



# Distant Reading in

## Topic Modeling

Simone Rebora & Giovanni Pietro Vitali

[simone.rebora@univr.it](mailto:simone.rebora@univr.it)

[giovannipetrovitali@gmail.com](mailto:giovannipetrovitali@gmail.com)



# TOPIC MODELS

According to David Blei:

“Topic models are a suite of algorithms that **uncover the hidden thematic structure** in document collections. These algorithms help us develop new ways **to search, browse and summarize** large archives of texts”

(<http://www.cs.columbia.edu/~blei/topicmodeling.html>)



# TOPIC MODELS

## Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

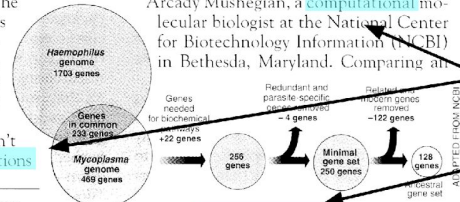
## Documents

### Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic numbers** game, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

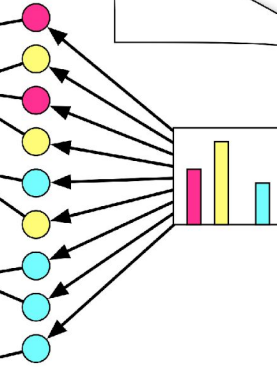


\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. **Computer analysis** yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

## Topic proportions and assignments





(Blei, 2012)



# LDA TOPIC MODELS

LDA = Latent Dirichlet Allocation

- 
- a topic is a distribution of probabilities of words
  - all words in a document can belong to all topics
  - a document is a distribution of probabilities of topics
- 



# LDA TOPIC MODELS

*a topic:*

sole (10.1%)  
cuore (6.4%)  
amore (4.7%)  
...

