Topic Modeling: Mallet Workshop

Fall 2016
INF383H Intro to DH

Agenda

- General overview of topic modeling
- Introduction to MALLET
- Review MALLET projects in action
- Introduce and contextualize demo data
- Demo MALLET
- Independent activity
- Group discussion

Workshop goals & learning objectives

- Become familiar with the basic concept of topic modeling, understanding its potential and limitations
- Explore how MALLET is used to generate topics for corpora of data
- Give you confidence to employ MALLET in your own research
- Engage in hands-on topic modeling exercise, working on the command line, and create basic visualizations using common applications
- Develop appreciation for analytical work of interpreting results from topics
- Start a conversation about next steps beyond topic modeling

Topic modeling

What is topic modeling?

Miriam Posner has described topic modeling as "a method for finding and tracing clusters of words (called 'topics' in shorthand) in large bodies of texts."

What, then, is a topic? One definition offered on Twitter during a conference on topic modeling described a topic as "a recurring pattern of co-occurring words."

-- Megan Brett, "Topic Modeling: A Basic Introduction"

Topic modeling is a way to use distance reading and a computer tool to data mine a large set of documents to find meaning throughout the entire data set. Setting the researcher up with a better idea of where to start on your research questions into that data or document set.

What is it for?

Topic modeling provides the researcher with a way to build links that s/he might not otherwise see with a close reading of the texts

The topic modeling tool identifies relationships between words for which the researcher then identifies overall themes

This is a shortcut that saves the researcher a great deal of time in becoming familiar with a corpus. However, the researcher still need some familiarity with the corpus to identify themes

How does it work?

Topic modeling tools use statistical methods to analyze the words of the original texts to discover themes

Ted Underwood provides this equation:

$$P(Z|W,D) = \frac{\#\ of\ \ words\ \ W\ \ in\ \ topic\ \ Z + \beta_w}{total\ \ tokens\ \ in\ \ Z + \beta} * (\#\ \ words\ \ in\ \ D\ \ that\ \ belong\ \ to\ \ Z + \alpha)$$

First, the program runs through the collection word-by-word and randomly assigns each word to a “basket,” then iteratively resamples topic assignments for each word given all other words and their current topic assignments

Model improves and becomes more consistent over time

What do you need to topic model?

Megan Brett defines four essentials:

1. A large **corpus**
2. **Familiarity** with the corpus
3. A **tool** to do the topic modeling
4. A way to **understand** your results

What are the limitations of topic modeling?

Requires the researcher to make judgement calls so the end results is based more on preferences than transparent characteristics

The results produced by topic modeling tools can be difficult to understand

The researcher still needs to be familiar with the corpus in order to use the tool. It isn’t a miracle cure that lets you find results without work

There is an element of randomness to the results. Even if the researcher does a second run on that same data set, it is unlikely that s/he will get exactly the same result again

MAchine Learning for LanguagE Toolkit

History of MALLET

According to the tool’s website, “MALLET is a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text.”

It was developed in 2002 by computer science professor Andrew McCallum at UMass Amherst, has received 8 updates since its original release, and is available free for use under the common public license.

Natural language processing

MALLET uses latent Dirichlet allocation (LDA) modeling

LDA is a generative statistical model that allows sets of observations to be explained by unobserved groups that determine why some parts of the data are similar.

MALLET software determines that the words collected in each document are a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics. LDA was first presented as a model for topic discovery by David Blei, Andrew Ng, and Michael I. Jordan in 2003.

MALLET's data pre-processing components

MALLET requires command line experience to use. To direct the computer to work with data, one enters specific **commands** and **operators.**

A **pipe** operator (|) takes the output of one command and directs it into the input of another command. Pipes are most often used for feature extraction.

A pipe operates on an **instance**, which is a carrier of data.

An instance contains four generic fields of a predefined name: "**data**," "**target**," "**name**," and "**source.**" "Data" holds the data represented by the instance, "target" is often a label associated with the instance, "name" is a short identifying name for the instance (such as a filename), and "source" is human-readable source information (such as the original text).

