University of Notre Dame Du Lac, LG-72-18-0221-18

Investigating the National Need for Library Based Topic Modeling Discovery Systems

Summary: The University of Notre Dame is seeking an IMLS planning grant of $48,968 to: 1) determine the national need for an automated tool that supports cross-disciplinary research by applying semantic analysis across disparate disciplinary, scholarly works and 2) to solidify partnerships for the collaborative development of tools to support those needs. The main components of the project include a comprehensive literature review, an environmental scan, and a series of workshops to begin a national conversation and to form a national team to support further research. The planning project should be considered as the foundation for further development of an innovative approach that we believe will strengthen support of interdisciplinary scholarship by advancing library knowledge classification practices into the era of machine learning and artificial intelligence.

As part of our own evolution towards providing services in support of new forms of scholarship, the Hesburgh Libraries recently partnered with the Notre Dame Center for Civil and Human Rights (CCHR) to create a research tool called *Convocate*. Scholars at Notre Dame wanted to provide a tool to identify relationships between Catholic social teaching and international human rights law. The affordance to view the documents side-by-side in a digital interface enhanced the ability to compare and contrast concepts from disparate bodies of work. Displaying the documents side-by-side was not enough. With the aid of topic modeling (metadata creation and textual association), *Convocate* is able to return over 11,000 paragraphs tagged against a list of 250 topics and help users truly bridge the gap between the ways in which different disciplines describe equivalent concepts and thus discover more robust search results. For example, users can select the topic "solidarity/cooperation." Solidarity is a pervasive concept in Catholic social teaching that recognizes the responsibility to work for the common good of others because we are all persons. Cooperation is a similar concept in legal documents that addresses the need for nations to work together to solve international problems of economic, social, cultural or humanitarian nature. By choosing "solidarity/cooperation", users will be able to explore meaningful search results that reflect the common threads from the two fields about working together to improve human lives. The topic search enables a user from one discipline to overcome the problem of nuanced vocabulary in the other discipline and, hence, uncover relevant information that might otherwise remain hidden within the context of current classification schema.

While our initial work has demonstrated the value and need of cross-disciplinary approaches to discovery tools, we feel it can have greater impact to extend the reach of *Convocate* to other cross-disciplines, e.g. psychology and economics, chemistry and art, etc. In the research project to follow the initial phase supported by the planning grant, we will focus on how we can extend the *Convocate* prototype to explore scalability, extensibility and automation.[1]

Statement of Need: During the first phase of *Convocate's* development, we discovered one of the biggest limitations was the arduous and time-consuming task of human creation and association of metadata. While it always will be necessary to train topic modeling systems to develop classification and concept associations, the current level of human intervention cannot scale to millions of documents. It is clear that automation will be required. Gleaning from existing scholarly research on automated metadata creation and topic modeling, we know that libraries and organizations like OCLC have made progress in this area (Golub 2006). Caragea describes an automatic classification method applied to documents harvested from the Web (2014). Danilevsky outlines a framework for topical keyphrase generation and ranking, based on the output of a topic modeling of short documents (2014). Bijalwan's approach is similar to what we are considering, namely, "first categorize the documents using KNN based machine learning approach and then return the most relevant documents."(2014) From our perspective, none of the literature posits methods for comparing & contrasting seemingly disparate corpora.

The means in which academic research libraries create value for the academy has been rapidly evolving over the past 10 years. We are transitioning from organizations that support research by providing access to information and acquiring resources that are produced in the research process,to organizations that are more collaboratively

[1] Research questions include: 1. How the methodology used in Convocate can extend to a larger body of texts in the areas of international human rights and Catholic social teaching? 2. How to extend the *Convocate* tools and approaches to other cross-disciplinary works beyond international human rights and religion, e.g. sociology and art, comparative literature and physics, etc.? 3. What is needed to scale and automate (as much as possible) the current time-consuming task of human creation of topic/content associations and metadata?

immersed in the creation of scholarship itself. We also recognize an increased need to provide more sophisticated and innovative tools to support university trends in cross disciplinary research. To meaningfully expose relationships across disciplinary domains, library discovery systems and metadata creation must leverage emerging technologies for greater sophistication. Both the need to support interdisciplinary scholarship as well as the gap in utilizing topic modeling methods to bridge the related knowledge between two different corpora present a unique opportunity to augment the library classification toolbox with emerging artificial intelligence technologies. A planning grant would enable us to seek input from and establish a network with other academic institutions to develop a clear list of requirements . The outreach to librarians, scholars, and other researchers will ensure that the next phase is accurately informed about their needs, especially within the Digital Humanities. By strategically aligning the workshops, we intend to ensure that smaller institutions have a voice in determining what is needed to support scholarly research in their unique situations.

National Impact: In this brave new world, information professionals in all corners of the academic research library are impacted by the trends in shifting priorities to meet emerging scholarly needs while finding new ways to support teaching, learning, and research in the face of automation of traditional library tasks. We believe the work on *Convocate* can have a significant impact both for scholars and libraries. For researchers, we believe that scaling the capabilities of *Convocate* and broadening its scope of cross-disciplinary content to science, social sciences, and humanities, could augment library science theory, knowledge, and practices with computational methodologies for supporting interdisciplinary discovery and scholarship. For libraries, there is a significant opportunity to leverage traditional library skills such as classification, metadata creation, and disciplinary stewardship in ways that create new value and can augment the developing skill profile of library employees such as catalogers, subject liaisons, and digitization specialists to help transition into the knowledge work of the future. We believe, additionally, that this work provides an opportunity to revitalize the perception of libraries as collaborators in scholarly research.

Project Personnel: Zheng (John) Wang - Associate University Librarian, Hesburgh Libraries, University of Notre Dame (PI). John's scholarly work is grounded in the areas of web services and analytics, digital library development, project management, and digital scholarship. Christina Leblang - *Convocate* Project Manager, Center for Civil and Human Rights, University of Notre Dame, M.T.S., Moral Theology, University of Notre Dame, M. Eng., Biomedical Engineering, Cornell University, B.S., Biological and Environmental Engineering, Cornell University.

Workplan: *Environmental Scan and Literature Review:* Environmental scans and literature reviews will gather additional background information, identify potential open-source programming resources, and assess comparable applications and tools to inform development decisions.
Workshops: Conduct at least 3 workshops over the year, with locations and timing selected based on regional variety (East Coast, West Coast, and Mid regions) so that smaller institution can participate, as well as an associated event - such as ALA, CNI, DLF, and digital humanities conferences. The workshops will serve to: 1) engage a national and multi-disciplinary discussion about cross-disciplinary research and research tools using the *Convocate* prototype to help frame the discussion, 2) build a diverse network of collaborators and stakeholders, 3) assess the value of such tools for libraries and colleges/universities, and 4) outline best methods for project dissemination, sustainability, and support for higher ed institutions of all sizes and type.
Networking/Advisory Board: Key project personnel will continue to network with collaborators throughout the project. The advisory board for the next phase of work will be finalized.

Budget: We are seeking $48,968. $5,033 for salary and fringe benefits for 10% of a project manager's time. $19,685 for workshop participant support and $6,250 directed to other workshop expenses, including meals. $10,000 will be available for travel and expenses for ND personnel to host workshops and network with collaborators. $1,000 for professional consultation on survey questions that will be utilized in workshops. $6,760 for indirect costs.