Topic Modeling for Digital Collections Appraisal

Thomas G. Padilla
University of Illinois at Urbana Champaign
Graduate School of Library and Information Science
tpadill2@illinois.edu | @thomasgpadilla

This project aims to improve the appraisal of large collections of heterogeneous digital files with respect to time and efficiency by applying latent dirichlet allocation (LDA) to a corpus of an unprocessed personal digital archive. Having secured permission from the University of Illinois at Urbana Champaign Archives, an image was made of one of the hard drives of Dr. Carl Woese, a preeminent scholar at the university. Dr. Woese famously discovered the third domain of life, Archaea.

All text documents were extracted from the hard drive, converted to .txt format, file names were normalized, and the corpus was subjected to topic modeling. The results indicate that topic modeling could be a viable option to approach appraising and providing access to unprocessed digital collections thus saving time, money, and shortening the distance between acquisition of and public access to digital collections.

Tools:

Forensic Toolkit Imager
http://www.forensicswiki.org/wiki/FTK_Imager

JDiskReport: http://www.jgoodies.com/freeware/jdiskreport/

MyMorph
http://docmorph.nlm.nih.gov/docmorph/mymorph.htm

NameChanger
http://www.mrrsoftware.com/MRRSoftware/NameChanger.html

Topic Modeling Tool
https://code.google.com/p/topic-modeling-tool/

Images:

Big Tree of Life
Norm Pace, The Pace Lab, http://pacelab.colorado.edu/

Document
Edward Boatman, http://thenounproject.com/edward/

Network
Semilla Solar, http://thenounproject.com/semilla.solar.10/

Bibliography:

Blei, David. "Topic Modeling and Digital Humanities." *Journal of Digital Humanities*. no. 1 (2012). http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/(accessed April 20, 2013).

Blei, David. "Probalistic Topic Models." Last modified April 2012. Accessed April 20, 2013. http://www.cs.princeton.edu/~blei/papers/Blei2012.pdf

Goldstone, Andrew, and Ted Underwood. "What Can Topic Models of PMLA Teach Us About the History of Literary Scholarship?" Journal of Digital Humanities. no. 1 (2012). http://journalofdigitalhumanities.org/2-1/what-can-topic-models-of-pmla-teach-us-by-ted-underwood-and-andrew-goldstone/ (accessed April 20, 2013).

Mimno , David. "The Details: Training and Validating Big Models on Big Data." Journal of Digital Humanities. no. 1 (2012). http://journalofdigitalhumanities.org/2-1/the-details-by-david-mimno/ (accessed April 20, 2013).

Posner, Miriam and Andy Wallace. "Very basic strategies for interpreting results from the Topic Modeling Tool." *Miriam Posner* (blog), 10 29, 2012. http://miriamposner.com/blog/?p=1335(accessed April 20, 2013).

Rhody, Lisa. "Topic Modeling and Figurative Language." *Journal of Digital Humanities*. no. 1 (2012). http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/(accessed April 20, 2013).

Schmidt, Benjamin. "When you have a MALLET, everything looks like a nail." *Sapping Attention*(blog), 11 2, 2012. http://sappingattention.blogspot.com/2012/11/when-you-have-mallet-everything-looks.html (accessed April 22, 2013).

Underwood, Ted. "What kinds of “topics” does topic modeling actually produce?." *The Stone and the Shell* (blog), 04 1, 2012. http://tedunderwood.com/2012/04/01/what-kinds-of-topics-does-topic-modeling-actually-produce/(accessed April 24, 2013).