Understanding Scholarly Information Seeking Behavior

**Background**: Background material is essential for learning about new things. Think of how much time people spend on Wikipedia just clicking through links to learn about related topics. Scholars too follow these information threads by searching for keywords or following the citation trails in papers. In general, this behavior is known as **information seeking** and has been studied from multiple vantage points in Information Science, Anthropology, and Psychology to name a few. Knowing which kinds of articles people like to read next can have significant benefits for recommender systems (people who read X read Y next), for understanding trends in what content is most important, and for modeling how to best get people up to speed on a topic.

**Project goals**: This project aims to data mine a large collection of real-world online accesses of a major scholarly publication website in order to identify trends in how people browse its contents. Using millions of traces that show which articles are accessed among the thousands possible, we'll aim to quantify how each trace is meeting a particular information need and discover what trends emerge. As a part of this project, we'll use data mining and machine learning techniques on both the behavior traces and the content of the articles themselves to understand what is being read. A stretch goal is to build a recommendation system that suggests what articles a reader might consider reading *before* or *after* the currently selected article.

**Student Role**: A student will participate in a new research project focused on data mining behavior traces in a large scholarly collection and be the primary research developer for the software that drives the project. The student should have a strong programming background in python, familiarity working in a UNIX environment, and some experience with machine learning is a plus. As a part of the project, the student will help write up results in both a scientific publication format and for a general audience as a blog. Specifically, a student will:
- Develop new data mining models to predict and measure article access
- Use machine learning and natural language processing techniques to analyze article content and relate it to access patterns.
- Assist in performing a research survey and writing up results
- As time allows, develop a recommendation system

**Mentorship Role**: I ([David Jurgens](#)) will meet one on one with the student weekly and then we’ll also meet weekly with the whole [Blablablab](#) group for a general lab meeting. Individual meetings will cover both technical and professional skills needed to be a successful researcher in industry or academia. As a part of the summer project, I’ll work with the student on how to design a research study, develop reproducible research, write results for both a scholarly and general audience, and give presentations on their work. I will help guide the student to obtain concrete outcomes they can show off to others to help them get to the next stage of their career.