*a word:*

amore

4.7%

7.1%

5.8%

12.4%

5.2%

15.8%

*bad  
poetry*

*sentiments*

*very bad  
poetry*

*a document:*



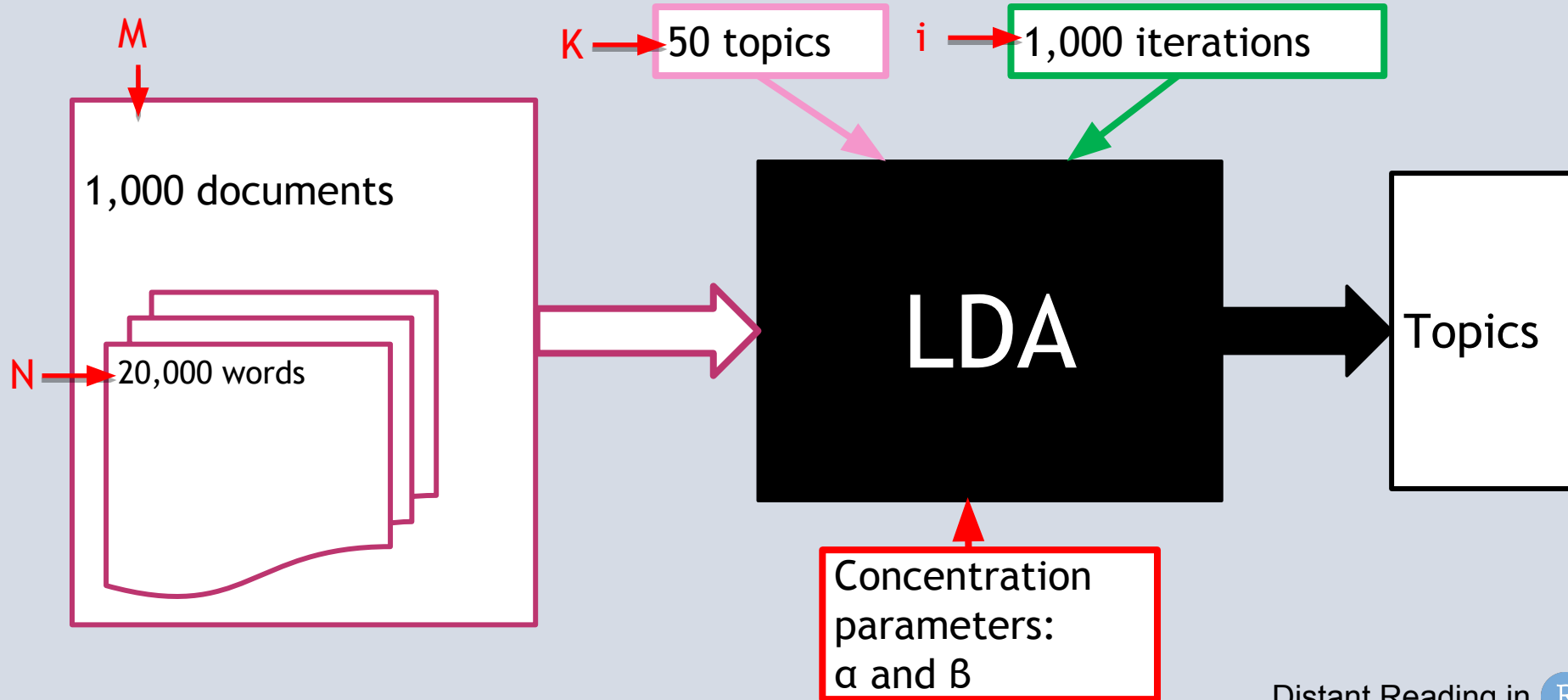


# LDA: HOW DOES IT WORK?

- ▶ Initialize topic assignments **randomly**
- ▶ **For each word** in each document:
  - ▶ **re-sample topic** for word,  
given all other words and their current topic assignments
- ▶ Iterate  $n$  times!



# LDA: HOW DOES IT WORK?





iteration #1,456

## LDA'S "BLACK BOX"

document #151

7

12

12

5

2

17

9

branch lake como turns south unbroken chains

7

10% (cinema 12%, branch 10%, movie 9.5%, actor 9.5%...)

12

9% (Como 15%, lake 14%, meatball 10.5%, branch 9%...)

2

6%





iteration #1,456

## LDA'S "BLACK BOX"

document #151



12

12

5

2

17

9

branch lake como turns south unbroken chains

7

10% (cinema 12%, branch 10%, movie 9.5%, actor 9.5%...)

12

9% (Como 15%, lake 14%, meatball 10.5%, branch 9%...)

