Making Sense of Large-scale Collections of Unstructured, Text-Based, Interview Responses

Research Experience for Master’s Students

Gabriel Harp, Research Director
ArtsEngine/a2ru, University of Michigan
gharp@umich.edu

Analysis of textual data is a growing modality for researchers. For example, one estimate anticipates 246 billion emails will be received by the end of 2019. For scientific research, it is becoming much easier to assemble large-scale datasets that consist of unstructured texts from web and forms, open-ended survey responses, social media, and from captured audio or video. Large-scale datasets also come with additional challenges for managing those data: from the interoperability of data formats to end-to-end integration of workflows to the disciplinary practices and the tools needed to accomplish dedicated tasks.

This research project (research.a2ru.org) aims to make sense of ~600 long-form, semi-structured interviews about the role of the arts in research universities, with the end goal of learning about the principles and practices that can lead to a more dynamic research university system. To support that goal we have built a knowledge engine (the SPARC platform; currently in a closed initial release) for indexing, searching, browsing, and visualizing unstructured text-based interview responses and documents. Researchers from different disciplines and universities already use the knowledge engine to learn from the data, but there are some fantastic options for improving the usability and accessibility of such large quantities of unstructured text. Some of the immediate possibilities include:

- Information design and visualization of analyzed text responses (Java, javascript, D3),
- Designing analytics and/or machine learning approaches for unstructured text responses such as topic analysis, summarization, or entity extraction,
- Developing and documenting recipes for end-to-end text processing, integration among toolsets, and data transformations to facilitate and enable the ingestion and acquisition of new data.

The research outcomes of student projects in these areas will contribute directly to the project, enhancing our knowledge about how to handle large sets of unstructured text for learning about what people think, feel, observe, and act — using their own words as the primary basis for interpretation by researchers and policy makers. The outcomes will also advance our understanding about how to communicate complicated information and data science techniques to mostly novice users, either as explicit recipes for data curation and/or by “hiding” technical choices within an application interface.
**Student role:**
The student will serve as a research assistant for the project, will be primarily self-directed, and will be involved in research meetings with our research team (other undergraduate, post-doc, and professionals) working on the project. In this way s/he will peripherally participate in a research team and be exposed to all aspects of the research. The student will be specifically responsible for her/his own role, project scope, outcomes, and coordinating with others when needed and necessary.

The choice of specific research projects and outcomes will depend on the student's specific skills (or desire to learn such skills), their entrepreneurship and problem-solving, and domain area or technical interests.

Specific project outcomes would be anticipated to include:
- interviews or focus groups with leading users of the prototype knowledge engine to learn more about existing use-cases, unfulfilled needs, and prospective use-cases
- experimentation and development of visualizations, analyses, or recipes
- application of visualizations, analyses, or recipes to a demonstration question
- accessing and utilizing the knowledge engine's API as a data source
- documentation (visualization how-to, workflow, recipes and processes)
- contributing issues and bug reports to usability discussions and development tasks

**Mentoring plan:**
I will meet with the student several times weekly throughout the project. Additionally, I and my post-doctoral fellow will monitor and provide constructive and iterative feedback to the student closely during several key phases of the project, including data collection, analysis of needs, workflow or visualization design, documentation, etc.

Opportunities for presenting work in progress to various data science communities on campus will be possible (e.g. the Michigan Data Science Initiative or the Computational Social Sciences workshop), as well as the opportunity for consultations with the leading users of the knowledge engine who are using the data around the country.

Additional professional opportunities can include consultations with the Research Data Services group at the UM Library, the Digital Education and Academic Innovation group, data scientists and researchers at the Inter-University Consortium of Political and Social Research, the Statistics, Consulting, and Analytics consulting group (CSCAR), and other faculty and researchers from UM and a2ru’s national network.