Using the command line

The command line interface is a tool where you can type text commands to perform specific tasks

The Mac command line is a program called Terminal, which lives in the **/Applications/Utilities/** folder

Basic commands:

- **ls** displays a list of files and folders in the current directory
- **cd** instructs the computer to change directory

Examples of commands in MALLET

The command to import data in MALLET:

```
bin/mallet import-dir --input sample-data/web/* --output web.mallet
```

The command to change your corpus of data to a MALLET-friendly file to run through the algorithm:

```
./bin/mallet import-dir --input sample-data_one --output tutorial.mallet
--keep-sequence --remove-stopwords
```

Here you see an instance (the file name sample-data)

Challenges

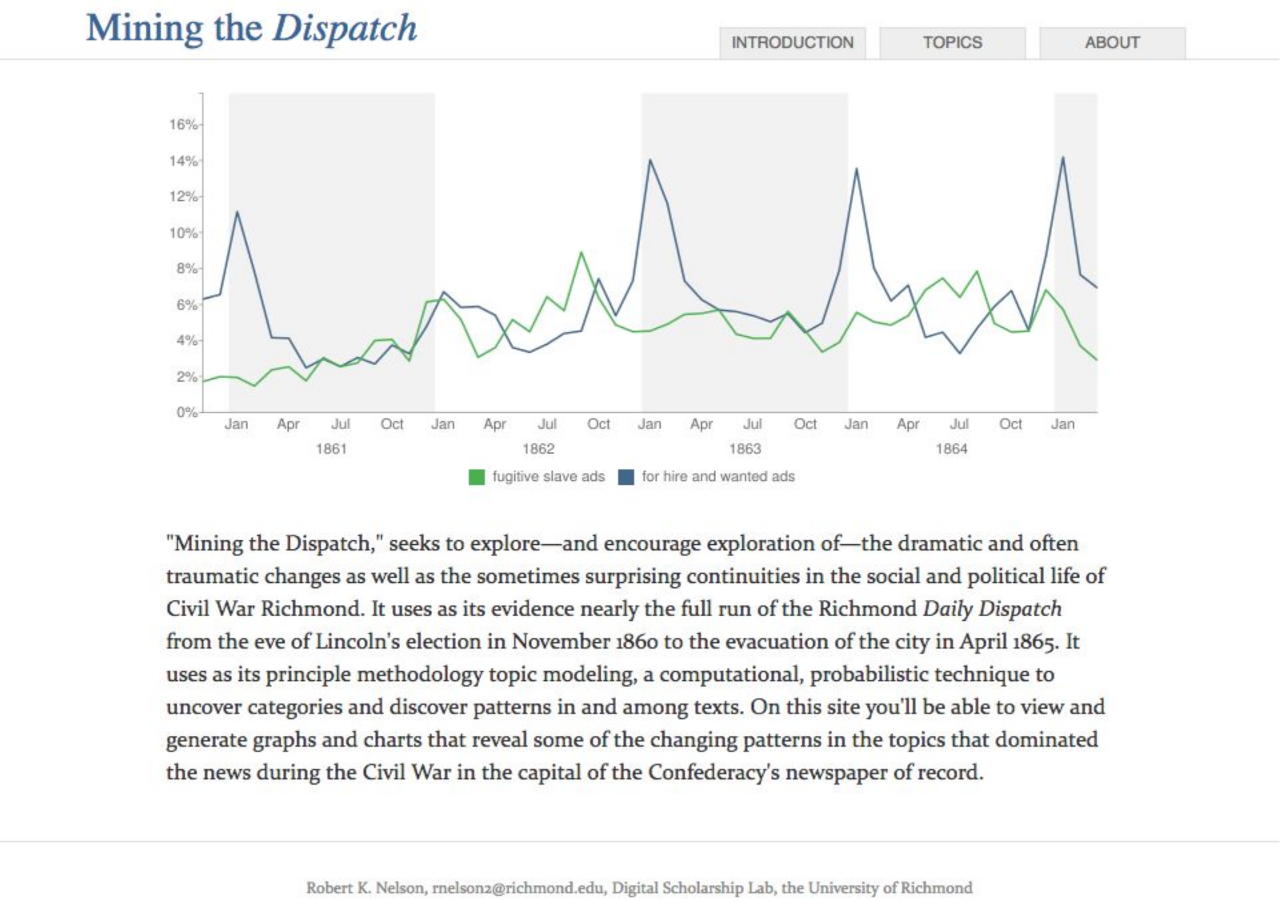
Lack of documentation

Steep learning curve and hidden steps for novice in the command line

Command line commands differ based on operating system

Topic modeling is a method of distance reading, but in order to use MALLET, the researcher still need at least a reasonable familiarity with the files running through the tool