2

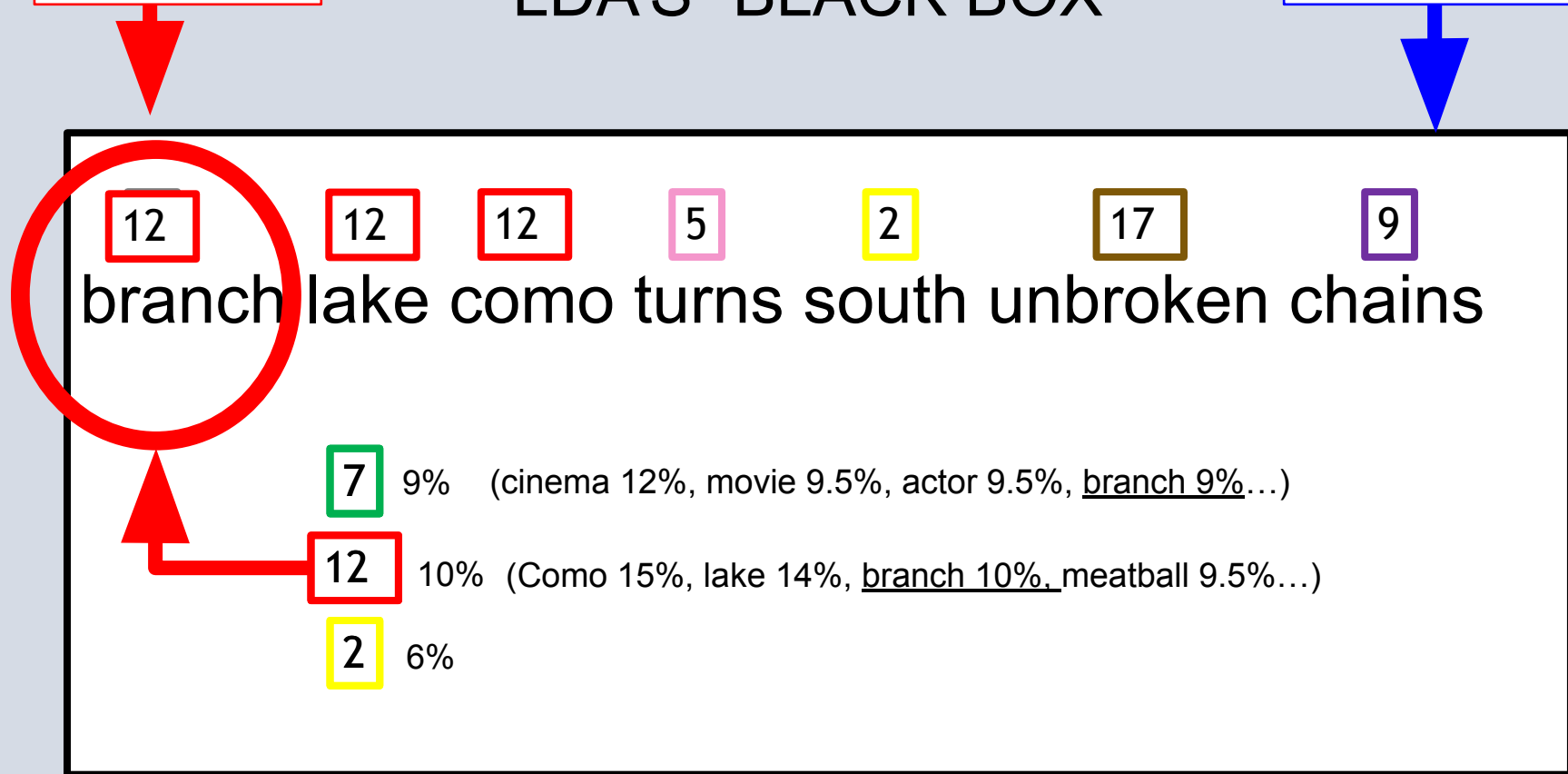
6%



iteration #1,456

## LDA'S "BLACK BOX"

document #151





# APPLICATIONS

[home](#) | [submissions](#) | [about dhq](#) | [dhq people](#) | [contact](#) [Search](#)

## Current Issue

[» 2017: 11.4](#)

## Preview Issue

[» 2018: 12.1](#)

## Previous Issues

[» 2017: 11.3](#)[» 2017: 11.2](#)[» 2017: 11.1](#)[» 2016: 10.4](#)[» 2016: 10.3](#)[» 2016: 10.2](#)[» 2016: 10.1](#)[» 2015: 9.4](#)[» 2015: 9.3](#)[» 2015: 9.2](#)[» 2015: 9.1](#)[» 2014: 8.4](#)[» 2014: 8.3](#)[» 2014: 8.2](#)[» 2014: 8.1](#)

2017

Volume 11 Number 2

[2017 11.2](#) | [XML](#) | [Discuss](#) ([0 Comments](#))

## Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama

[Christof Schöch](#), [<christof\\_dot\\_schoech\\_at\\_uni-wuerzburg\\_dot\\_de>](#), University of Würzburg, Germany

### Abstract

The concept of literary genre is a highly complex one: not only are different genres frequently defined on several, but not necessarily the same levels of description, but consideration of genres as cognitive, social, or scholarly constructs with a rich history further complicate the matter. This contribution focuses on thematic aspects of genre with a quantitative approach, namely Topic Modeling. Topic Modeling has proven to be useful to discover thematic patterns and trends in large collections of texts, with a view to class or browse them on the basis of their dominant themes. It has rarely if ever, however, been applied to collections of dramatic texts.

In this contribution, Topic Modeling is used to analyze a collection of French Drama of the Classical Age and the Enlightenment. The general aim of this contribution is to discover what semantic types of topics are found in this collection, whether different dramatic subgenres have distinctive dominant topics and plot-related topic patterns, and inversely, to what extent clustering methods based on topic scores per play produce groupings of texts which agree with more conventional genre distinctions. This contribution shows that interesting topic patterns can be detected which provide new insights into the thematic, subgenre-related structure of French drama as well as into the history of French drama of the Classical Age and the Enlightenment.



# APPLICATIONS

“The data used in this study comes from the Théâtre classique collection maintained by Paul Fièvre (2007-2015). At the time of writing, this continually-growing, freely available **collection of French dramatic texts contained 890 plays published between 1610 and 1810**, thus covering the Classical Age and the Enlightenment.” **(Schöch, 2017)**



