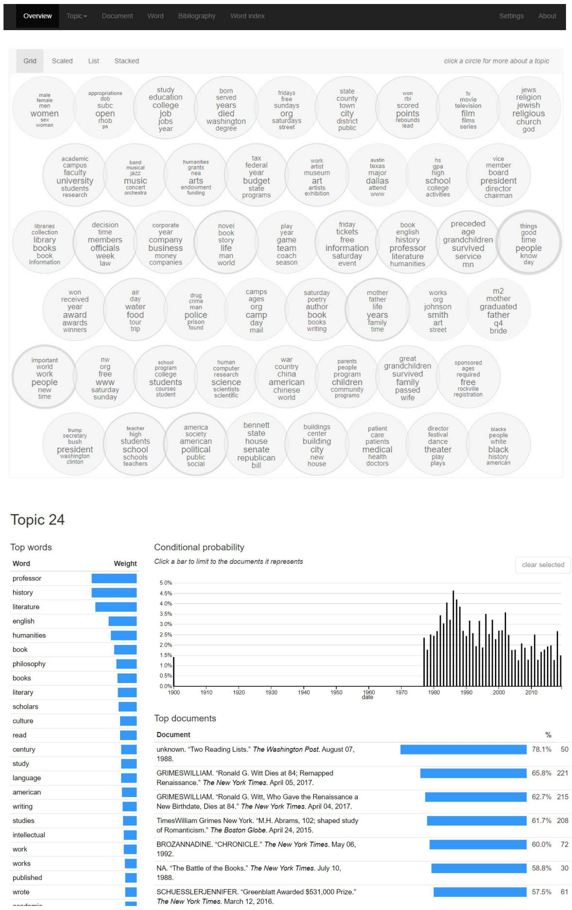
WE1S Methods

"Topic modeling" is WE1S's main method for studying collections of documents at big data scale.

Topic modeling is a leading method of machine learning that discovers "topics" in texts by analyzing the statistical co-occurrence of words. It finds out which words tend to come up together in a document set (and in individual texts) when people discuss something or, as in newspapers, many things. An article about a political election, for example, might frequently use words naming a nation's capital city (e.g., "London") with other words like "Parliament." Co-occurring words suggest "topics."

Topic modeling not only identifies topics but also indicates their weights. (For example, politics might be more important than economics in one newspaper.) It also shows which specific documents participate strongly in a topic (or several topics at once, since an article that mentions "London" might talk partly about politics but also partly about economics because London is also a financial capital). Topic models thus not only help with "distant reading" large collections but also guide researchers to specific documents to "close read" because of their statistical association with a topic.

Topic modeling is an "unsupervised" machine-learning method, meaning that it does not require pre-training of the computer on a gold standard of texts that humans previously analyzed. The specific kind of topic modeling that WE1S uses is Latent Dirichlet Allocation (LDA) as implemented in the MALLET Machine Learning for Language Toolkit.


Overview
Topic
Document
Word
Bibliography
Word index
Settings
About
Grid
Scaled
List
Stacked
click a circle for more about a topic
male female men women sex woman
appropriations dob subc open rhob ps
study education college job jobs year
born served years died washington degree
fridays free sundays org saturdays street
state county town city district public
won rbi scored points rebounds lead
tv movie television film films series
jews religion jewish religious church god
academic campus faculty university students research
band musical jazz music concert orchestra
humanities grants nea arts endowment funding
tax federal year budget state programs
work artist museum art artists exhibition
austin texas major dallas attend www
hs gpa high school college activities
vice member board president director chairman
libraries collection library books book information
decision time members officials week law
corporate year company business money companies
novel book story life man world
play year game team coach season
friday tickets free information saturday event
book english history professor literature humanities
preceded age grandchildren survived service mn
things good time people know day
won received year award awards winners
air day water food tour trip
drug crime man police prison found
camps ages org camp day mail
saturday poetry author book books writing
mother father life years family time
works org johnson smith art street
m2 mother graduated father q4 bride
important world work people new time
nw org free www saturday sunday
school program college students courses student
human computer research science scientists scientific
war country china american chinese world
parents people program children community programs
great grandchildren survived family passed wife
sponsored ages required free rockville registration
trump secretary bush president washington clinton
teacher high students school schools teachers
america society american political public social
bennett state house senate republican bill
buildings center building city new house
patient care patients medical health doctors
director festival dance theater play plays
blacks people white black history american
Conditional probability
Click a bar to limit to the documents it represents
clear selected
5.0%
4.5%
4.0%
3.5%
3.0%
2.5%
2.0%
1.5%
1.0%
0.5%
0.0%
1900
1910
1920
1930
1940
1950
1960
date
1970
1980
1990
2000
2010

Topic 24

Top words

Word	Weight
professor	
history	
literature	
english	
humanities	
book	
philosophy	
books	
literary	
scholars	
culture	
read	
century	
study	
language	
american	
writing	
studies	
intellectual	
work	
works	
published	
wrote	

Top documents

Document	%	
unknown. "Two Reading Lists." *The Washington Post*. August 07, 1988.	78.1%	50
GRIMESWILLIAM. "Ronald G. Witt Dies at 84; Remapped Renaissance." *The New York Times*. April 05, 2017.	65.8%	221
GRIMESWILLIAM. "Ronald G. Witt, Who Gave the Renaissance a New Birthdate, Dies at 84." *The New York Times*. April 04, 2017.	62.7%	215
TimesWilliam Grimes New York. "M.H. Abrams, 102; shaped study of Romanticism." *The Boston Globe*. April 24, 2015.	61.7%	208
BROZANNADINE. "CHRONICLE." *The New York Times*. May 06, 1992.	60.0%	72
NA. "The Battle of the Books." *The New York Times*. July 10, 1988.	58.8%	30
SCHUESSLERJENNIFER. "Greenblatt Awarded $531,000 Prize." *The New York Times*. March 12, 2016.	57.5%	61

Topic model of newspaper articles visualized in Dfr-browser (a topic model interface by Andrew Goldstone)

Audience for key finding: Scholarly, Digital Humanists

Introductions toTopic Modeling: WE1S bibliography

Recommended articles:

* Blei, David M. "Probabilistic Topic Models." (2012)

* Underwood, Ted. "Topic Modeling Made Just Simple Enough" (2012)

* Mohr, John, and Petko Bogdanov, "Topic Models: What They Are and Why They Matter" (2013)

30 December 2019 (Alan Liu)