Topic Modeling

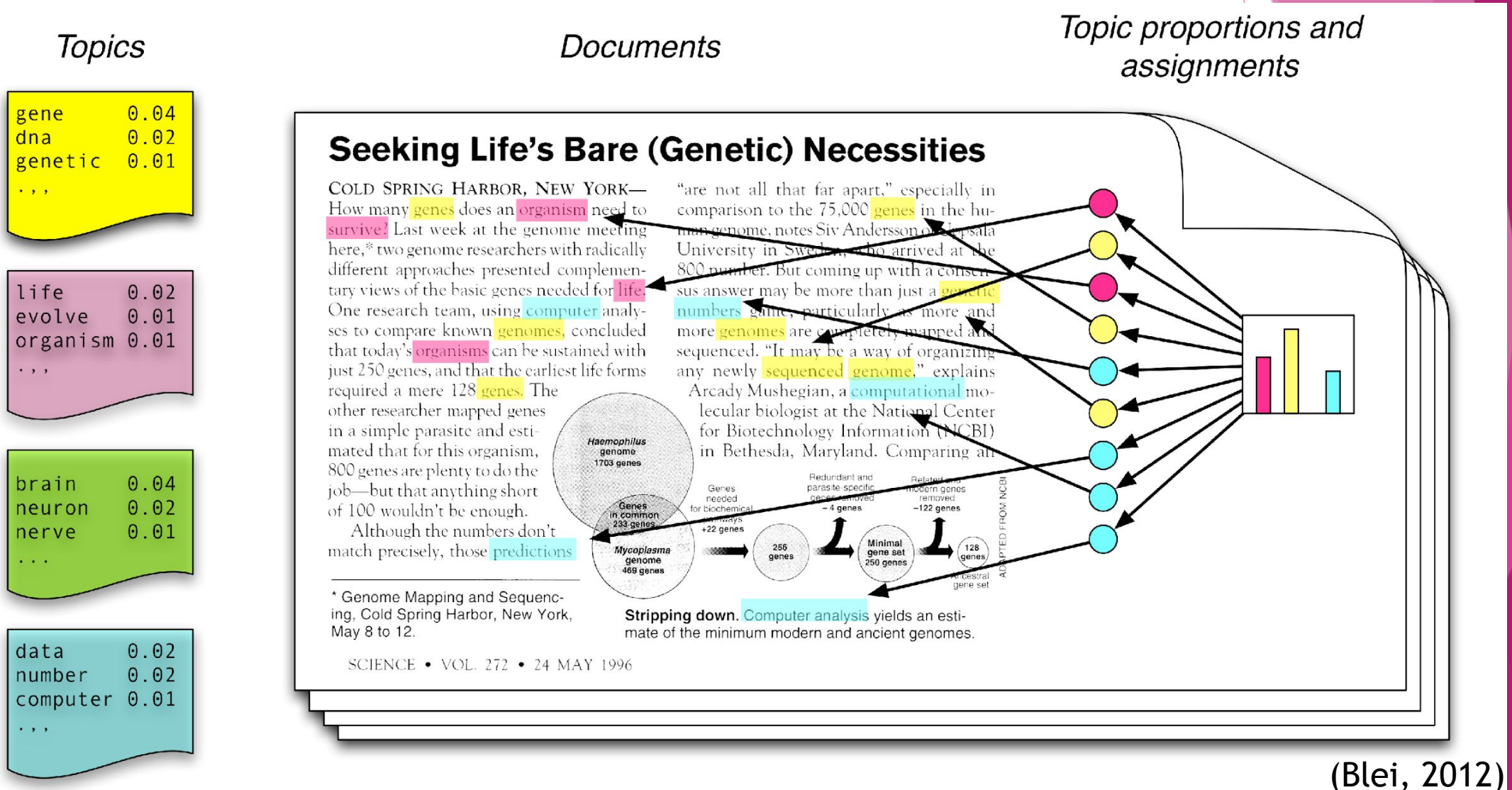
Topic Models

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



(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%

