Collecting and Data Mining Online Text Data

Daniel Kerchner
Justin Littman
Laura Wrubel
GW Libraries

David A. Broniatowski, Ph.D.,
Asst. Prof., GW Department of Engineering Management and Systems Engineering,

Despite the increasing popularity of social media as a topic of study, most current open source tools for social media collecting do not adequately support today's researchers in building robust collections for analysis and archiving. GW Libraries has developed Social Feed Manager (http://go.gwu.edu/sfm), an open source application that allows researchers to define and build collections from the public APIs of Twitter, Tumblr, Flickr, and Sina Weibo. Social Feed Manager provides a user interface for multiple researchers at an institution to build collections with their own social media credentials; tracks changes to the collecting over time, in support of data provenance; and provides export options for JSON, CSV, Excel, and other formats.

In this tutorial, the Social Feed Manager project team will demo the software, walking through its capabilities for building collections and exporting data in multiple formats. A sandbox instance of SFM will be available for attendees to build sample collections and gain an understanding of how to collect with the tool.

Next, Dr. Broniatowski will discuss several text mining techniques that may be applied to text data, including the social media data collected from SFM. Using the Python libraries scikit-learn and NLTK (Natural Language Toolkit), Dr. Broniatowski will provide several examples of how text data may be collected from several other sources including websites, RSS feeds, PDFs, and other venues. Examples of insights from both supervised and unsupervised learning algorithms will be discussed. Case studies include the use of Twitter data to conduct influenza surveillance and to understand rationales for vaccine refusal and the use of government transcript data to understand social dynamics on FDA advisory panels. Finally, Dr. Broniatowski will discuss challenges with current research tools, and opportunities to use social media to extend these tools in a synergistic manner.