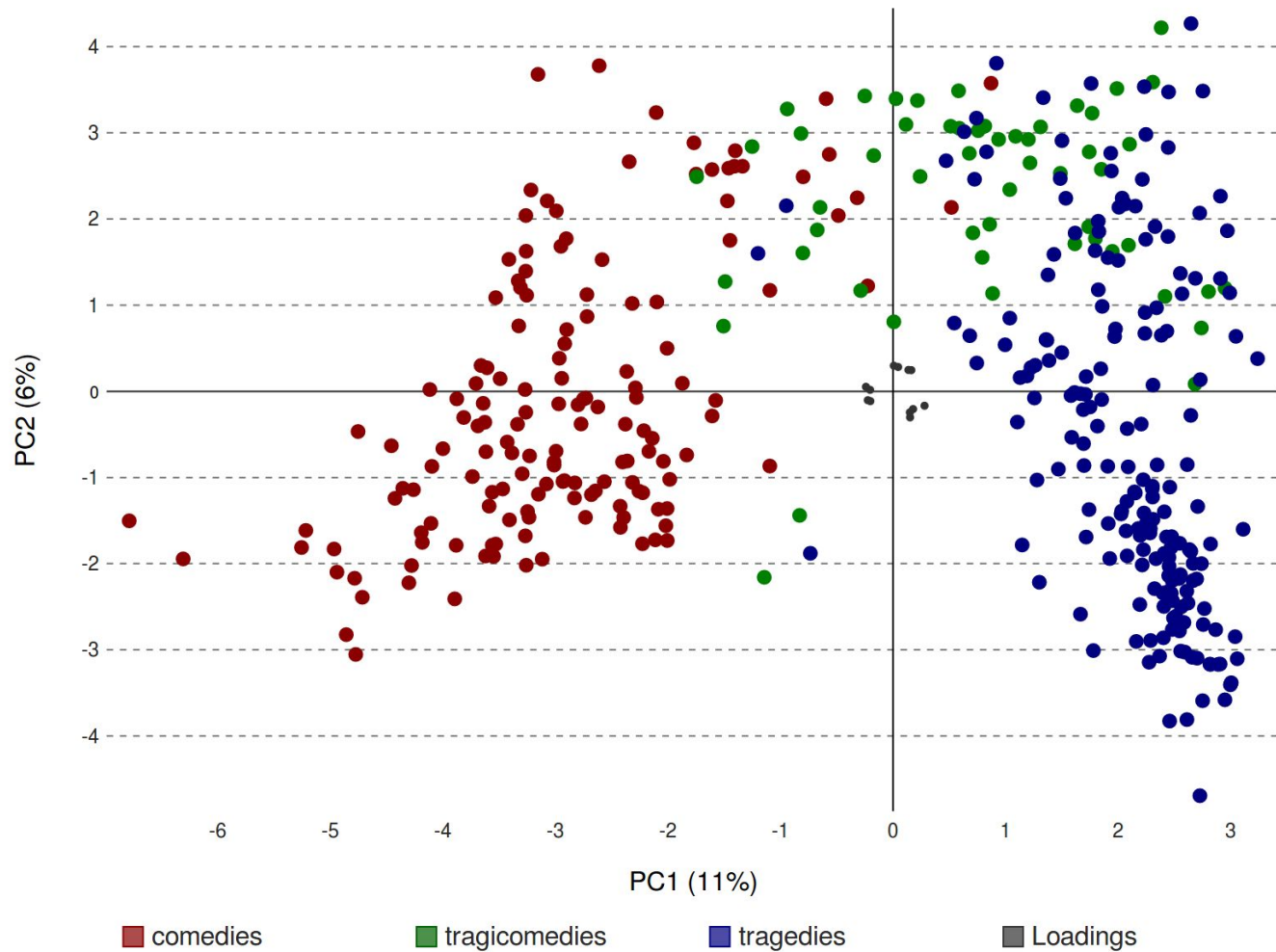
topic 32 (1/60)  
oser aimer souffrir espoir connaître  
seul âme peine craindre doux hymen souffrir écart  
secret vain offrir laisser sort foi  
croire plaire amour prix princesses plaindre voeu  
intérêt soigner ardeur mériter gloire flamme  
madame choix

topic 3 (6/60)  
être même expliquer attendre surprendre chercher esprit  
trouver effet passer aimer seul crois entendre moins penser besoin  
mystère ignorer doute soupçon temps part  
soin oser paraître taire tenir croire  
ami avoir cacher apprendre découvrir avouer sembler

topic 30 (53/60)  
poète seul sujet génie muse mauvais talent  
goût art rôle scène  
bon théâtre jouer nom merveille  
temps ouvrage acteur écrire rime premier public  
nouveau trouver représenter  
esprit beau lire pièce  
écrit prose commencer nommer sonnet plaisir

topic 34 (57/60)  
remède chose vif malade savant  
science fou statue connaître  
astrologie effet jours art charlatan vapeur docteur  
santé seul maladie corps folie cause homme  
médecin guérir habile mourir  
vrai goutte secours soutenir raison sentir

(Schöch,  
2017)





(Schöch, 2017)





