

# Topic Modeling

# 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  
...

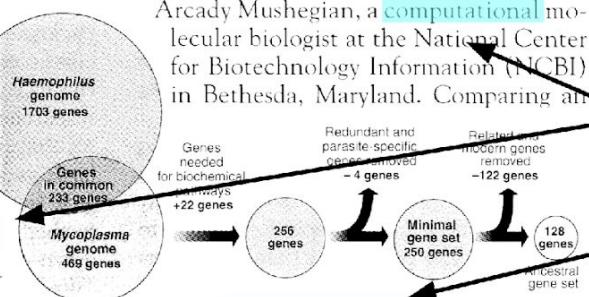
Documents

## 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

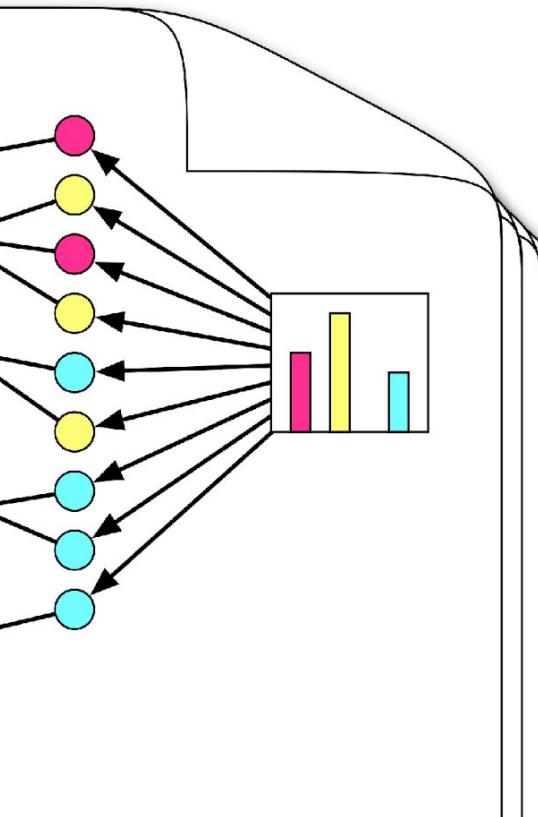
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes 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 all



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%)
- ...