bad poetry

sentiments

very bad poetry

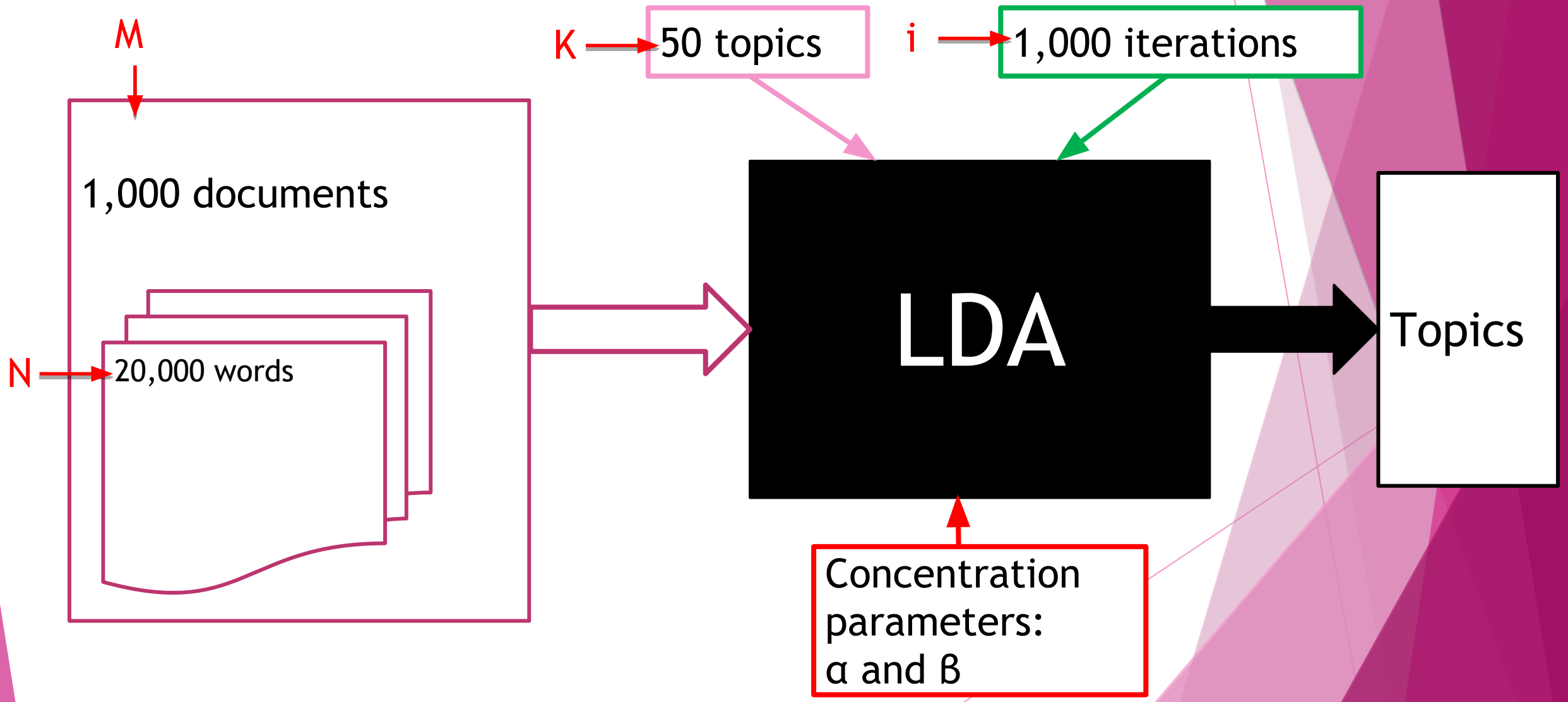
12.4%
5.2%
15.8%

a document:

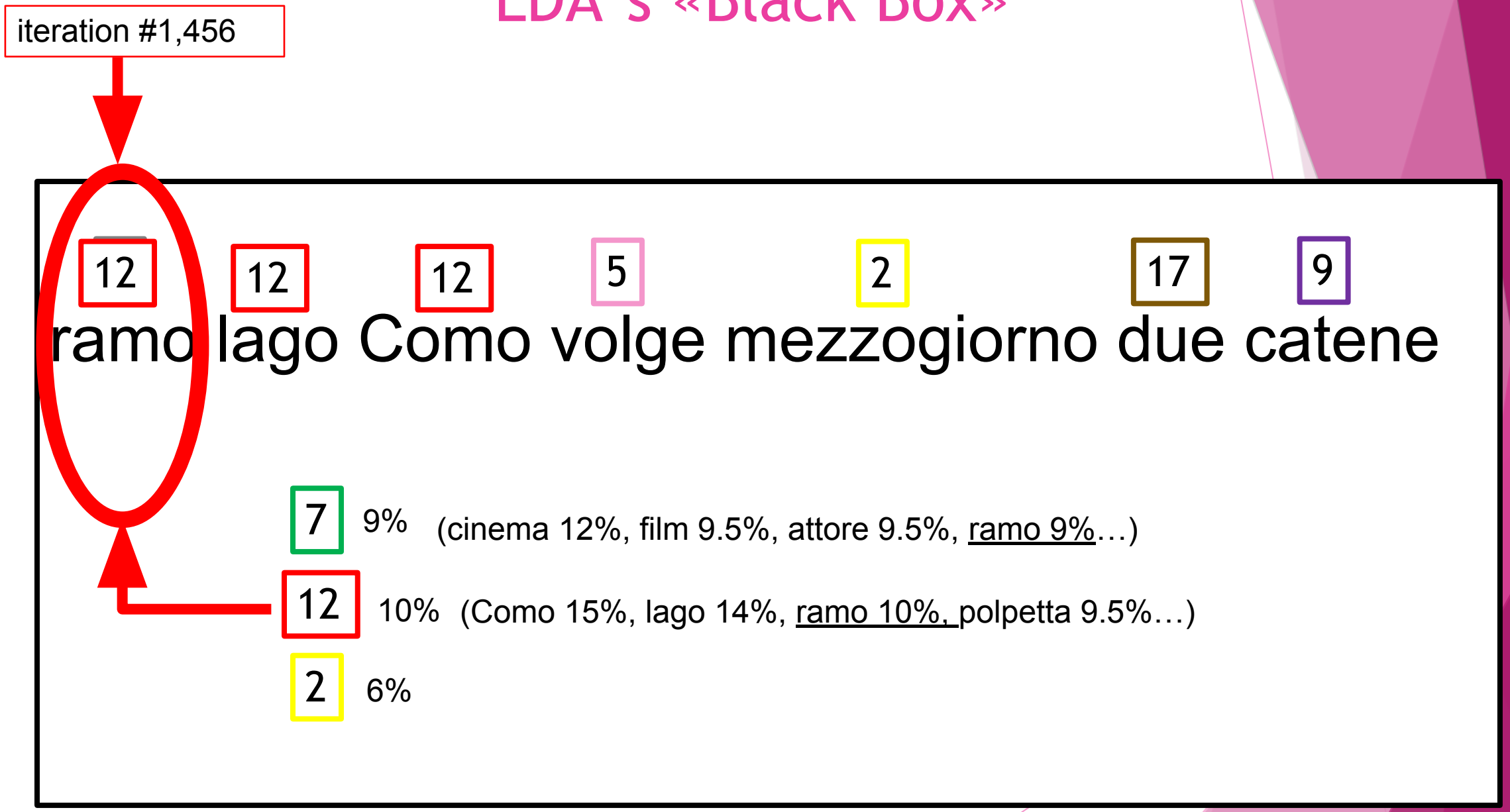
LDA: How Does it Work?

