Finding the Canary for Text Mining: Analysis of the use and users of MONK text mining research software

Harriett Green, English and Digital Humanities Librarian
University Library, University of Illinois at Urbana-Champaign
green19@illinois.edu
2010 Chicago Colloquium on Digital Humanities and Computer Science Poster Presentation

MONK is a textual analysis research tool hosted by the University of Illinois Library that enables humanities scholars to mine data from digitized texts in select literary databases and archives. this poster presents initial analysis on the use of MONK by researchers using web usage data tracked on MONK during the first 8 months of 2010.

BACKGROUND

MONK (http://www.monkproject.org/) is an advanced text-mining software program built with the support of a $1 million grant from the Andrew W. Mellon Foundation from 2007 to 2009. MONK was built by researchers at the University of Illinois at Urbana-Champaign, University of Alberta, University of Maryland, McMaster University, the National Center for Supercomputing Applications, University of Nebraska-Lincoln, and Northwestern University.

MONK builds upon two previously developed text mining programs NORA (http://www.noraproject.org/) and WordHoard (http://wordhoard.northwestern.edu/), and builds upon them to create a powerful new environment that "lets users carry out complex data-mining and query operations across collections that contain nearly 200 million words" (MONK documentation, http://www.monkproject.org/background.html). The SEASR (http://seasr.org/) environment provides the tools for statistical analyses in MONK.

MONK transitioned from a research project to University of Illinois Library-supported research resource in late December 2009 and January 2010. All scholars can access MONK to conduct textual analysis on publicly available digital collections such as Indiana University's "Wright American Fiction 1850-1875" collection, University of North Carolina at Chapel Hill's "Documenting the American South" collection, and the University of Virginia's "Early American Fiction" collection. Scholars affiliated with institutions in the Committee for Institutional Cooperation (CIC) consortium can additionally access texts from proprietary databases such as *Eighteenth Century Collections Online*, *Early English Books Online*, and Proquest's Chadwyck-Healey *Nineteenth-Century Fiction*. Researchers can also import texts into MONK with the use of Zotero and a MONK Firefox extension.

DATA AND METHODOLOGY

Statistics about the web traffic and usage of MONK were gathered using AWStats, a web log analyzer used to track the web statistics for MONK. The statistics used in this study were gathered January 2010 through August 2010, the first eight months of MONK's release as a public instance.

The statistics include the number of visits on each webpage within MONK, the amount of data processed through each page, and length of the visits. Other available statistics that are currently being analyzed include the number of entry and exit visits, users' geographic locations, and times of day when MONK was accessed. For the next stage of the research, qualitative data will also be gathered from users of MONK.

The preliminary statistical analyses conducted on the data included calculating the mean of users that accessed each section of MONK; the mean the amount of data flowing through each section MONK; standard deviation of the number of users; and the distribution of the MONK sections as divided into the three types of web pages: Orientation, Workset, and Toolset pages.

ANALYSIS

Early analysis has revealed that the text mining tools in MONK are being accessed and utilized at varying frequencies by users. The most frequent tools used on average were:

https://monk.library.illinois.edu/secure/get/CorpusManager.getWorkList = compiling the worksets
https://monk.library.illinois.edu/secure/get/ProjectManager.getToolSets = selecting a toolset
https://monk.library.illinois.edu/cic/public = the opening page

One hypothesis that might be drawn from this initial data is that many users are in the initial exploratory steps of using MONK by creating accounts and putting together their first worksets to analyze. Another point of note is a comparison of the accessed MONK pages, which reveals that the use of analytics toolsets and the use of tools for creating worksets were accessed by researchers at a proportion of 2 to 1. In another analysis, the largest amount of data was utilized for a tool comparing the frequency of word features, with a usage of 272.1 MB on average and 15% of the total data processed. These are only a sample of the analyses conducted so far.

CONCLUSION THUS FAR

This initial examination has begun to reveal several early insights on how scholars are conducting textual analysis research in MONK. Projected in-depth analyses will show how users employing MONK for their textual analysis research. I anticipate that further analysis of the usage data of MONK will critically reveal new avenues of examining the workflows of humanist scholars, and how they integrate digital tools with traditional modes of scholarship.

REFERENCES

Burrows, John and Hugh Craig. 2001. Lucy Hutchinson and the Authorship of Two Seventeenth-Century Poems: A Computational Approach. *Seventeenth Century*. 16: 259-283.

Sinclair, Stéfan. 2003. Computer-assisted reading: Reconceiving textual analysis. *Literary and Linguistic Computing*. 18: 175-184.

Sperberg-McQueen, C. M. 1991. Text in the electronic age: textual study and text encoding, with examples from medieval texts. *Literary and Linguistic Computing*. 6: 263-279.

Warwick, Claire. 2004. Print Scholarship and Digital Resources. In *A Companion to Digital Humanities*, ed. Susan Schreibman, Ray Siemens, John Unsworth. Oxford: Blackwell, 2004.

Zillig, Brian L. Pytilk. 2009. TEI Analytics: converting documents into TEI format for cross-collection text analysis. *Literary and Linguistic Computing*. 24: 187-192.

MONK documentation, http://monkpublic.library.illinois.edu/monkmiddleware/public/index.html, accessed October 31, 2010.