*bad poetry*

*a word:*

**amore**

4.7%  
7.1%  
5.8%

12.4%  
5.2%  
15.8%

*sentiments*

*very bad poetry*

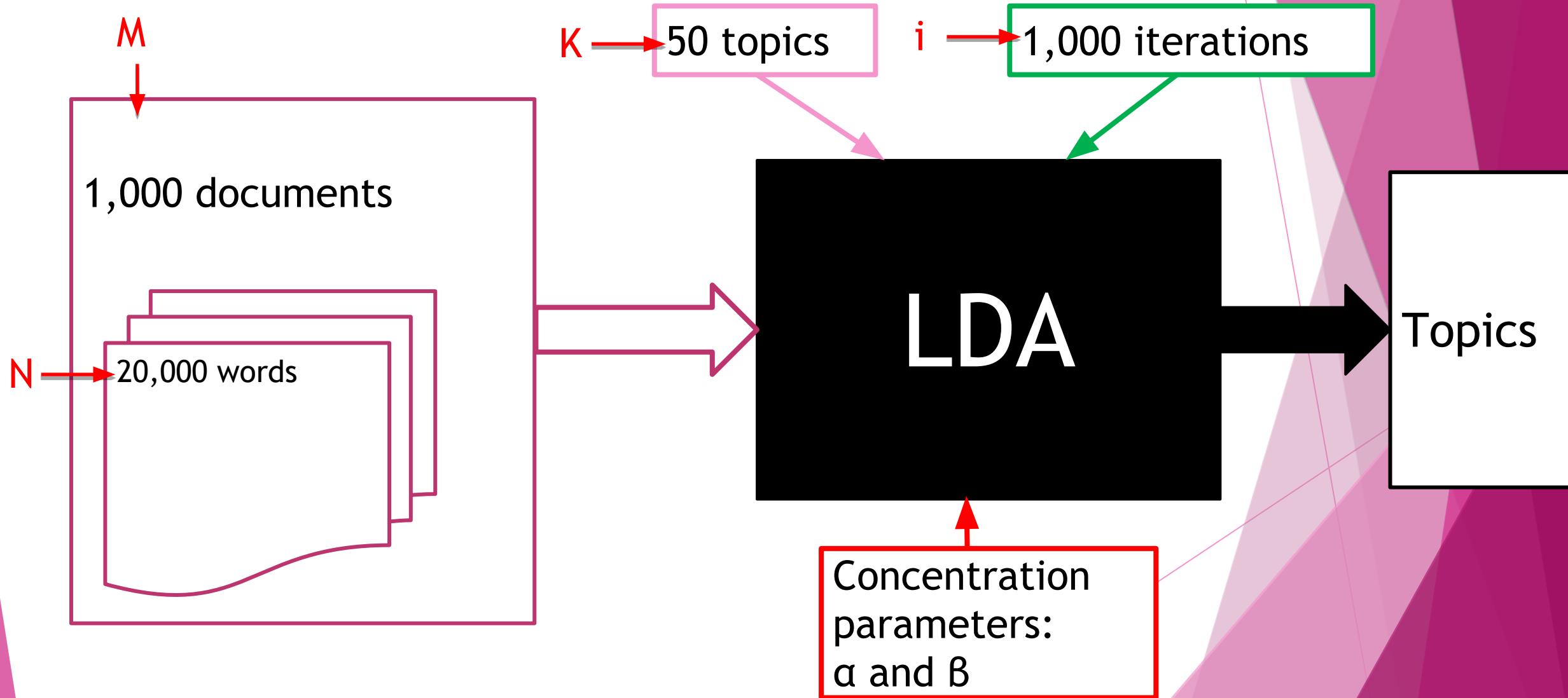
*a document:*



# LDA: How Does it Work?

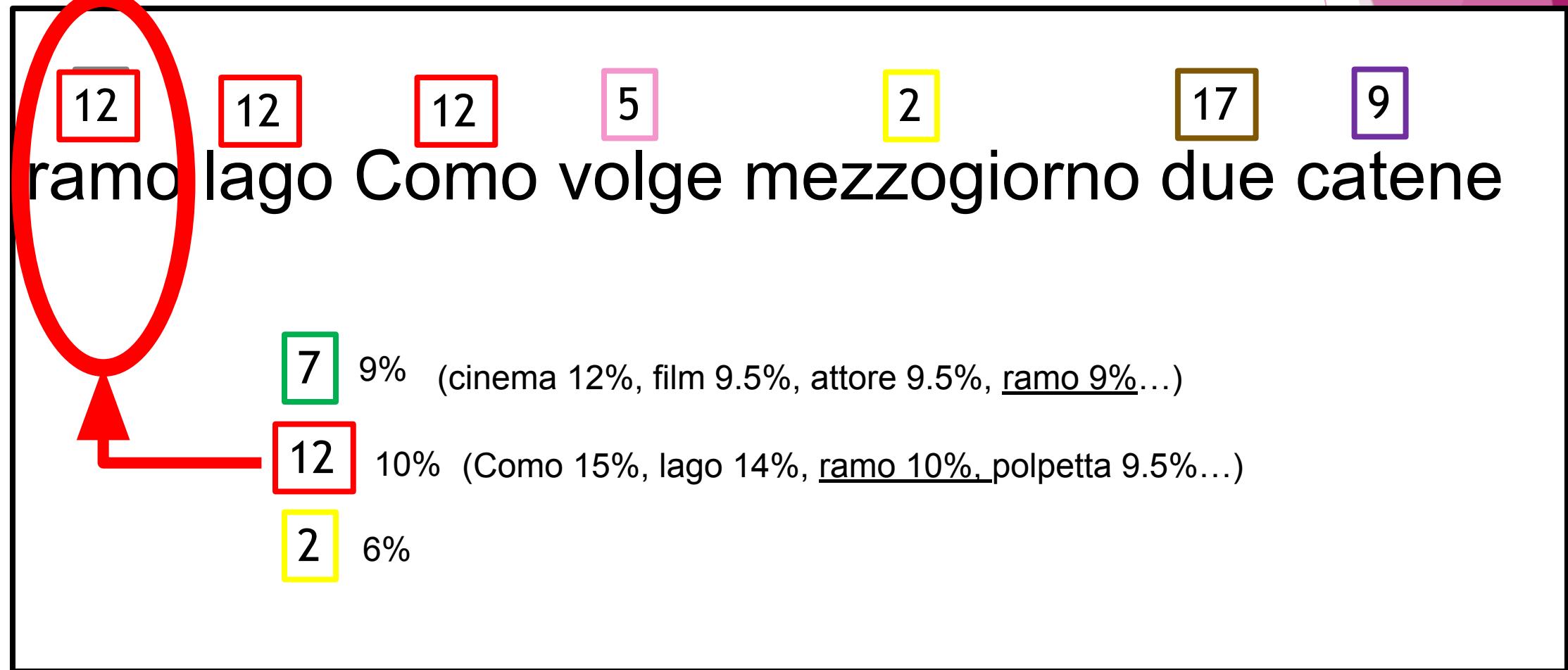
- ▶ Initialize topic assignments **randomly**
- ▶ **For each word** in each document:
  - ▶ **re-sample topic** for word,  
given all other words and their current topic assignments
- ▶ Iterate  $n$  times!