- 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»



TM Applications in Literature

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

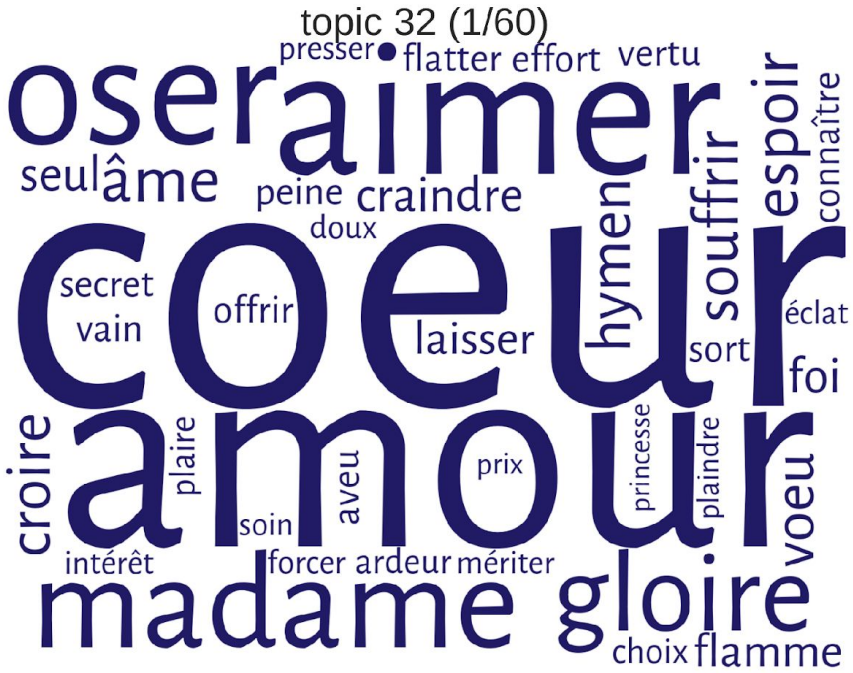
Abstract

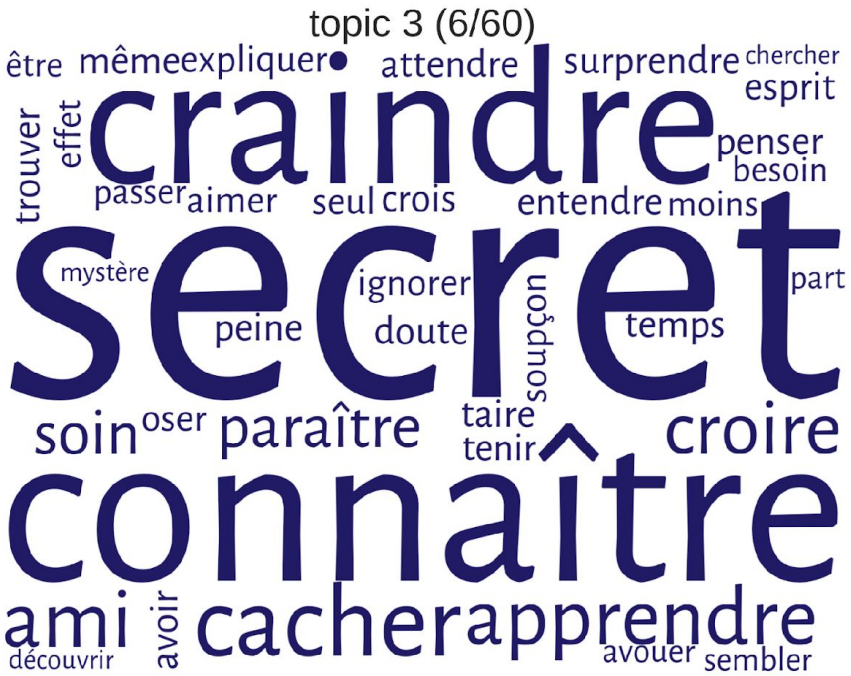
The concept of literary genre is a highly complex one: not only are different genres frequently defined on several, but not necessarily the same levels of description, but consideration of genres as cognitive, social, or scholarly constructs with a rich history further complicate the matter. This contribution 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.

