THATCamp ASEEES 2016: Topic Modeling

Carlotta Chenoweth
carlotta.chenoweth@yale.edu

1

What is Topic Modeling?

Why is it relevant for DH?
When might it be useful?

What is Topic Modeling (TM)?

- TM is a form of **large-scale** textual analysis
- Most TM uses an algorithm called Latent Direchlet Allocation (LDA)
- LDA iterates over a corpus, measuring the mutual proximities of words
- Words that frequently appear near other words are clustered together creating a “topic”
- The sum total of topics is referred to as the “topic model”

What parameters matter in TM?

- This will not work on a small corpus.
- As the executor of LDA, you decide how loosely these topics are bound: the more topics, the closer the mutual proximity of words in the corpus.
- Generally, larger corpora require more topics (but not always - it might take some finagling before you have meaningful results).

Why is TM useful?

- The computer has not sense of semantic meaning when examining a text and thus may make draw connections between terms or texts we would not see with the human eye.
- TM can be used to trace trends across enormous volumes of text, making large-scale analysis far less time consuming.

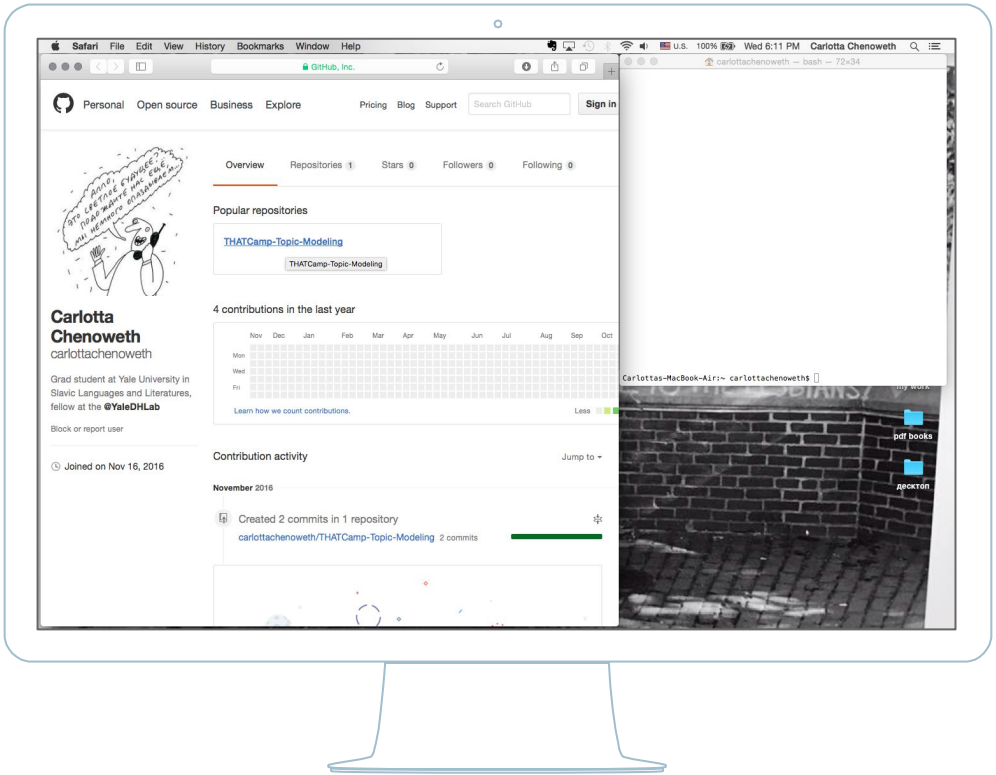
2

How to Implement TM

What software do you need?
Where do you get your corpus?

Today's object of study:
167 poems by Vladimir Mayakovsky

- Scraped from wikisource.org
- Using MALLET, LDA software developed by Andrew McCallum at UMASS
- There are other algorithms and/or programs that can be used. MALLET is the most common and user-friendly.



Go to: github.com/carlottachenoweth

Open Terminal (Mac) or Command Prompt (PC)

Online resources:

Detailed explanations for setting up MALLET on PC:

http://programminghistorian.org/lessons/topic-modeling-and-mallet

MALLET Homepage (UMASS):

http://mallet.cs.umass.edu/

List of other TM software applications from David Blei (Princeton):

https://www.cs.princeton.edu/~blei/topicmodeling.html

Stanford TM Toolbox:

http://nlp.stanford.edu/software/tmt/tmt-0.4/

Publications on TM:

- Blei, D. M., Andrew Y. Ng, and Michael I. Jordan. “Latent Dirichlet Allocation.” *The Journal of Machine Learning Research* 3 (2003): 993–1022.
- Blevins, C. 2010. *Topic Modeling: Martha Ballard’s Diary*. April 1. http://historying.org/2010/04/01/topicbmodelingbmarthabballardsbdiary/
- Block, S. 2006. “Doing More with Digitization: An Introduction to Topic Modeling of Early American Sources.” *Common_place: The Interactive Journal of Early American Life* 6, no. 2.
- Buzzetti, Dino. 2002. “Digital Representation and the Text Model.” *New Literary History* 33.1: 61-88.
- Chang, J., J. Boyd, Graber, S. Gerrish, C. Wang, and D. Blei. 2009. “Reading Tea Leaves: How Humans Interpret Topic Models.” In *Advances in Neural Information Processing Systems* 22: 288–96. Proceedings of the 2009 conference. Norwich, UK: Curran Associates.
- Epstein, Joshua M. 2008. “Why model?” *Journal of Artificial Societies and Social Simulation* 11.4.
- Frye, Northrop. 1991. “Literary and Mechanical Models.” In *Research in Humanities Computing 1*. Selected papers from the 1989 ACH-ALLC Conference. Ed. Ian Lancashire. Oxford: Clarendon Press: 3-12.
- Goldstone, Andrew, and Ted Underwood. "What can topic models of PMLA teach us about the history of literary scholarship." *Journal of Digital Humanities* 2.1 (2012): 40-49.
- Stubbs, Michael. 2005. “Conrad in the computer: Examples of quantitative stylistic method.” *Language and Literature* 14.1: 5-24.
- Underwood, Ted. 2015. “Seven Ways Humanists Are Using Computers to Understand Text.” *The Stone and the Shell*. http://tedunderwood.com/2015/06/04/seven-ways-humanists-are-using-computers-to-understand-text/