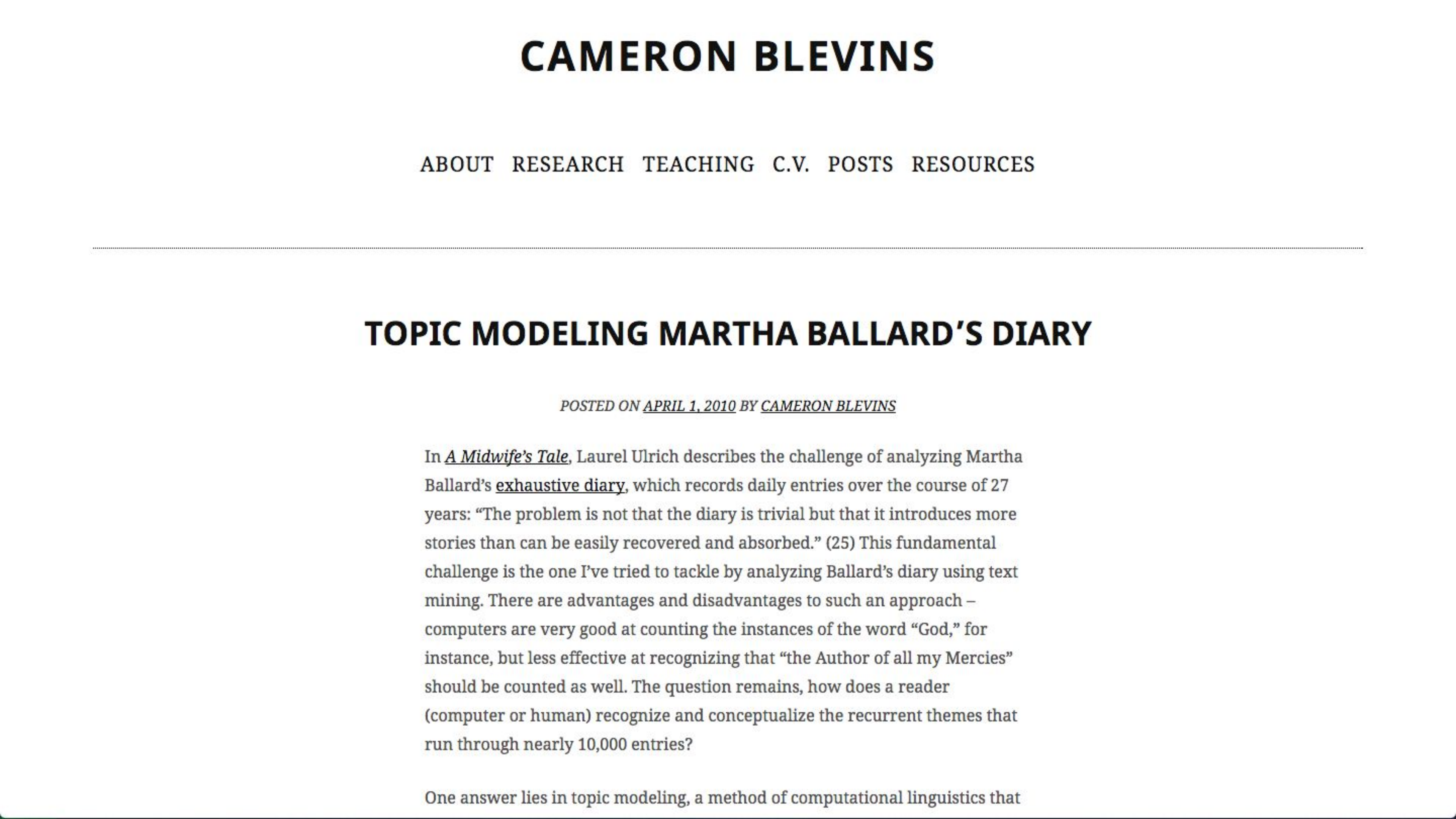
MALLET projects



Mining the *Dispatch*

"Mining the Dispatch," seeks to explore—and encourage exploration of—the dramatic and often traumatic changes as well as the sometimes surprising continuities in the social and political life of Civil War Richmond. It uses as its evidence nearly the full run of the Richmond *Daily Dispatch* from the eve of Lincoln's election in November 1860 to the evacuation of the city in April 1865. It uses as its principle methodology topic modeling, a computational, probabilistic technique to uncover categories and discover patterns in and among texts. On this site you'll be able to view and generate graphs and charts that reveal some of the changing patterns in the topics that dominated the news during the Civil War in the capital of the Confederacy's newspaper of record.