# LDA: How Does it Work?



# LDA's «Black Box»

iteration #1,456



# TM Applications in Literature



home | submissions | about dhq | dhq people | contact

Search

## Current Issue

[» 2017: 11.4](#)

2017

Volume 11 Number 2

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

## 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](#)

[» 2013: 7.3](#)

[» 2013: 7.2](#)

[» 2012: 7.1](#)

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

[Christof Schöch](#) <[christof.dot.schoech@uni-wuerzburg.de](mailto:christof.dot.schoech@uni-wuerzburg.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.

# TM Applications in Literary Studies

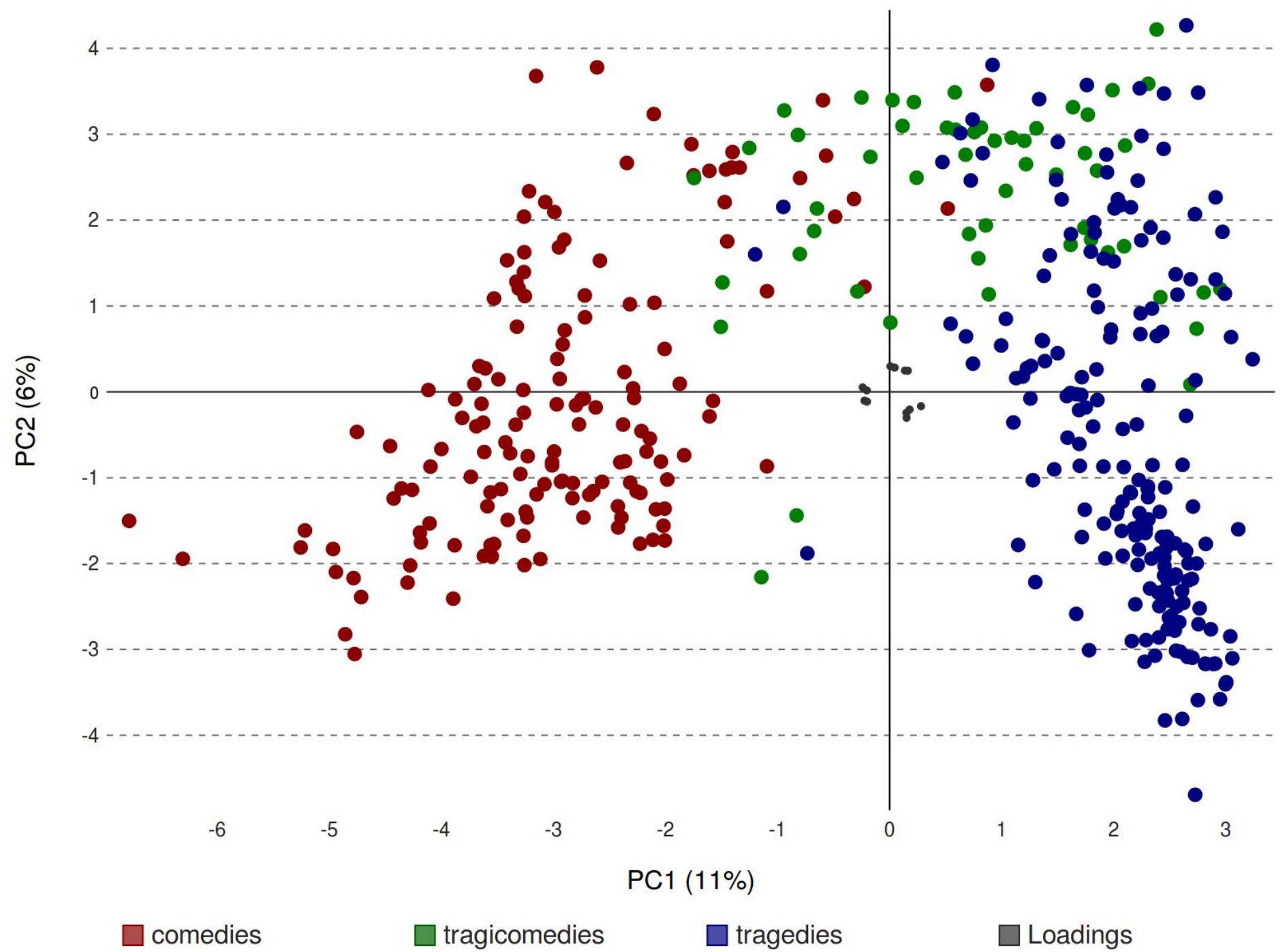
“The data used in this study comes from the Théâtre classique collection maintained by Paul Fièvre (2007-2015). At the time of writing, this continually-growing, freely available collection of French dramatic texts contained 890 plays published between 1610 and 1810, thus covering the Classical Age and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