# APPLICATIONS

**Journal of  
Digital Humanities**

Search JDH 

AboutVolumesSubmissions


 Subscribe to the RSS

Table of Contents for  
Vol. 2, No. 1 Winter  
2012

Introductions

Beginnings

**Applications and Critiques**

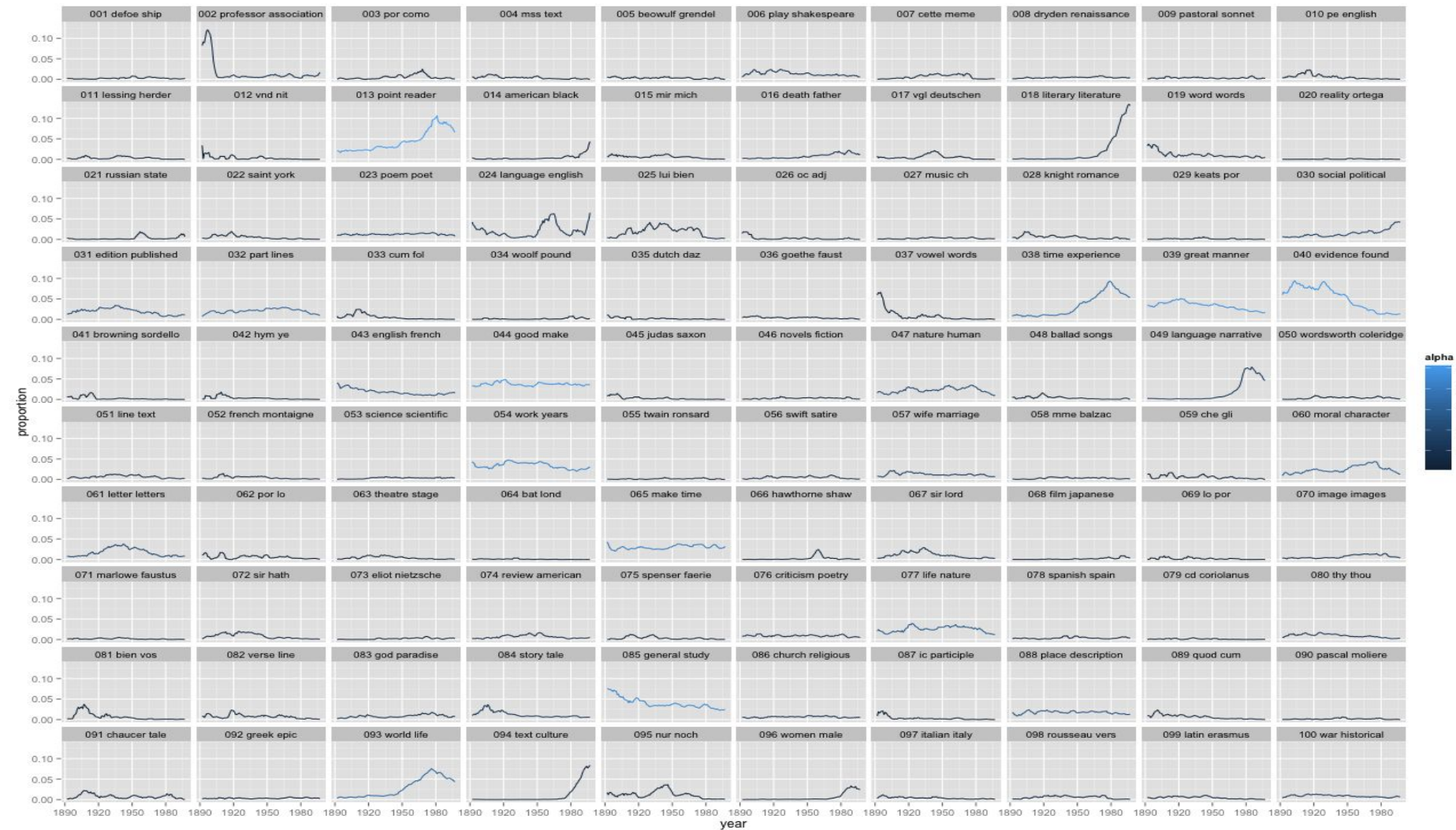
Topic Modeling and Figurative  
Language  
Lisa M. Rhody

