CS 12200: Group Project Proposal

Ruvim Ginzburg, Mark Saddler, Vishok Srikanth
January 24, 2017

Objectives

We propose to create software that can aid scientists in determining which federal government funding agencies are most likely to award grants in particular research areas. To do this, we propose to compile a searchable database of funded grant applications to major federal funding agencies such as the National Science Foundation (NSF), National Institutes of Health (NIH), and Centers for Disease Control and Prevention (CDC).

We envision the final product will allow a user to query our compiled database with keywords related to their research and be presented with summary statistics for each agency. These statistics will include the number of projects funded, the total amount of funding, and the average grant size for projects relating to these keywords. Furthermore we expect to be able to provide temporal and geographical information that may be of interest to the user. The proposed software will generate time series graphs to visualize trends in funding for projects that relate to the keywords, such as number of funded grants and total amounts awarded in previous fiscal years. Geographical information such as the states, cities, and institutions receiving the most funding for related projects will also be conveyed. A convenient yet detailed breakdown of awarded grants by funding source will provide valuable information to scientists looking to apply for federal funding for their own projects.

In addition to comparing funding from different sources for projects related to the user's keywords, our proposed software will also aid users by reporting commonly used language in the descriptions of successfully funded projects. This will be done by searching for frequently found phrases in the titles, abstracts, and descriptions of projects that match the query. By providing short phrases that occur frequently in relevant successful grants, we hope to give the user a sense of the kind of language that funding agencies have historically responded well to when awarding grants. This information may prove helpful for the writing of grant proposals.

Data

This project will rely on several government databases that record federal expenditures on scientific research. The NSF maintains a publically available record of all grants awarded since 1959 containing information such as project titles, abstracts, organizations, and amounts. All award information is available for download from the NSF website [1] in XML format. The NIH also maintain the Research Portfolio Online Reporting Tools Expenditures and Results

(RePORTER) database, which allows funded projects and abstracts to be downloaded in XML or CSV format. The NIH also provides outcomes information such as publications and patents linked to funded projects in similar downloadable formats. The Department of Health and Human Services also maintains the Tracking Accountability in Government Grants System (TAGGS), which is a database that reports all grant funding awarded by the operating divisions within the Department of Health and Human Services. This database provides basic information for grants awarded through the CDC among other agencies. The TAGGS database also allows for exporting to CSV files. Though we will initially focus our efforts on compiling a database of NSF, NIH, and CDC awards, if time permits, we will expand our search to include grants from other organizations such as the Department of Defense, Department of Energy, NASA, which can be found through usaspending.gov.

Outline

A basic outline of the required steps and the expected timeline for the proposed project are provided below:

1) Download NSF, NIH, and CDC award databases as XML and CSV files
    Expected: ~ 2 days

2) Extract and clean relevant information (e.g. titles, abstracts, organizations, award amounts, awarding agencies, timestamps) from downloaded files to assemble a single database
    Expected: ~ 1-2 weeks (and on-going if possible)

3) Develop code to efficiently query assembled database (using keywords to find related projects)
    Expected: ~ 1 week

4) Develop code to analyze results from queries and generate a visually appealing report (calculating summary statistics, time series graphs, geographical information, and recurring phrases)
    Expected: ~ 1-2 weeks

5) Make our code interface with a Django web application
    Expected: ~ 1 week

Note that some these steps can be worked on simultaneously.