topic 32 (1/60)  
presser • flatter effort vertu  
**oser** seul âme peine craindre doux  
**aimer** secret vain offrir laisser hymen  
**coeur** croire plaisir aveu prix  
**amour** intérêt soin forcer ardeur mériter  
**madame** gloire choix flamme  
**gloire** princesse plaindre voeu

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

topic 3 (6/60)  
être même expliquer attendre surprendre chercher esprit  
trouver effet passer aimer seul crois entendre moins  
**craindre**神秘 peine ignorer doute  
**secret** soin oser paraître taire tenir croire  
**connaître** ami avoir cacher apprendre avouer sembler  
découvrir

topic 34 (57/60)  
remède chose vif malade savant  
**monsieur** manquer science fou statue connaître  
astrologie effet jour charlatans seul corps folie  
**ma** art vapeur docteur avis cause  
maladie santé humeur  
**médecin** goutte secours soutenir sentir raison  
**guérir** habile mourir

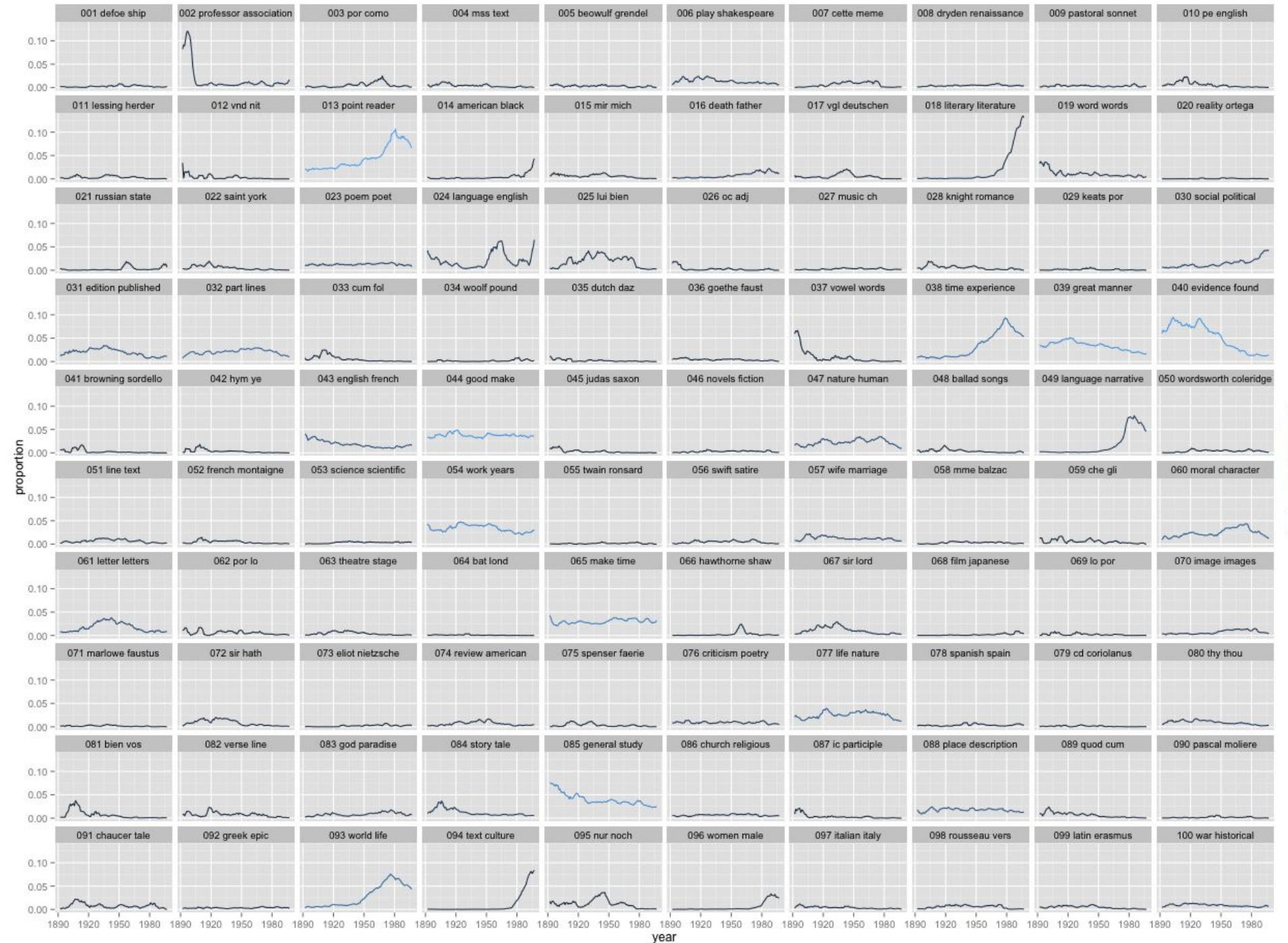


# TM Applications in Literary Studies

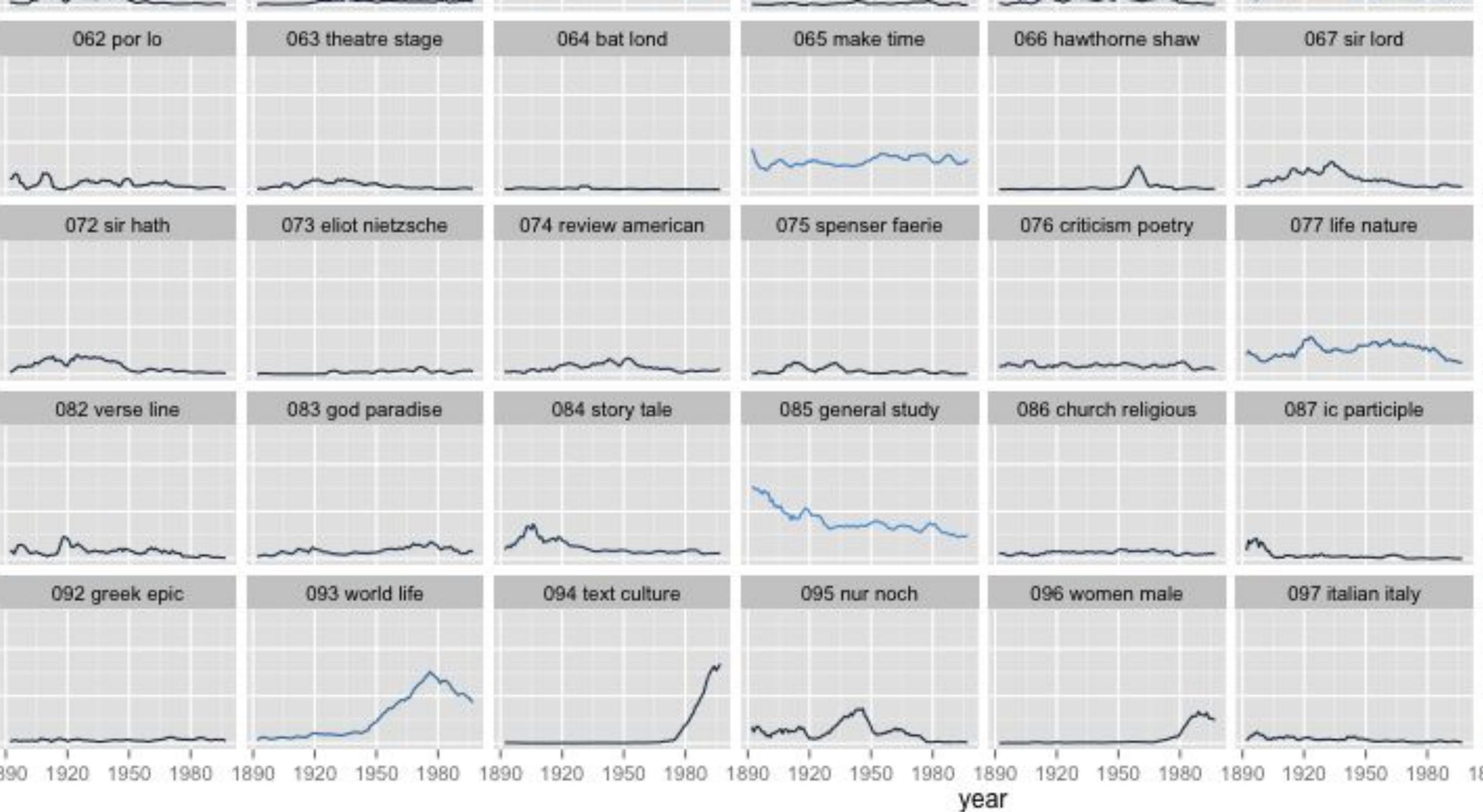
The screenshot shows the homepage of the **Journal of Digital Humanities**. At the top right is a search bar with the placeholder "Search JDH" and a magnifying glass icon. Below the search bar is a purple navigation bar with links for "About", "Volumes", "Submissions", and "Subscribe to the RSS". The main content area features a large, bold title: "What Can Topic Models of PMLA Teach Us About the History of Literary Scholarship?". Below the title, the authors "TED UNDERWOOD AND ANDREW GOLDSTONE" are listed. To the left of the main article, there is a sidebar with a "Table of Contents for Vol. 2, No. 1 Winter 2012" and several article titles and authors:

- Introductions
- Beginnings
- Applications and Critiques**
- Topic Modeling and Figurative Language  
Lisa M. Rhody
- Topic Model Data for Topic Modeling and Figurative Language  
Lisa M. Rhody

The main article's text begins with: "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 understanding of itself."



(Underwood  
and  
Goldstone,  
2012)

