DSCC 440: Data Mining

Final Project Proposal

Thomas Durkin and Sarah Siddiqui

Due Date: 10/28/2021

Problem statement

"How have the funding and associated research at the University of Rochester (and beyond) evolved during the past 20 years?"

For this project, we propose to analyze grants received by researchers at the University of Rochester and try to identify trends in the funding received as well any changes in the direction of research. Our primary focus will be the years between 2000-2020. Some variables included within the dataset are the grant's title, abstract (not all records), names of researchers along with affiliations, funding amount, funding institution, field of research, and the identifier (Dimensions publication id) for any associated publications. The available abstracts will be useful for identifying any emerging topics.

There are some instances of similar research [1], but these documents tend to either have a broad focus or are used for internal purposes or specific institutes [2]. In this case our focus is also narrowed to the University of Rochester, although we hope to analyze funding information for all the units including River Campus, Medical Center, Laboratory for Laser Energetics and Eastman School of Music. We want to do a comparison for different periods, over 5 year intervals, then 10 years, with emphasis on 2019-2020 since there could be a shift towards COVID related research that is also reflected in grant awards.

There is a possibility that the data generated (next section) is not sufficient for the data mining algorithms. In that event, we will expand our focus to funding received by Computer Science researchers in the United States, and see if that aligns with major areas listed in CSRankings [3], which is based on publications data from DBLP.

Data collection

The data for this project will be extracted from the Dimensions database (https://app.dimensions.ai/discover/grant) and analyzed using Python. Dimensions is well-known for its vast collection of grants, a lot of which are collected from the funders. We will start with University of Rochester data (over 3800 records), but also have access to the Dimensions API if additional customized records are needed.

Methodologies

We propose to utilize topic modeling and pattern mining for the following database fields: abstract, funders, and collaborators. Also, we will perform network analysis to identify collaborators within and outside of the UofR. Text mining techniques will help discover the differences based on discipline or subject category and affiliation data.

Algorithms:
The Latent Dirichlet Allocation (LDA) model will be useful for topic modeling with the grants' titles and assigned subject areas.We can also combine it with clustering and other preprocessing methods from Natural Language Processing [4].

Packages from Python:

- Scikit learn (clustering, preprocessing, etc.)
- Gensim (LDA)
- PyChart to display graphical comparisons

References

[1] 2021. Discovering Funding Sources with Dimensions at University of Colorado, Boulder. *Dimensions*. Retrieved October 18, 2021 from https://www.dimensions.ai/resources/discovering-funding-sources-with-dimensions-at-university-of-colorado-boulder/

[2] 2019. Dimensions - addressing analytical needs across campus at UC San Diego. *Dimensions*. Retrieved October 18, 2021 from https://www.dimensions.ai/resources/dimensions-addressing-analytical-needs-across-campus-at-uc-san-diego/

[3] Emery D. Berger. 2020. CSRankings. (2020). Retrieved October 27, 2021 from https://csrankings.org

[4] Sanjaya Subedi. 2021. Natural Language Processing with Python. *Sanjaya's Blog*. Retrieved October 27, 2021 from https://sanjayasubedi.com.np/series/nlp/