Literature Scraper
PHAC
Requested by James Robertson
Written by Justin Schonfeld

Specification:

Extract information from Pubmed or Google Scholar based on a search.

*Author
*DOI
*Year
*Journal
*Title

*Abstract

Google Scholar

- Google Scholar robot.txt disallows bots
- Google Terms of Service says : Don't misuse our Services. For example, don't interfere with our Services or try to access them using a method other than the interface and the instructions that we provide.
- Several GitHub projects: scholar.py - https://github.com/ckreibich/scholar.py
- Google specifically disallows bots: https://scholar.google.ca/robots.txt
- Google uses Captcha to check for scraping of Google Scholar

---

PubMed

- PubMed E-Utils (Entrez Library)
- Now offers API keys (May 2018)
- NCBI Book on Entrez - https://www.ncbi.nlm.nih.gov/books/NBK25497/
- Biopython Tutorial and Cookbook - http://biopython.org/DIST/docs/tutorial/Tutorial.htm

---

Design Notes

---

*Went with PubMed for implementation

Examples

Get help on how the tool works
>python model_scraper.py —help

>python model_scraper.py --output e_coli_conjugation_one_term.tsv --pia "conjugation" "conjugation e coli"

>python model_scraper.py --output e_coli_conjugation.tsv --pia "conjugation,frequency" "conjugation e coli"

Cleanup

- Change scripts so they are no longer dependent on Justin Schonfeld's credentials.