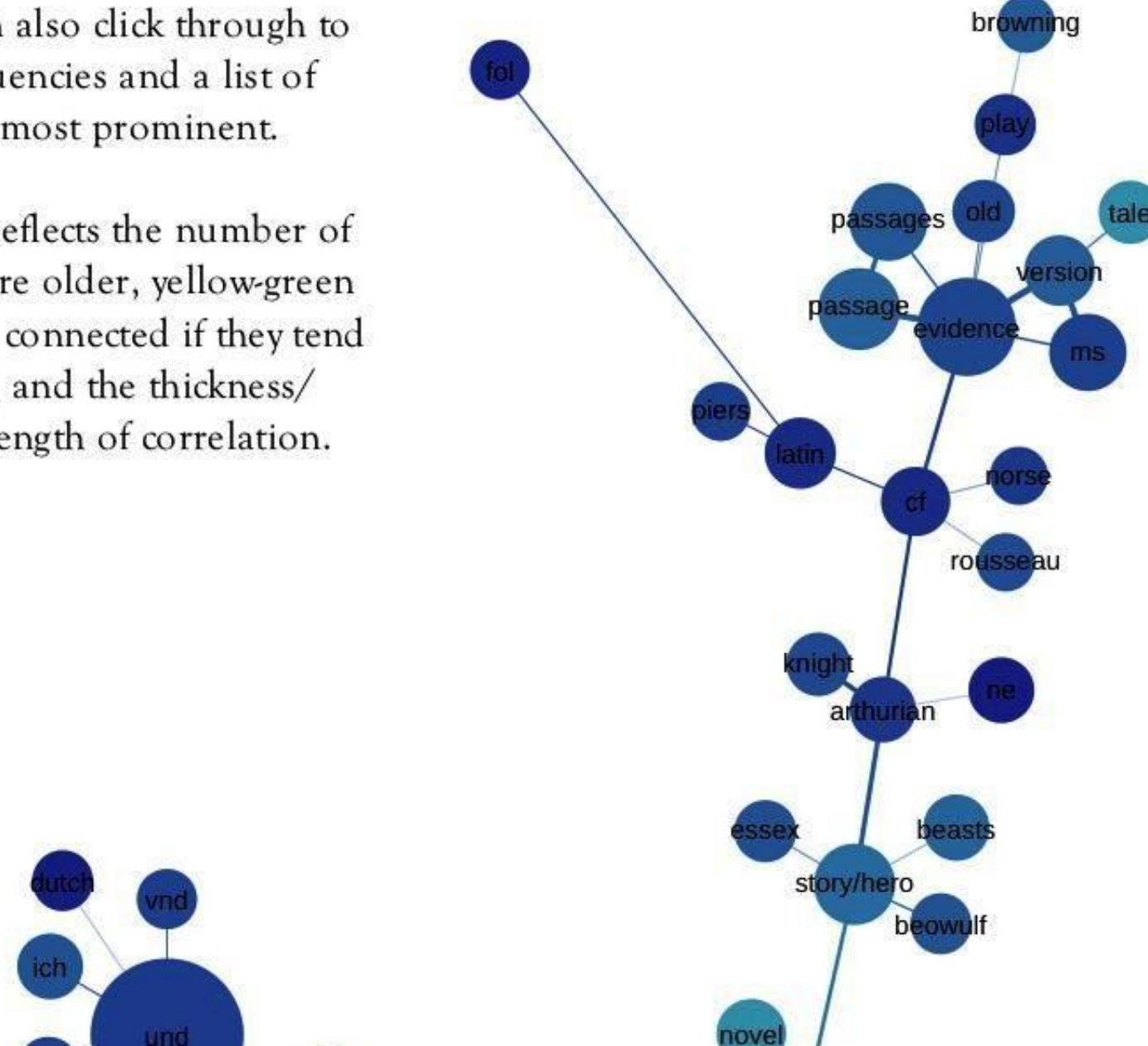


## Underwood's model of PMLA 1924-2006.

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.

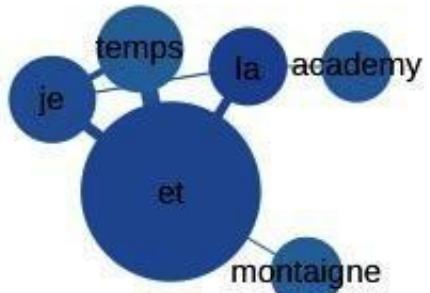
The size of each circle (loosely) reflects the number of words in the topic. Blue topics are older, yellow-green topics closer to 2006. Topics are connected if they tend to appear in the same articles, and the thickness/closeness of the link reflects strength of correlation.

