` will save the current progress. A cron job later attempts to resume the process with `manage.py scrape`. Thus, the total process could take several days, but will only need to be finished a single time.

The format of the Google Scholar CSV is:

1. URL
2. title
3. number of citations
4. journal edition info
5. journal name
6. journal volume
7. journal issue
8. year
9. authors

NIH

Once all CAF-related papers have been collected from Google Scholar, CAFDC uses the NIH ExPORTER to download CSV files of the NIH RePORTER database, one per year. See `nih_data.py` for the locations of these files.

The format of the NIH CSV is:

1. URL
2. title
3. funded researcher

4. funding amount
5. year

Team

- Matthew Pfeiffer
- Hunter Gatewood
- Srinivas Kaza