Finding the Canary in Text Mining
Analysis of the use and users of MONK text mining research software
Harriett Green, English and Digital Humanities Librarian  ❖  green19@illinois.edu  ❖  University of Illinois at Urbana-Champaign

BACKGROUND
MONK (http://www.monkproject.org/) is an advanced text-mining software program created with the support of a $1 million grant from the Andrew W. Mellon Foundation from 2007 to 2009. Researchers from seven institutions, including UIUC, the NCSA, and Northwestern University, collaborated to build MONK, and built it upon two previously developed text mining programs NORA and WordHoard. The SEASR environment provides the tools for statistical analyses in MONK. MONK became a University of Illinois Library-supported research software in late December 2009 and January 2010. All researchers now utilize MONK to conduct textual analysis on public digital text archives such as Indiana University’s “Wright American Fiction 1850-1875” collection and UNC-Chapel Hill’s “Documenting the American South” collection. Faculty and students affiliated with the Committee for Institutional Cooperation (CIC) consortium also can access texts from select proprietary databases such as Eighteenth Century Collections Online and Chadwyck-Healey’s Nineteenth-Century Fiction. Researchers can also import texts into MONK with the use of Zotero and a Firefox extension.

DATA AND METHODOLOGY
Usage statistics for the MONK website were gathered from January 2010 through August 2010, the first eight months of MONK’s initial release as a public instance. The statistical analyses conducted included calculating the mean numbers of users that accessed each section of MONK; mean amounts of data processed; standard deviation of the number of users; the distribution of the web pages as divided into the categories of Orientation, Toolsets, and Worksets; and various distributions of users.

ANALYSIS
Early analysis has revealed that the text mining tools in MONK are being accessed and utilized at varying frequencies by users. Two graphs shown here reflect early analysis of the usage data.

The bar graph provides a comprehensive overview of the most frequently accessed and least frequently accessed pages by users of MONK. This graph also enables a comparison of the types of pages accessed.

The pie graph displays the average amount of data processed through each MONK page section, and this graph also reveals the broad distribution of the pages accessed by users in MONK. These graphs are but two analyses conducted on early usage statistics for MONK.

FINDINGS
This initial examination reveals several early insights on how researchers conduct data mining of text in MONK. Notable insights include how the frequency of use decreased as the data mining tools increased in degree of complexity, but a higher number of Toolset pages were accessed overall.

The investigator anticipates that further in-depth 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

https://www.monk.library.illinois.edu