Robert K. Nelson, rnelson2@richmond.edu, Digital Scholarship Lab, the University of Richmond



CAMERON BLEVINS

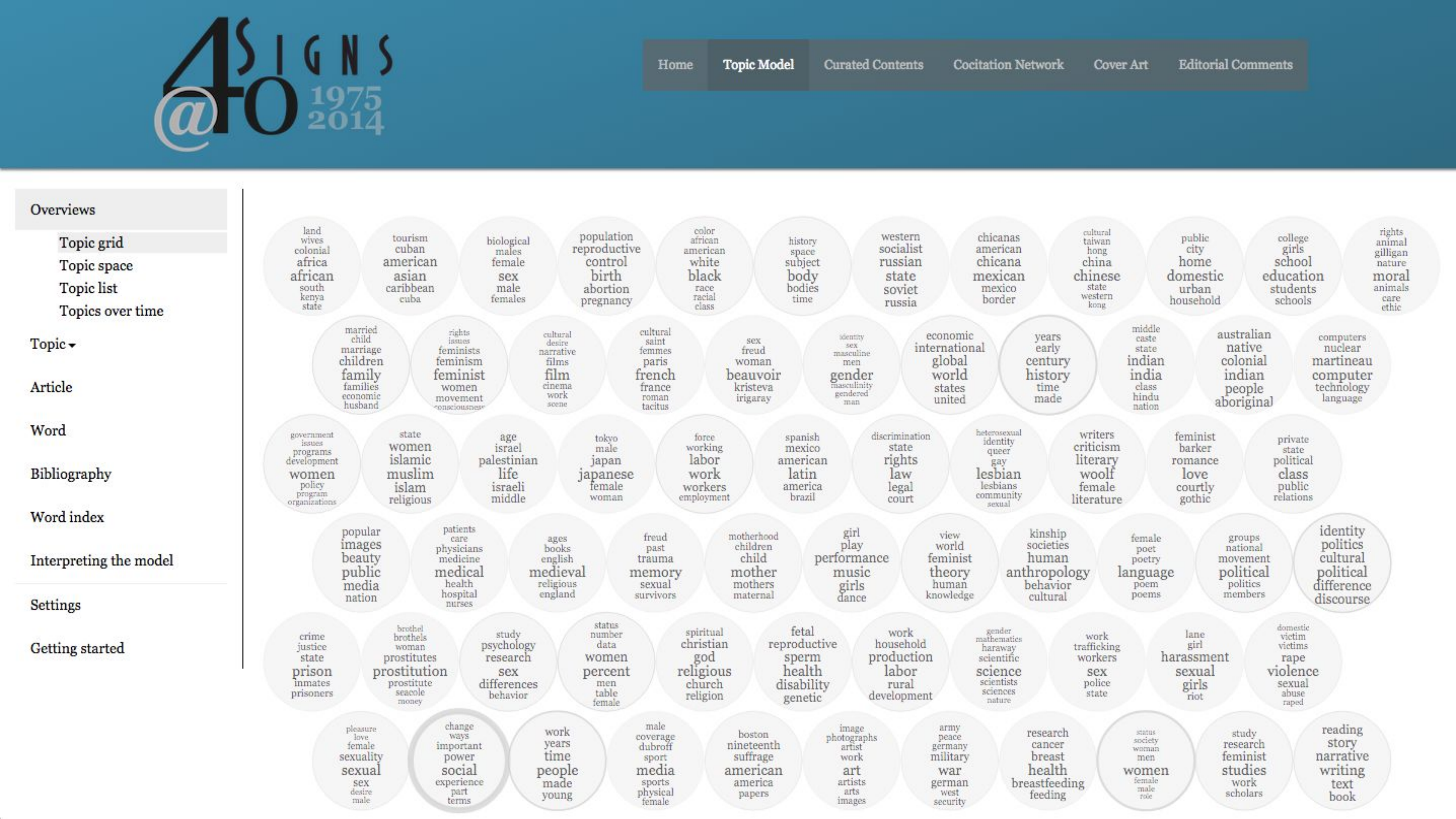
ABOUT RESEARCH TEACHING C.V. POSTS RESOURCES

