Group and members

Group 2

Peter Laurin (peterlaurin@uchicago.edu)

Lily Mansfield (lilymansfield@uchicago.edu)

Sadie Morriss (sadiemorriss@uchicago.edu)

Max Kay (maxwellkay@uchicago.edu)

Project description and goals

This project hopes to analyse gender disparity across academic disciplines and institutions. We will scrape authorship data from journal websites and article search engines, assign gender to author names, and analyse the distribution of gender in specific academic disciplines, institutions, and geographic regions. We also plan on getting author data from Google Scholar, including common coauthors and citation data, and answering some of the following questions:

At higher impact journals, are women less likely than men to get published?

Are men more likely to cite men than women?

How often are women lead author in a paper versus first author?

How much do gender proportions vary across institutions and fields of study?

In specific fields how have gender disparities changed over time?

We hope to present this data in a website using Django, and have interactive features such as filtering results by geographic region, institution, and field of study, as well as a map of pins of different institutions.

Data sources

Authorship data: Jstor, Science (suite of journals), Nature (suite of journals), PLOS journals, Google Scholar

Useful Python libraries: Gender API, Namsor (gender and ethnicity look-up), Beautiful Soup

Tasks and timeline

Week 5: Implement web crawler to scrape authorship data from above journals and search engines - get author name, paper title and citations, journal published in, and google scholar page for authors (common coauthors, citation data, keywords in paper titles). Decide how to determine field (i.e. if we will search by field when crawling the web)

Week 6: Build database with: Author identifier, institution, geographic location, gender, fields, paper identifiers; and a second database with: paper identifier, paper title, authors, number of authors, year of publication, journal, field of study, length, authors of cited papers (author identifiers)

Week 7: Start analyzing data, creating visualizations, exploring the above questions to find robust statistical answers

Week 8: Start building website, create easy user interface to search database based on field, institution, geographical region

Week 9: Troubleshooting and testing