Topic Model Data for Topic  
Modeling and Figurative

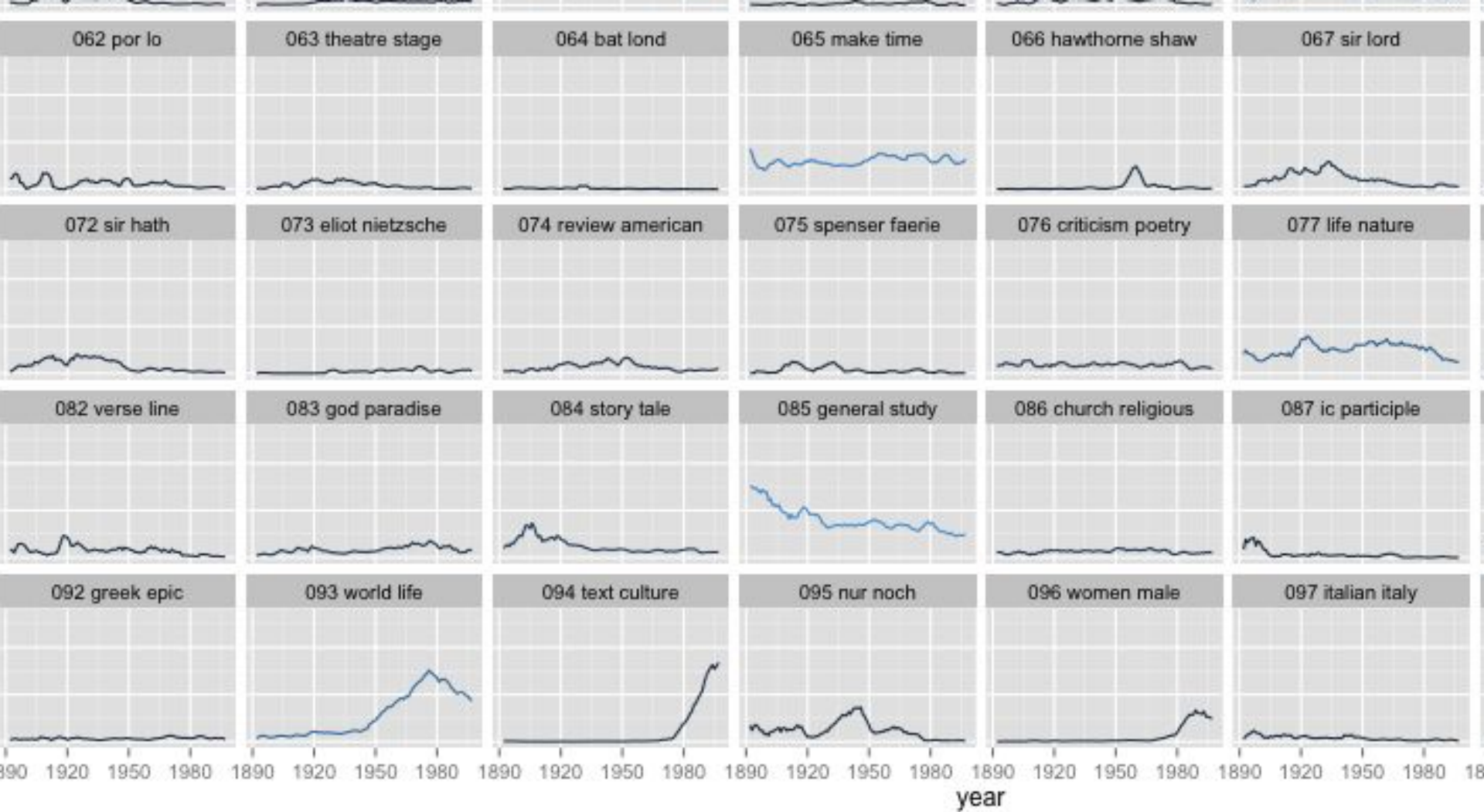
What Can Topic Models of PMLA  
Teach Us About the History of  
Literary Scholarship?

TED UNDERWOOD AND ANDREW GOLDSTONE

Of all our literary-historical narratives it is the history of criticism itself that seems most wedded to a stodgy history-of-ideas approach — narrating change through a succession of stars or contending schools. While scholars like John Guillory and Gerald Graff have produced subtler models of disciplinary history, we could still do more to complicate the narratives that organize our discipline's







Underwood's model of PMLA 1924-2006.

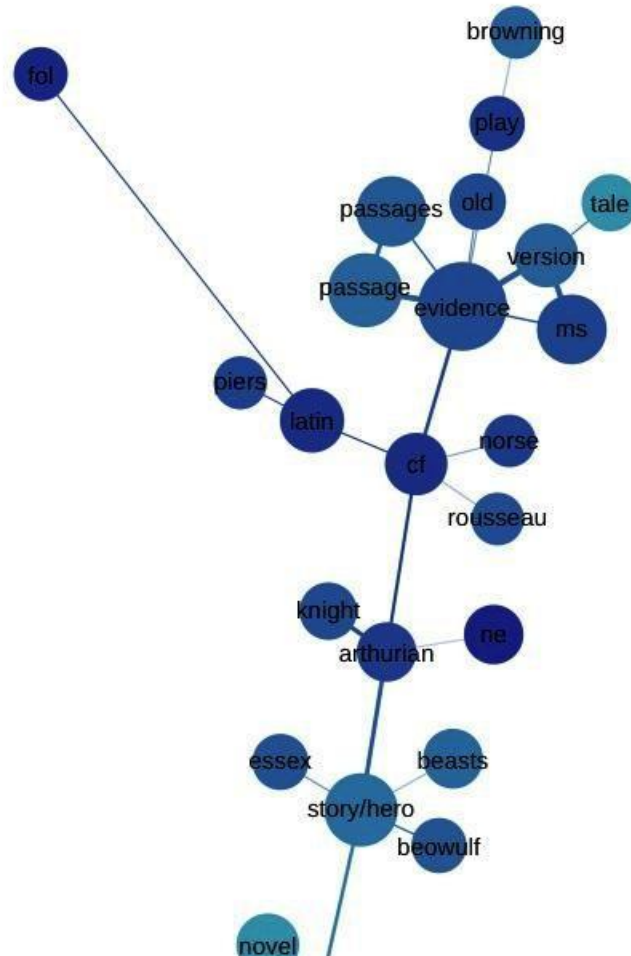
Mouse over any circle to get a longer list of words in that topic; in many cases you can also click through to get a scatterplot of yearly frequencies and a list of articles where the topic was most prominent.