TOPIC MODELING MARTHA BALLARD’S DIARY

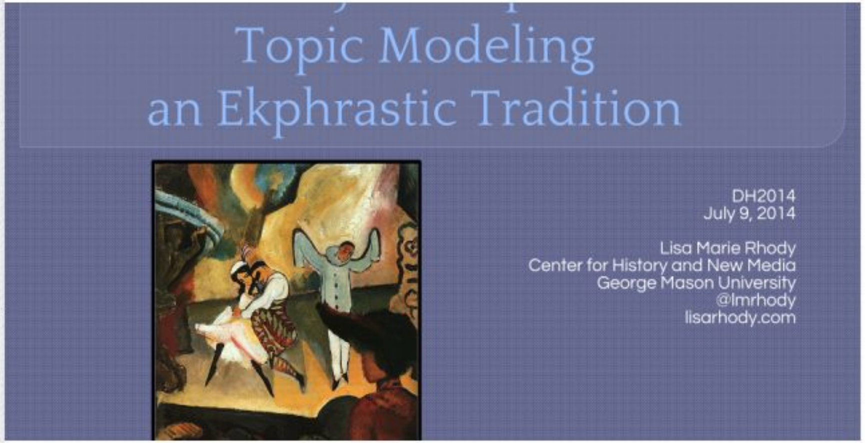
POSTED ON APRIL 1, 2010 BY CAMERON BLEVINS

In *A Midwife’s Tale*, Laurel Ulrich describes the challenge of analyzing Martha Ballard’s exhaustive diary, which records daily entries over the course of 27 years: “The problem is not that the diary is trivial but that it introduces more stories than can be easily recovered and absorbed.” (25) This fundamental challenge is the one I’ve tried to tackle by analyzing Ballard’s diary using text mining. There are advantages and disadvantages to such an approach – computers are very good at counting the instances of the word “God,” for instance, but less effective at recognizing that “the Author of all my Mercies” should be counted as well. The question remains, how does a reader (computer or human) recognize and conceptualize the recurrent themes that run through nearly 10,000 entries?

One answer lies in topic modeling, a method of computational linguistics that


SIGNS
@40
1975
2014
Home
Topic Model
Curated Contents
Cocitation Network
Cover Art
Editorial Comments
Overviews
Topic grid
Topic space
Topic list
Topics over time
Topic ▾
Article
Word
Bibliography
Word index
Interpreting the model
Settings
Getting started
land wives colonial africa african south kenya state
tourism cuban american asian caribbean cuba
biological males female sex male females
population reproductive control birth abortion pregnancy
color african american white black race racial class
history space subject body bodies time
western socialist russian state soviet russia
chicanas american chicana mexican mexico border
cultural taiwan hong china chinese state western kong
public city home domestic urban household
college girls school education students schools
rights animal gilligan nature moral animals care ethic
married child marriage children family families economic husband
rights issues feminists feminism feminist women movement consciousness
cultural desire narrative films film cinema work scene
cultural saint femmes paris french france roman tacitus
sex freud woman beauvoir kristeva irigaray
identity sex masculine men gender masculinity gendered man
economic international global world states united
years early century history time made
middle caste state indian india class hindu nation
australian native colonial indian people aboriginal
computers nuclear martineau computer technology language
government issues programs development women policy program organizations
state women islamic muslim islam religious
age israel palestinian life israeli middle
tokyo male japan japanese female woman
force working labor work workers employment
spanish mexico american latin america brazil
discrimination state rights law legal court
heterosexual identity queer gay lesbian lesbians community sexual
writers criticism literary woolf female literature
feminist barker romance love courtly gothic
private state political class public relations
popular images beauty public media nation
patients care physicians medicine medical health hospital nurses
ages books english medieval religious england
freud past trauma memory sexual survivors
motherhood children child mother mothers maternal
girl play performance music girls dance
view world feminist theory human knowledge
kinship societies human anthropology behavior cultural
female poet poetry language poem poems
groups national movement political politics members
identity politics cultural political difference discourse
crime justice state prison inmates prisoners
brothel brothels woman prostitutes prostitution prostitute seacole money
study psychology research sex differences behavior
status number data women percent men table female
spiritual christian god religious church religion
fetal reproductive sperm health disability genetic
work household production labor rural development
gender mathematics haraway scientific science scientists sciences nature
work trafficking workers sex police state
lane girl harassment sexual girls riot
domestic victim victims rape violence sexual abuse raped
pleasure love female sexuality sexual sex desire male
change ways important power social experience part terms
work years time people made young
male coverage dubroff sport media sports physical female
boston nineteenth suffrage american america papers
image photographs artist work art artists arts images
army peace germany military war german west security
research cancer breast health breastfeeding feeding
status society woman men women female male role
study research feminist studies work scholars
reading story narrative writing text book