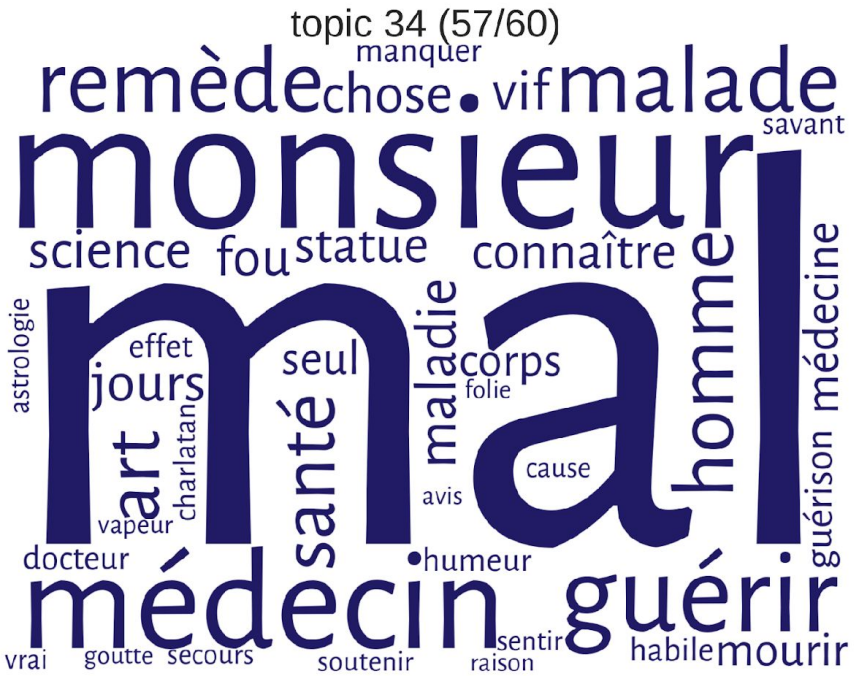
TM Applications in Literary Studies

"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)**

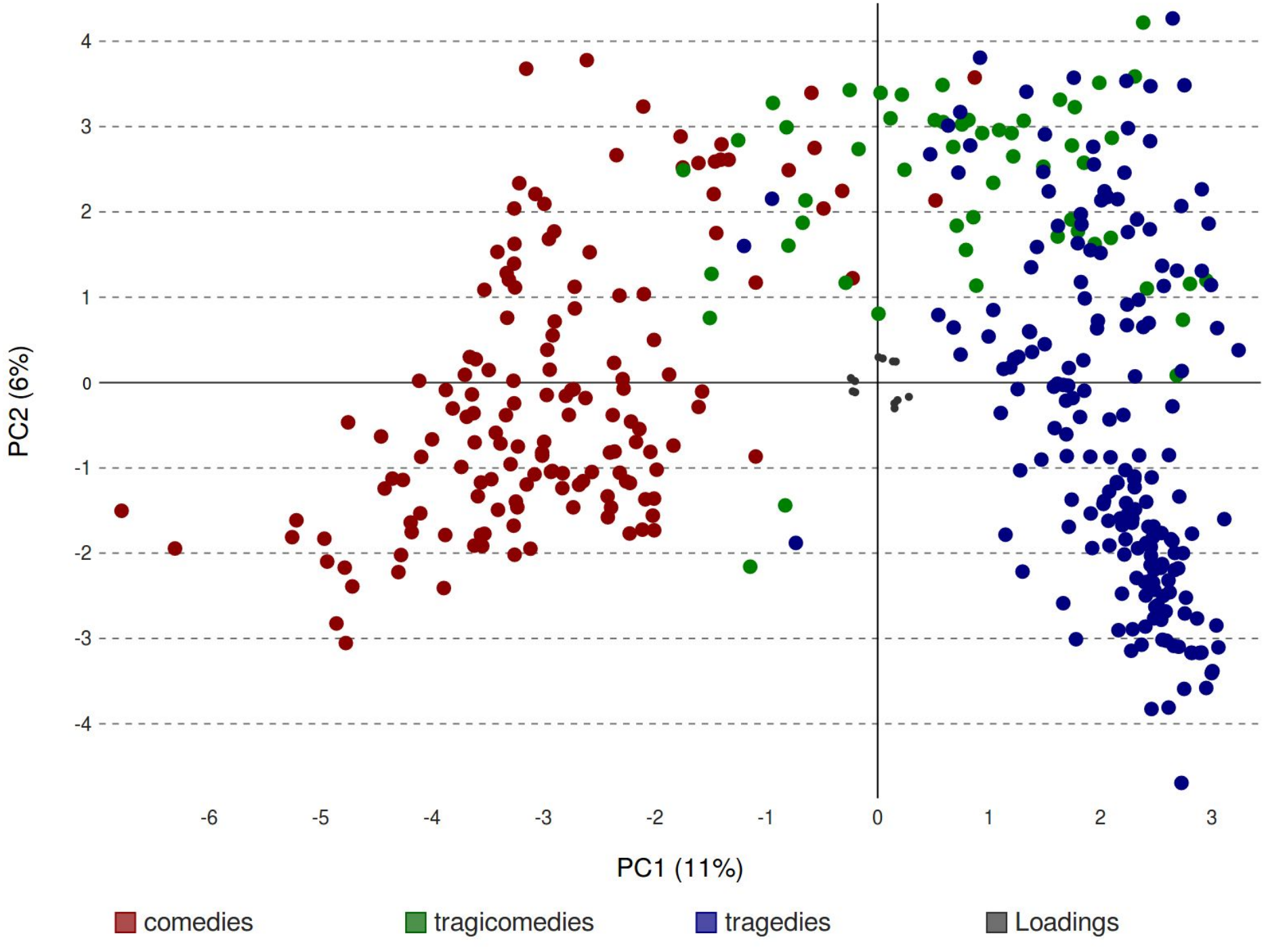

topic 32 (1/60)
oser
aimer
seul
âme
peine
craindre
doux
presser
flatter
effort
vertu
espoir
connaître
souffrir
hymen
coeur
secret
vain
offrir
laisser
éclat
sort
foi
amour
croire
plaire
soin
aveu
prix
princesse
plaindre
voeu
intérêt
forcer
ardeur
mériter
madame
gloire
choix
flamme