The size of each circle (loosely) reflects the number of words in the topic. Blue topics are older, yellow-green topics closer to 2006. Topics are connected if they tend to appear in the same articles, and the thickness/closeness of the link reflects strength of correlation.

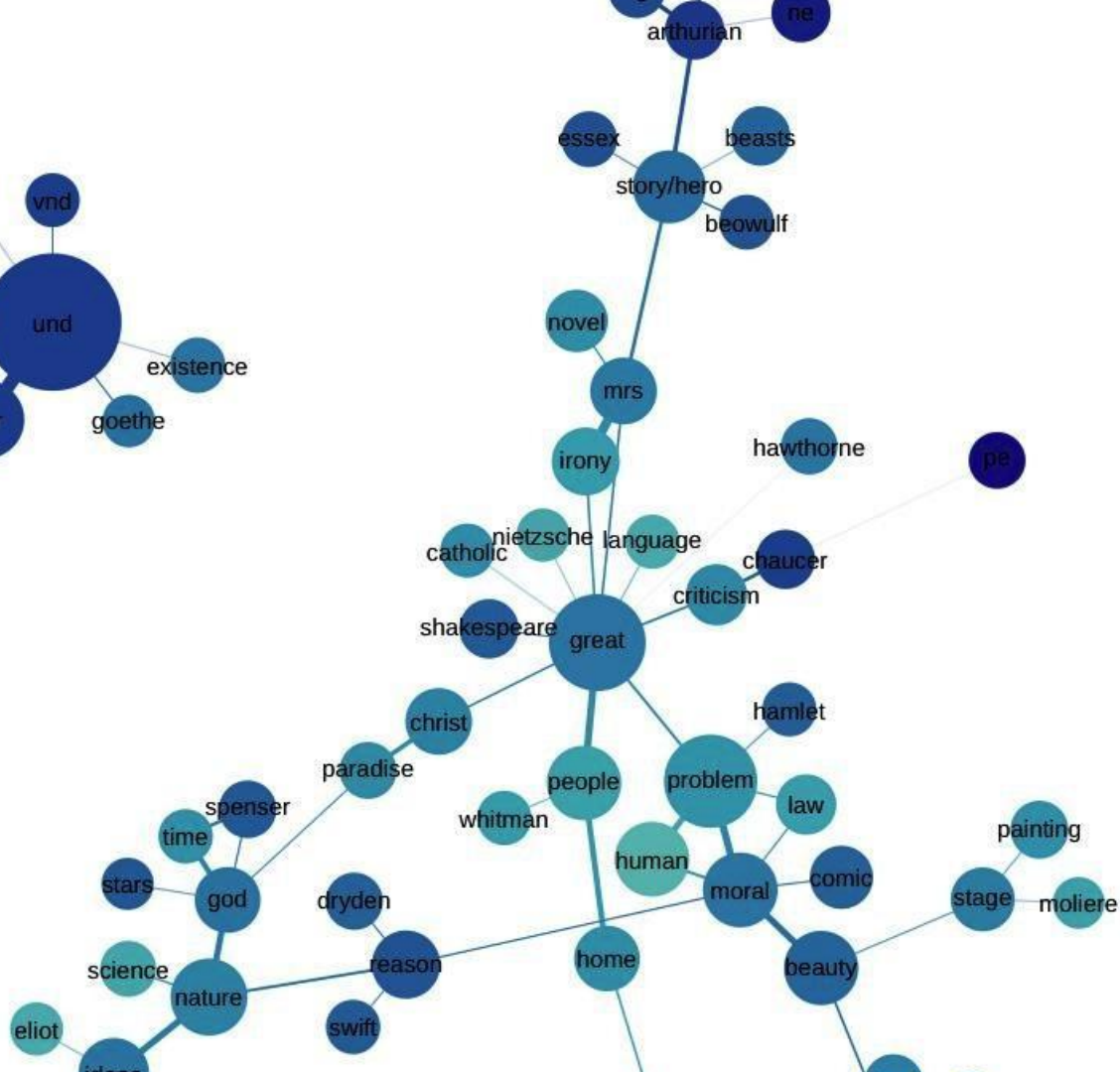
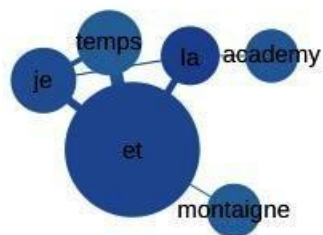
Very common English words are

