

The background is white with several large, colorful circles and dashed lines. In the top left, there is a large teal circle with a white center, and a smaller teal circle next to it. In the top right, there is a large lime green circle with a smaller green circle inside it, both with dashed outlines. In the bottom left, there is a large green circle with a white center, and a smaller yellow circle next to it. In the bottom right, there is a large yellow circle with a white center, and a smaller orange circle next to it. A dashed line forms a large circle around the text.

THATCamp ASEEEES 2016: Topic Modeling

Carlotta Chenoweth
carlotta.chenoweth@yale.edu

The background is white with several decorative elements: a large orange circle with a dashed red outline in the top left; a large yellow circle below it; a small pink circle below that; a large green circle with a dashed yellow outline in the top right; a small orange circle above it; a large blue circle with a dashed blue outline in the bottom right; a large green circle with a dashed green outline in the bottom left; and a small cyan circle above it. A large, light blue dashed circle is centered in the upper half of the slide.

1

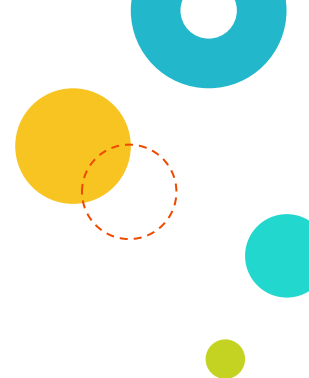

What is Topic Modeling?

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 Dirichlet 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”
- 
- 

A decorative graphic on the left side of the slide features several overlapping circles in pink, orange, teal, and lime green. Some circles have dashed outlines. At the bottom left, there is a green circle containing a white icon of an open book. On the right side, there are more circles in teal, yellow, and green, some with dashed outlines.

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).

A decorative graphic on the left side of the slide. It features a large teal ring at the top left, an orange circle below it, a yellow ring, a pink circle, and a green circle at the bottom. A green circle at the bottom left contains a white icon of an open book. On the right side, there are several smaller circles in teal, yellow, and green, some with dashed outlines.

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.

The background is white and decorated with various colorful circles and dashed lines. In the top left, there is a large orange circle with a dashed red outline, overlapping a yellow circle. Below the yellow circle is a small pink circle. In the top center, a large blue number '2' is centered within a large dashed light blue circle. In the top right, there is a green circle with a white dot in the center, a small orange circle, and a lime green circle with a dashed yellow outline. In the bottom left, there is a green circle with a dashed green outline, a large lime green circle, and a small cyan circle. In the bottom right, there is a large cyan circle with a white dot in the center, a small cyan circle with a dashed blue outline, and a small cyan circle.

2

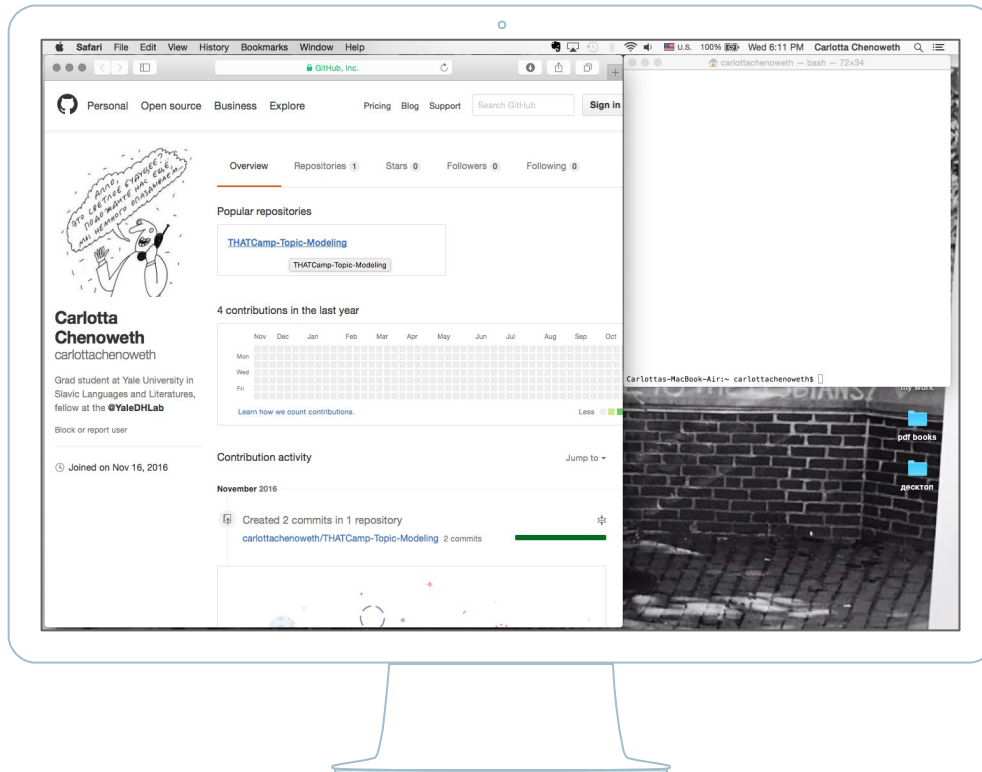
How to Implement TM

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)

A decorative graphic consisting of a large, light-blue dashed circle that frames the central text. Scattered around this circle are various smaller circles in different colors: teal, yellow, green, orange, and pink. Some of these circles are solid, while others are hollow or have a dashed outline. The overall aesthetic is modern and playful.

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/topicbmodelingbmarthabbballardsbdiary/>
- 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/>