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


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

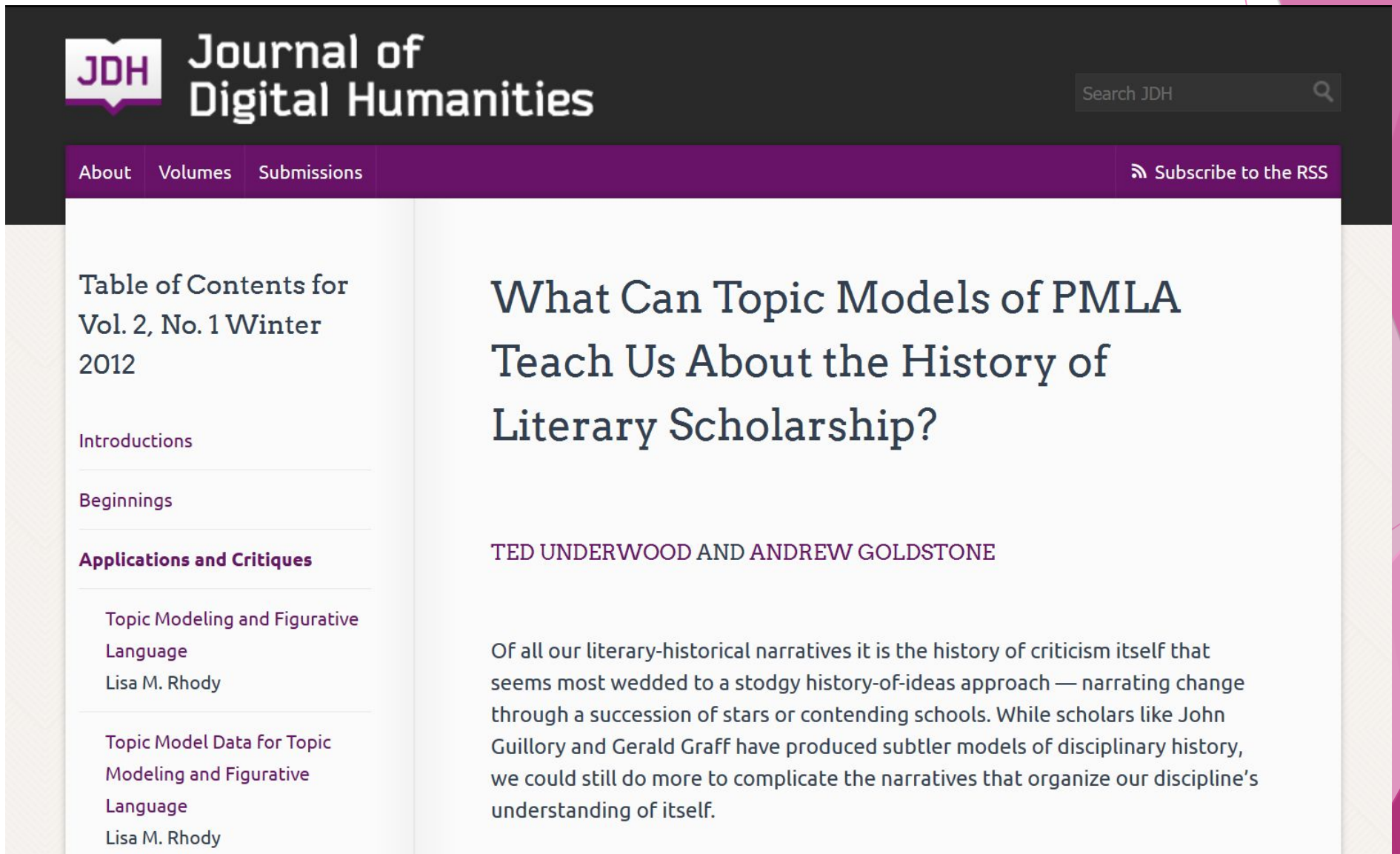

topic 34 (57/60)
remède
manquer
chose
vif
malade
savant
monsieur
science
fou
statue
connaître
mal
homme
médecine
astrologie
effet
seul
corps
jours
folie
maladie
art
charlatan
santé
cause
avis
guérison
vapeur
docteur
humeur
médecin
guérir
vrai
goutte
secours
soutenir
sentir
raison
habile
mourir