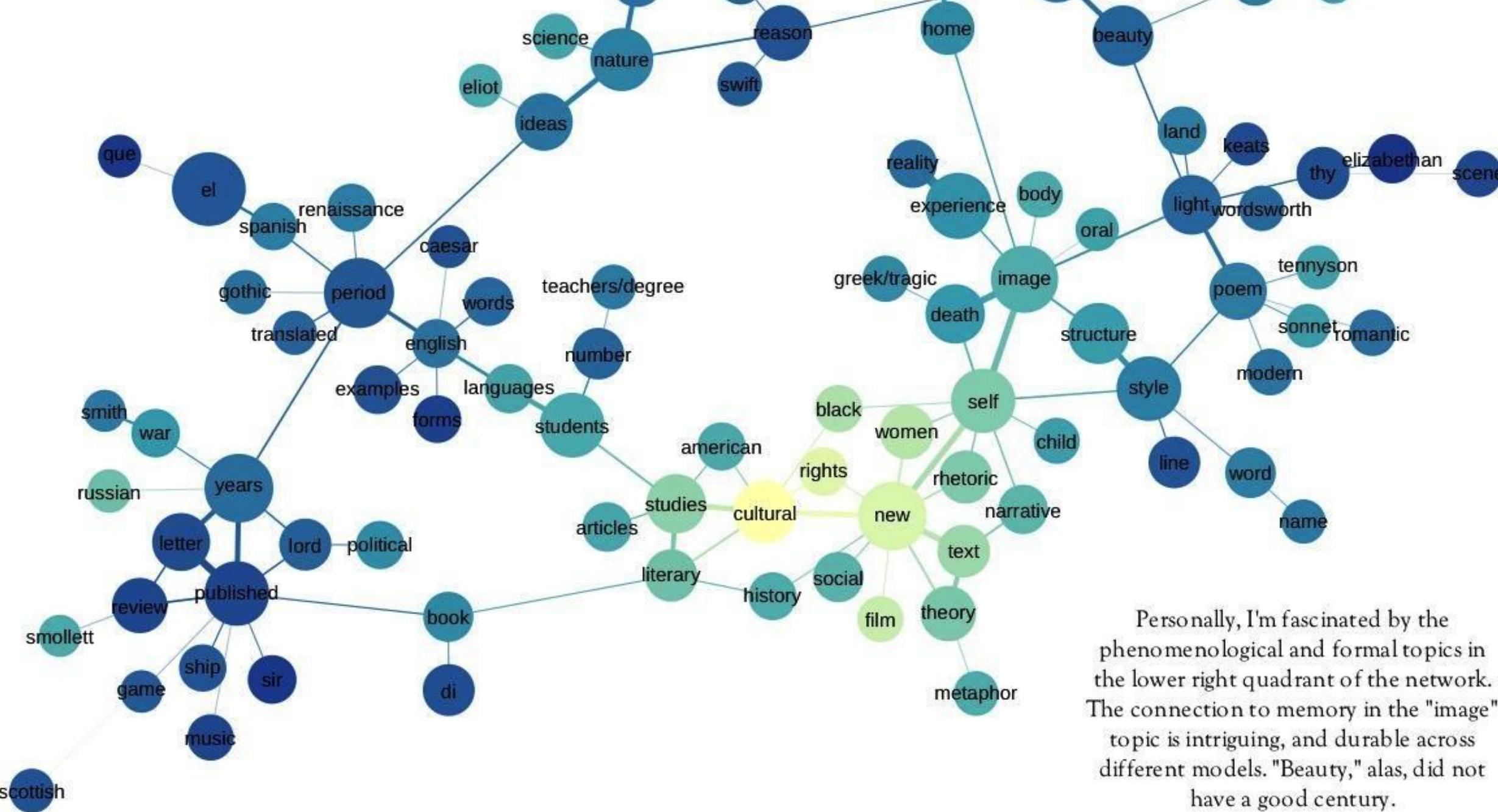
The window may have to be full width for click-through to work.



Very common English words are  
the, a, for, and, and, I,

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.





Personally, I'm fascinated by the phenomenological and formal topics in the lower right quadrant of the network. The connection to memory in the "image" topic is intriguing, and durable across different models. "Beauty," alas, did not have a good century.

# 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»?

“La discussione sulle possibili interpretazioni semiotico-letterarie della nozione di topic model e la constatazione della difficoltà teoriche che esse presentano ci porta ad affermare che in effetti **non è possibile trovare un unico e soddisfacente correlato teorico-letterario** dei risultati di questi metodi di analisi quantitativa”  
**(Ciotti, 2017)**