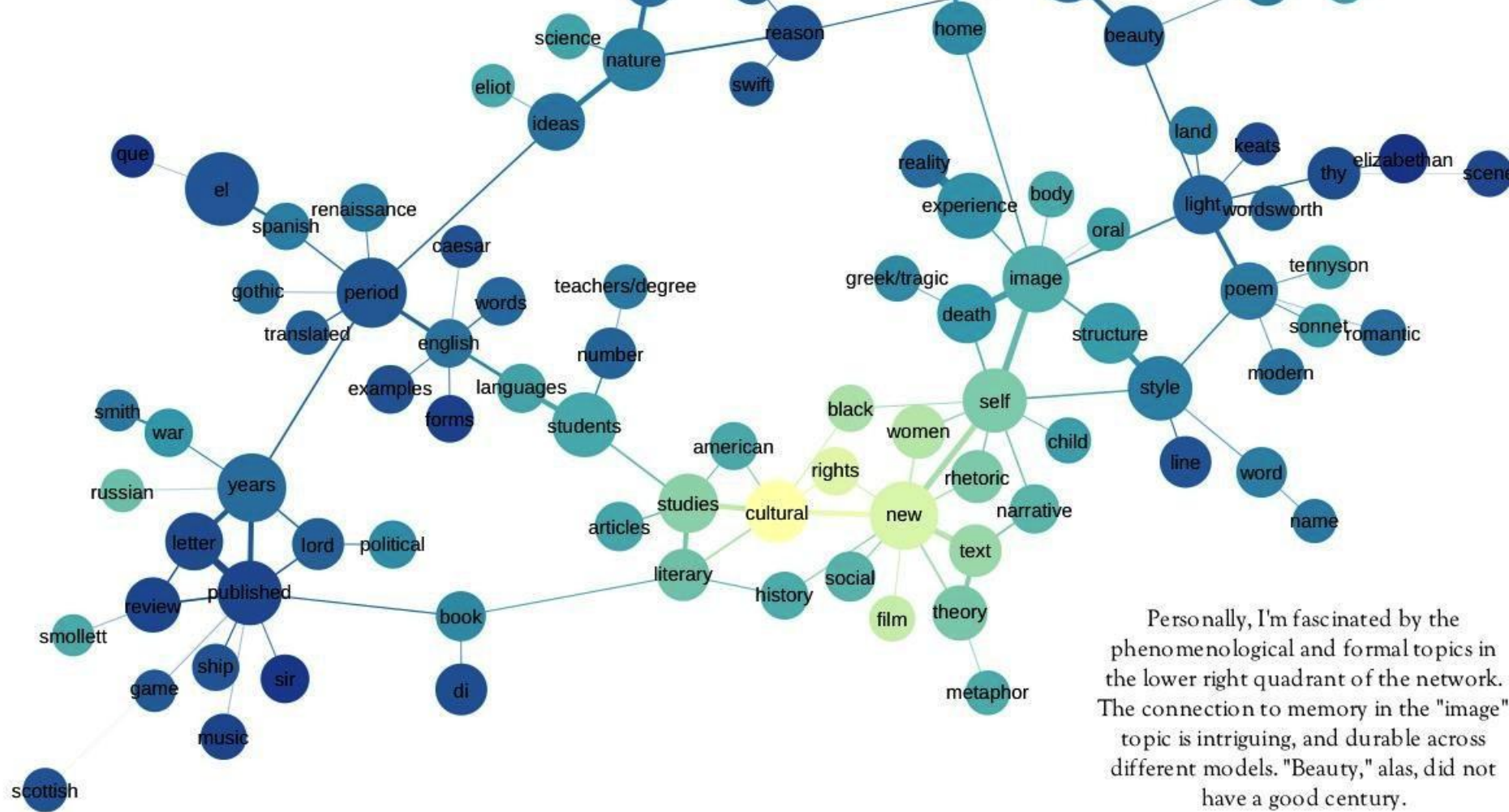


The window may have to be full width for click-through to work.



Very common English words are excluded from the model, but I didn't do the same thing with other languages, so these French and German networks tend to be dominated by uninformative function words.





# WHAT'S IN A TOPIC MODEL?

- The concept of topic (or thema) in **functionalist linguistics**?
  - The notion of isotopy in **structuralism and semiotics**?
  - The concepts of theme and motif in **thematic criticism**?
    - The **Foucaultian** notion of «discourse»?

“The discussion on the possible literary-semiotic interpretations of the notion of topic model and the observation of the theoretical difficulties they present leads us to affirm that in fact **it is not possible to find a single satisfactory theoretical-literary correlate** of the results of these methods of quantitative analysis”

**(Ciotti, 2017)**