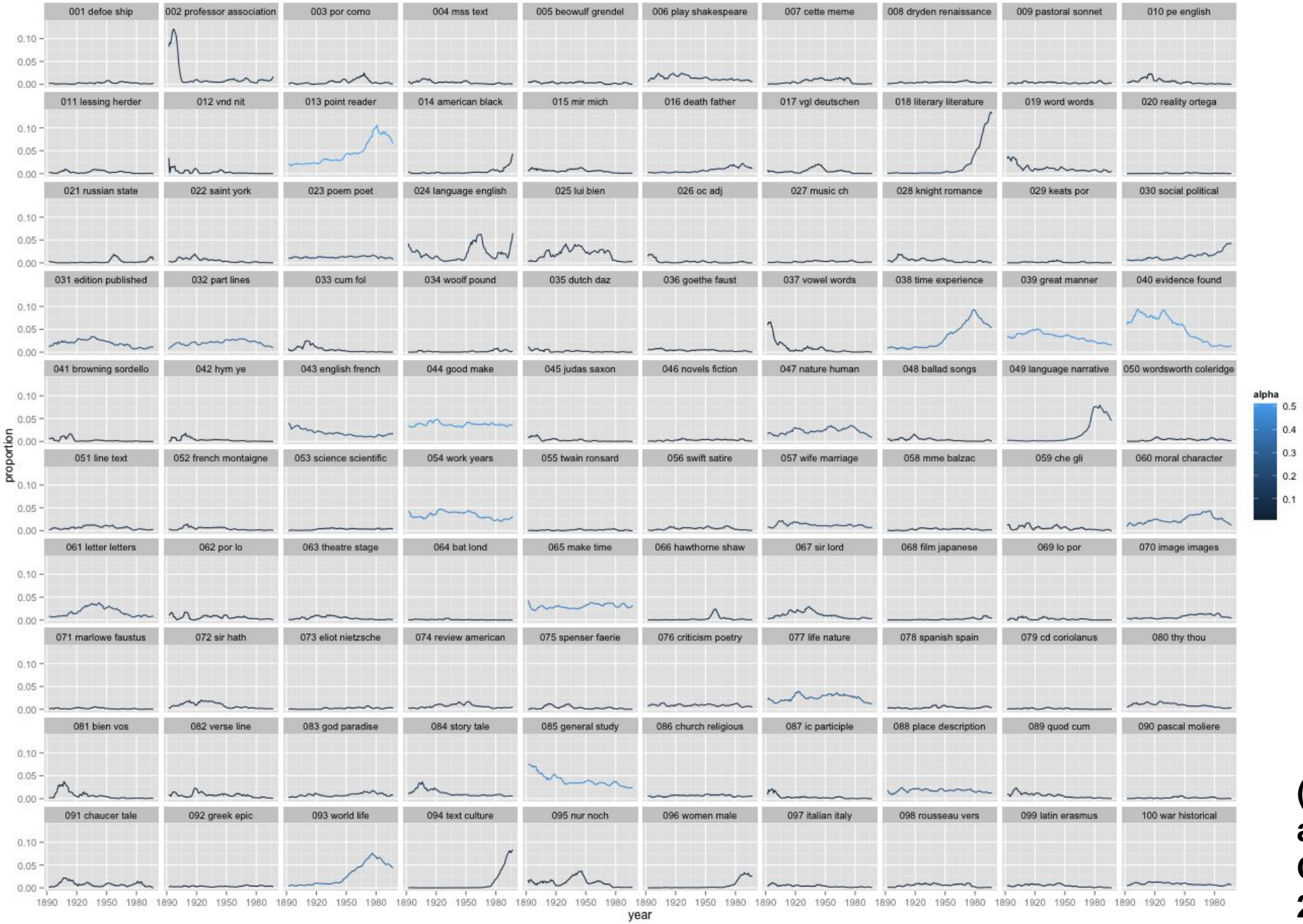
(Schöch, 2017)


4
3
2
1
0
-1
-2
-3
-4
PC2 (6%)
-6
-5
-4
-3
-2
-1
0
1
2
3
PC1 (11%)
comedies
tragicomedies
tragedies
Loadings

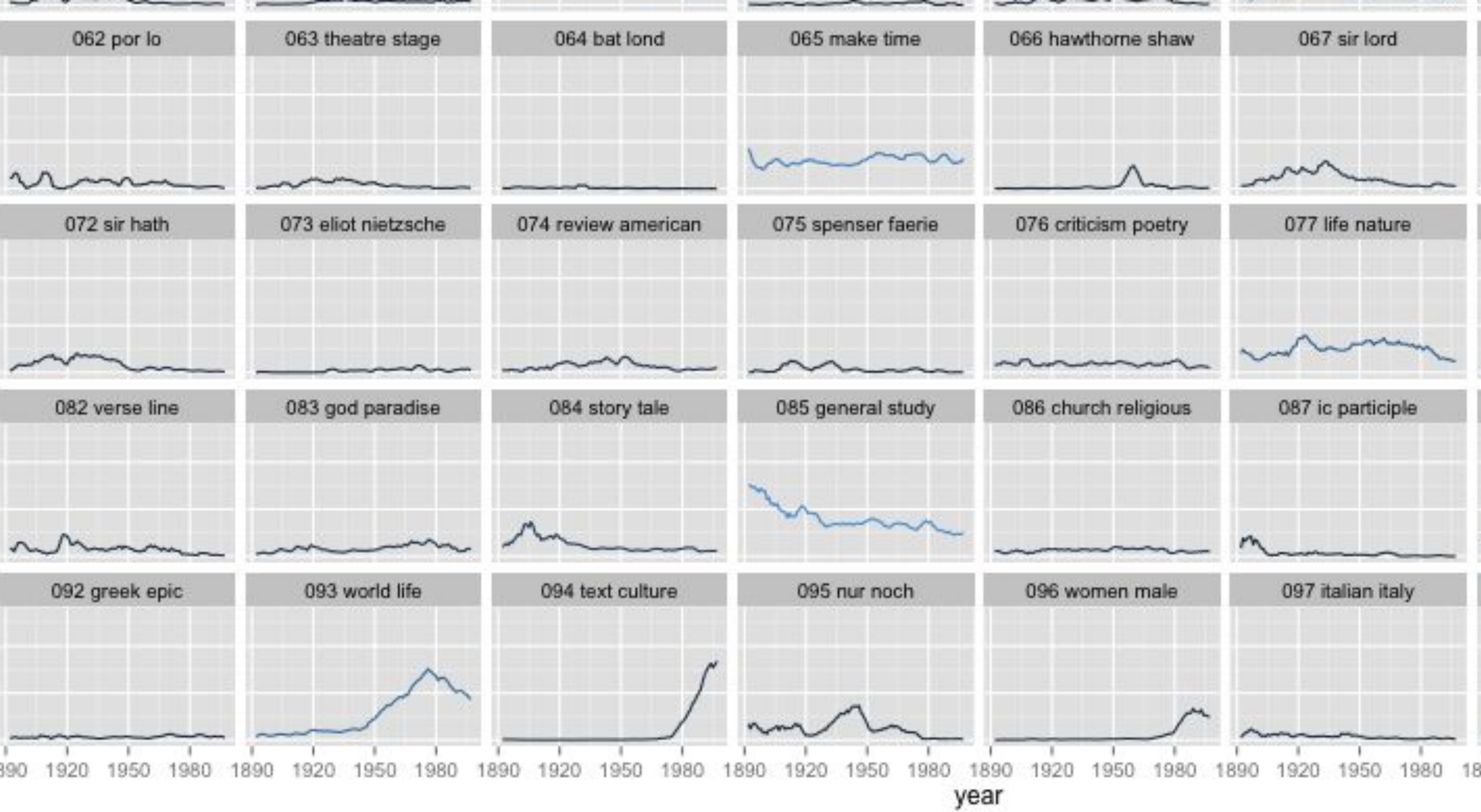
(Schöch, 2017)