THE STORY OF STOPWORDS: TOPIC MODELING AN EKPHRASTIC TRADITION



Topic Modeling
an Ekphrastic Tradition

DH2014
July 9, 2014

Lisa Marie Rhody
Center for History and New Media
George Mason University
@lmrhody
lisarhody.com

Posted by *Lisa Rhody* on May 6, 2015 in *Conferences*, *Digital Humanities*, *Reflection*, *Research*, *revisingekphrasis*, *Talks*, *Topic Modeling* |

The following paper was first presented on July 9, 2014 at DH2014 in Lausanne

Search for: Search

CALENDAR

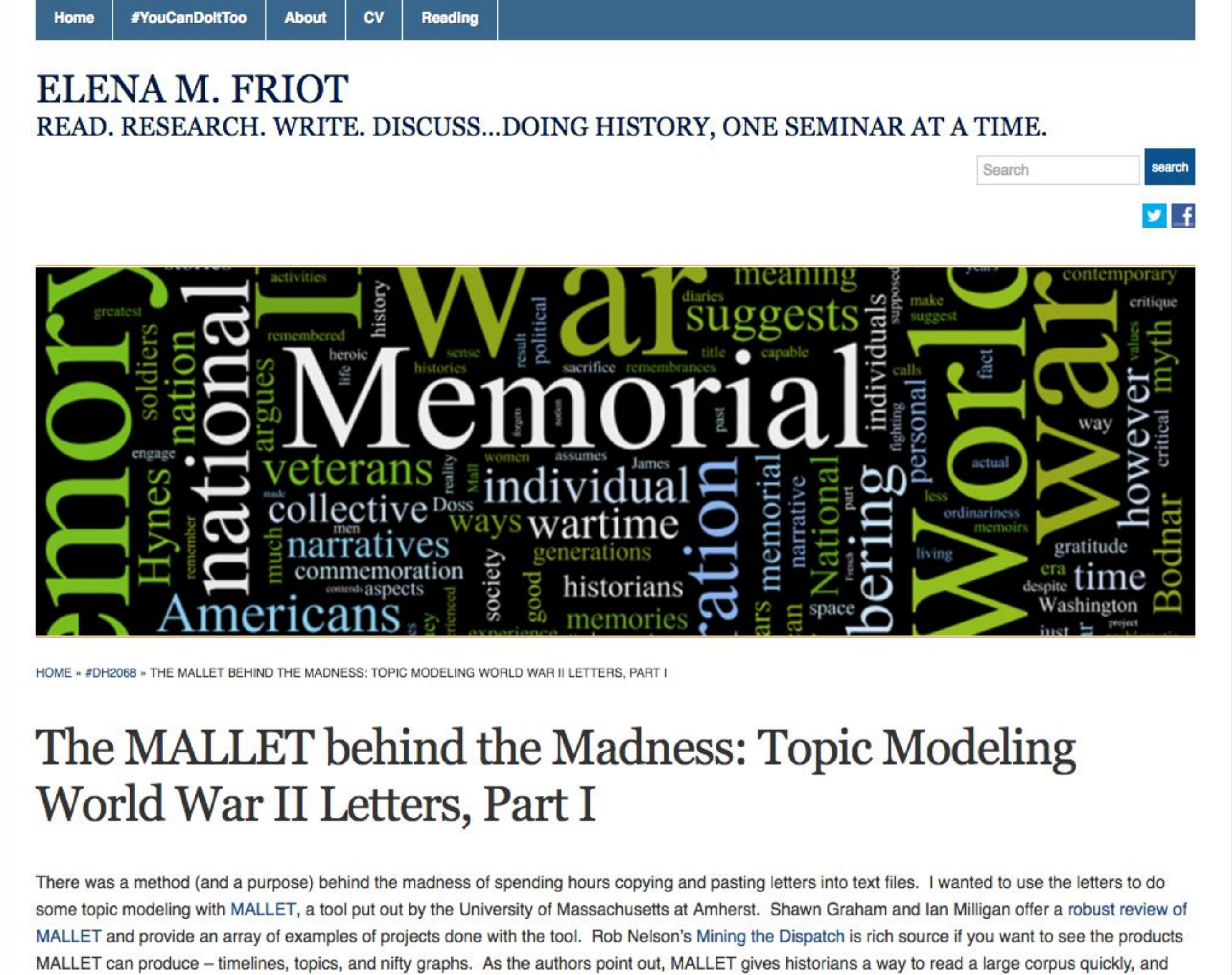
October 2016

M	T	W	T	F	S	S
					1	2
3	4		6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

« Oct

RECENT POSTS

- Gratitude



ELENA M. FRIOT

READ. RESEARCH. WRITE. DISCUSS...DOING HISTORY, ONE SEMINAR AT A TIME.

