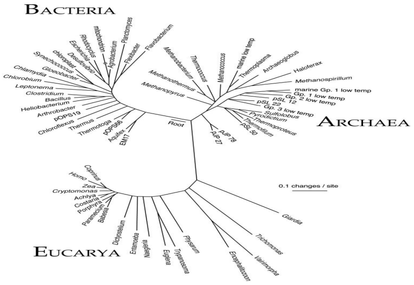
Topic Modeling for Digital Collections Appraisal

The Carl Woese Collection

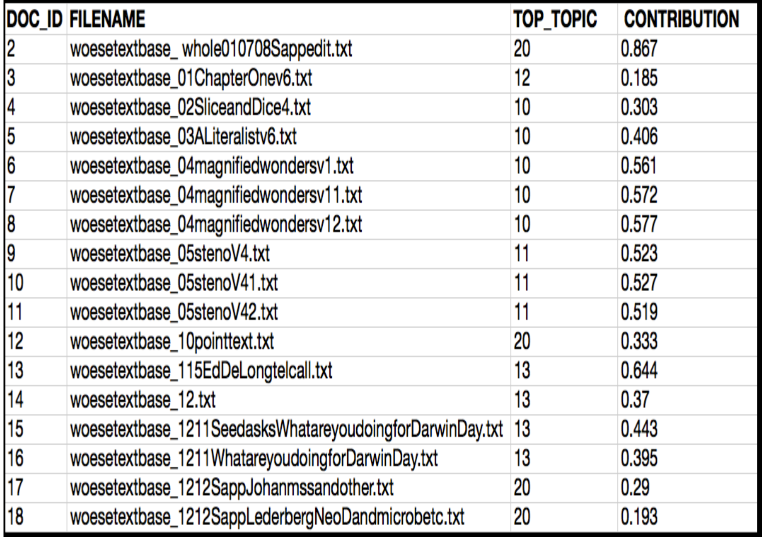
Thomas G. Padilla
Graduate School of Library and Information Science
University of Illinois at Urbana Champaign



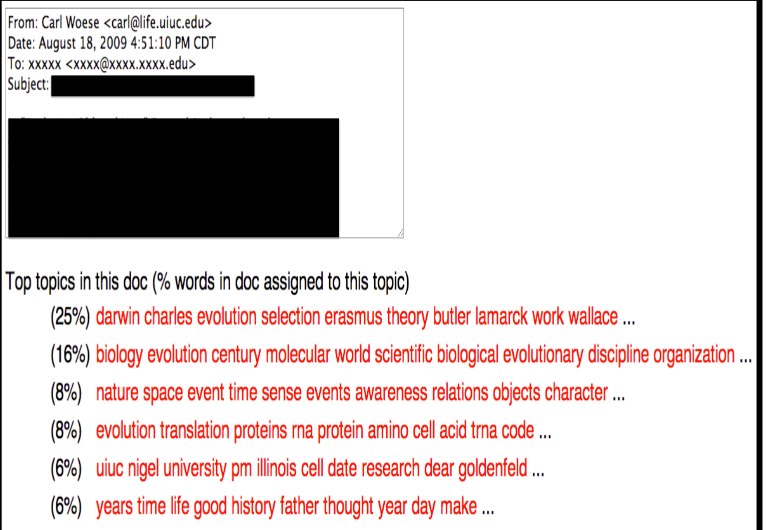
ABSTRACT

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 associated with the university. Dr. Woese famously discovered the third domain of life, Archaea.

All textual 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.



DOC_ID	FILENAME	TOP_TOPIC	CONTRIBUTION
2	woesetextbase_ whole010708Sappedit.txt	20	0.867
3	woesetextbase_01ChapterOnev6.txt	12	0.185
4	woesetextbase_02SliceandDice4.txt	10	0.303
5	woesetextbase_03ALiteralistv6.txt	10	0.406
6	woesetextbase_04magnifiedwondersv1.txt	10	0.561
7	woesetextbase_04magnifiedwondersv11.txt	10	0.572
8	woesetextbase_04magnifiedwondersv12.txt	10	0.577
9	woesetextbase_05stenoV4.txt	11	0.523
10	woesetextbase_05stenoV41.txt	11	0.527
11	woesetextbase_05stenoV42.txt	11	0.519
12	woesetextbase_10pointtext.txt	20	0.333
13	woesetextbase_115EdDeLongtelcall.txt	13	0.644
14	woesetextbase_12.txt	13	0.37
15	woesetextbase_1211SeedasksWhatareyoudoingforDarwinDay.txt	13	0.443
16	woesetextbase_1211WhatareyoudoingforDarwinDay.txt	13	0.395
17	woesetextbase_1212SappJohanmssandother.txt	20	0.29
18	woesetextbase_1212SappLederbergNeoDandmicrobetc.txt	20	0.193



From: Carl Woese <carl@life.uiuc.edu>
Date: August 18, 2009 4:51:10 PM CDT
To: xxxxx <xxxx@xxxx.xxxx.edu>
Subject: [redacted]

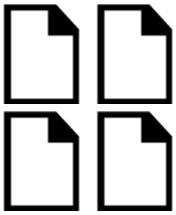
[redacted]

Top topics in this doc (% words in doc assigned to this topic)

- (25%) darwin charles evolution selection erasmus theory butler lamarck work wallace ...
- (16%) biology evolution century molecular world scientific biological evolutionary discipline organization ...
- (8%) nature space event time sense events awareness relations objects character ...
- (8%) evolution translation proteins rna protein amino cell acid trna code ...
- (6%) uiuc nigel university pm illinois cell date research dear goldenfeld ...
- (6%) years time life good history father thought year day make ...

42 GB

907 .TXT



20 TOPICS

EXTRACT	VISUALIZE	CONVERT	CLEAN	MODEL	REFINE
FTK IMAGER	JDISKREPORT	MYMORPH	NAMECHANGER	TMT	TMT

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/

QUESTIONS

tpadill2@illinois.edu
@thomasgpadilla