TM Applications in Literary Studies





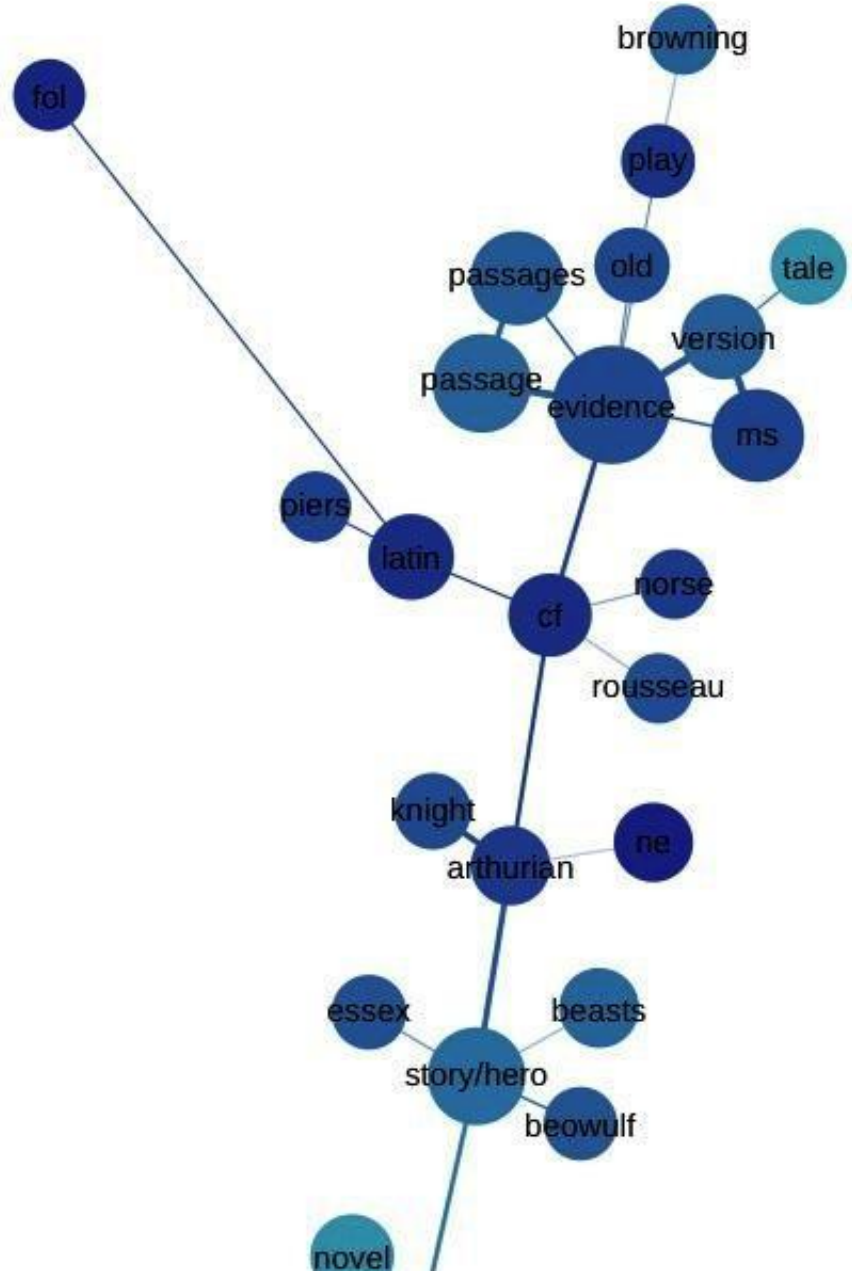
(Underwood and Goldstone, 2012)


062 por lo
063 theatre stage
064 bat lond
065 make time
066 hawthorne shaw
067 sir lord
072 sir hath
073 eliot nietzsche
074 review american
075 spenser faerie
076 criticism poetry
077 life nature
082 verse line
083 god paradise
084 story tale
085 general study
086 church religious
087 ic participle
092 greek epic
093 world life
094 text culture
095 nur noch
096 women male
097 italian italy
1890
1920
1950
1980
year

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.