HOME » #DH2068 » THE MALLET BEHIND THE MADNESS: TOPIC MODELING WORLD WAR II LETTERS, PART I

The MALLET behind the Madness: Topic Modeling World War II Letters, Part I

There was a method (and a purpose) behind the madness of spending hours copying and pasting letters into text files. I wanted to use the letters to do some topic modeling with MALLET, a tool put out by the University of Massachusetts at Amherst. Shawn Graham and Ian Milligan offer a robust review of MALLET and provide an array of examples of projects done with the tool. Rob Nelson's Mining the Dispatch is rich source if you want to see the products MALLET can produce – timelines, topics, and nifty graphs. As the authors point out, MALLET gives historians a way to read a large corpus quickly, and

Demo data

Grimm’s Fairy Tales

Collection of German fairy tales published in 1812 by Jacob and Wilhelm Grimm

Collected and preserved stories from the oral tradition of local storytellers

Common motifs include:

- Kings, magic, and talking animals
- Archetypical characters
- Morality
- Demonstrate values important to culture

Grimm's Fairy Tales in context

Influential work of folklore, important material for disciples of Germanic studies, history, literary studies, psychoanalysis, visual and performing arts, and cultural studies

Expression of Jungian collective unconscious, archetypes, and mythology

Among the best-known stories, popularized by Disney adaptations and influence children's books

Do archetypal patterns ignore certain communities or encourage stereotypes?

Our dataset

Obtained plain text file from Project Gutenberg, containing 62 stories translated to the English by Edgar Taylor and Marian Edwardes

MALLET is optimized for shorter chunks of text, so the single plain text file was split into 62 plain text files, one story per file, through extensive cut and paste. This allows MALLET to build more specific, meaningful topics

If the researcher really wanted to go in depth with MALLET, s/he could further parse each of these 62 plain text files into single paragraphs, as we have done with “Rapunzel” for the second dataset.

Demo time!

Demo instructions

- Open Mallet in the command line
- Download dataset into MALLET directory
- Use **cd** and **ls** commands
- Execute topic modeling commands
- Generate baskets of words
- Visualize the results in Word and Excel

Discussion

What challenges did you encounter when interpreting the meaning of topics? Could you generate meaningful, descriptive topic labels? To what extent is labeling a subjective task? Do you need to be a subject expert to interpret meaning?

Can you spot any trends or similarities between the projects that utilize MALLET?

How does Mallet compare with other distant reading tools like Voyant?

Literature is an artificial universe so how does using a scientific means of analysis (algorithms) make sense when the written word, unlike the natural world, can’t be counted on to obey a set of laws?

References

Brett, M. R. (2012). Topic modeling: a basic introduction. Journal of Digital Humanities, 2(1). Retrieved from http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/

Graham, S., Weingart, S., & Milligan, I. (2012). Getting started with topic modeling and MALLET. The Programming Historian. Retrieved from http://programminghistorian.org/lessons/topic-modeling-and-mallet

McCallum, A. K. (2002). "MALLET: A Machine Learning for Language Toolkit." Retrieved from http://mallet.cs.umass.edu

Newman, D. J., & Block Sharon. (2006). Probabilistic topic decomposition of an eighteenth-century American newspaper. *Journal of the American Society for Information Science and Technology, 57*(6): 753-767. Retrieved from http://www.ics.uci.edu/~newman/pubs/JASIST_Newman.pdf

Posner, M. (2012, October 29). “Very basic strategies for interpreting results from the Topic Modeling Tool.” Miriam Posner’s Blog. Retrieved from http://miriamposner.com/blog/very-basic-strategies-for-interpreting-results-from-the-topic-modeling-tool/

Underwood, T. (2012, April 7). “Topic modeling made just simple enough.” The Stone and the Shell Blog. Retrieved from https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/