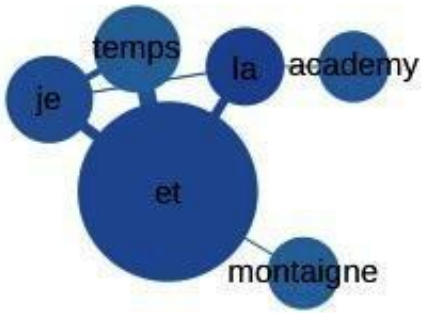
The size of each circle (loosely) reflects the number of words in the topic. Blue topics are older, yellow-green topics closer to 2006. Topics are connected if they tend to appear in the same articles, and the thickness/ closeness of the link reflects strength of correlation.

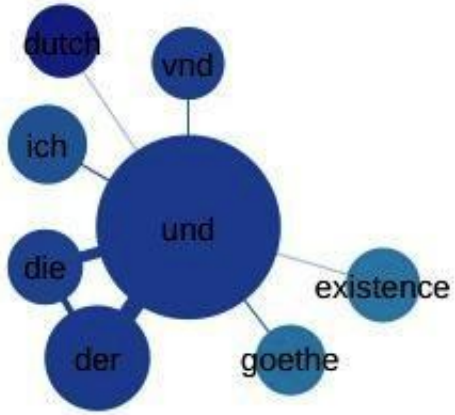
The window may have to be full width for click-through to work.

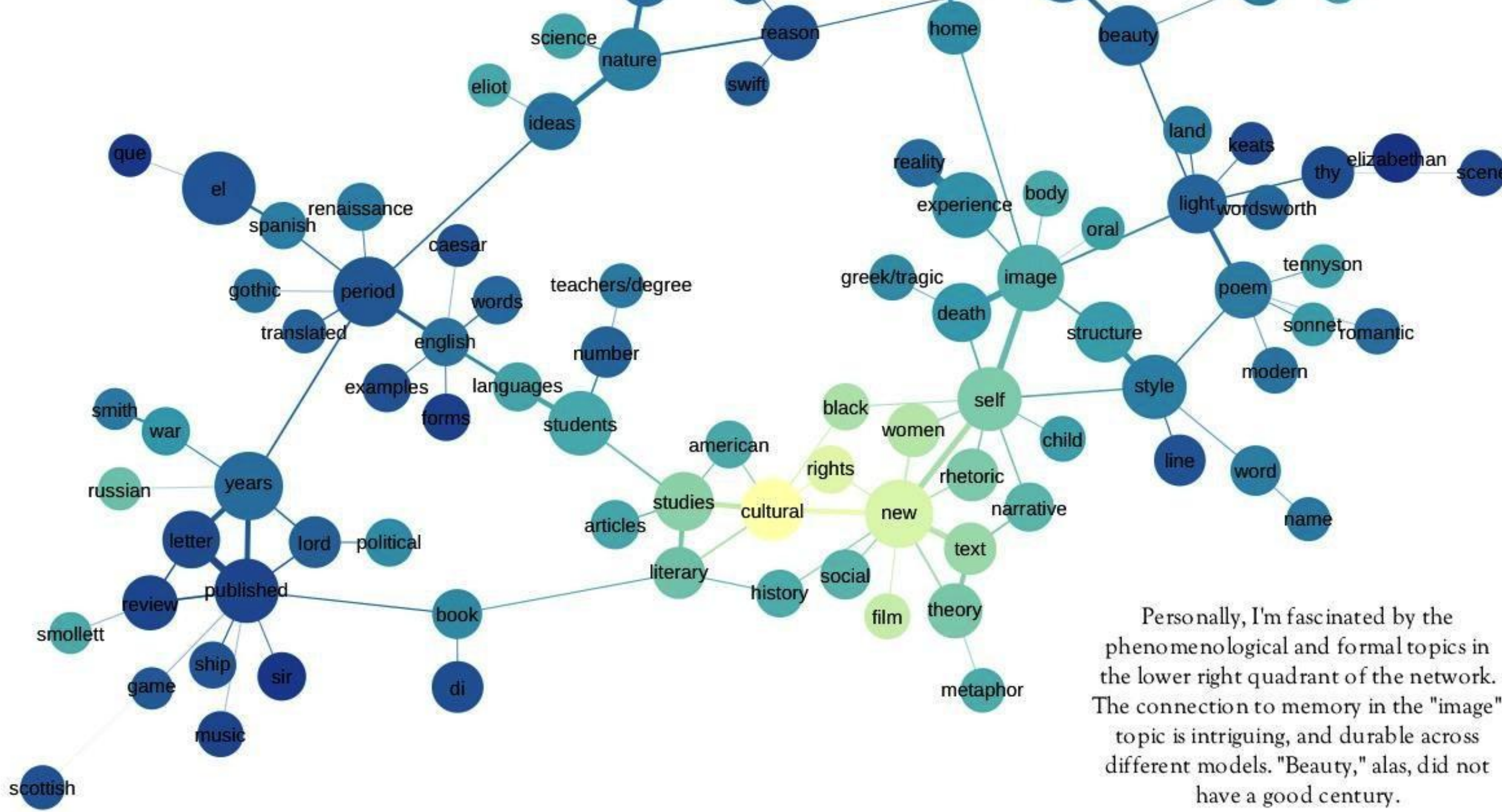


Very common English words are

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)