Distributional Semantics and Topic Modeling: Theory and Application

Baltic Summer School of Digital Humanities: Essentials of Coding and Encoding
Riga, July 2019

Christof Schöch
(Trier Center for Digital Humanities, Trier, Germany)

Universität Trier

Kompetenzzentrum
Trier Center For Digital Humanities (TCDH)

Overview

1. Introduction
2. Distributional Semantics: Principles and Methods
3. What are Word Embeddings?
4. What is Topic Modeling? Examples
5. Topic Models: the Theory
6. A Topic Modeling pipeline
7. First steps doing Topic Modeling
8. Advanced issues in Topic Modeling
9. Wrapping up

About this workshop

- Slides available online: https://christofs.github.io/riga/#/
- Download code and sample datasets: https://github.com/dh-trier/topicmodeling

About this workshop

- Context, examples, theory, demo, hands-on for Topic Modeling

---

- Slides available online: https://christofs.github.io/riga/#/
- Download code and sample datasets: https://github.com/dh-trier/topicmodeling

About this workshop

- Context, examples, theory, demo, hands-on for Topic Modeling
- Python-based, but not a Python workshop
  ("read and run" code, rather than write code)

- Slides available online: https://christofs.github.io/riga/#/
- Download code and sample datasets:
  https://github.com/dh-trier/topicmodeling

About this workshop

- Context, examples, theory, demo, hands-on for Topic Modeling
- Python-based, but not a Python workshop
  ("read and run" code, rather than write code)
- Learning goal: you understand how a Topic Model is created and can run your own Topic Modeling Pipeline

---

- Slides available online: https://christofs.github.io/riga/#/
- Download code and sample datasets:
  https://github.com/dh-trier/topicmodeling

About myself

- Professor of Digital Humanities
- Not a computer scientist, not a statistician
- French literary scholar by training
- Interests in corpus building and quantitative text analysis
- see: https://christof-schoech.de/en

About you: raise your hand if...

- ... you are a literary scholar

About you: raise your hand if...

- ... you are a literary scholar
- ... you are a historian

About you: raise your hand if...

- ... you are a literary scholar
- ... you are a historian
- ... you are a sociologist

About you: raise your hand if...

- ... you are a literary scholar
- ... you are a historian
- ... you are a sociologist
- ... you are a (computational / corpus) linguist

About you: raise your hand if...

- ... you are a literary scholar
- ... you are a historian
- ... you are a sociologist
- ... you are a (computational / corpus) linguist
- ... you are a computer scientist

About you: raise your hand if...

- ... you are a literary scholar
- ... you are a historian
- ... you are a sociologist
- ... you are a (computational / corpus) linguist
- ... you are a computer scientist
- ... you are a digital humanist

About you: raise your hand if...

- ... you are a literary scholar
- ... you are a historian
- ... you are a sociologist
- ... you are a (computational / corpus) linguist
- ... you are a computer scientist
- ... you are a digital humanist
- ... you are a librarian

About you: raise your hand if...

- ... you are a literary scholar
- ... you are a historian
- ... you are a sociologist
- ... you are a (computational / corpus) linguist
- ... you are a computer scientist
- ... you are a digital humanist
- ... you are a librarian
- ... you consider yourself to be a local

Distributional Semantics: Principles and Methods

Basic intuition about distributional semantics

- "Her friend's ..... was located on the second floor of the house."

Basic intuition about distributional semantics

- "Her friend's ..... was located on the second floor of the house."
- "apartment" !

Basic intuition about distributional semantics

- "Her friend's ..... was located on the second floor of the house."

- "apartment" !
- "room" !

Basic intuition about distributional semantics

- "Her friend's ..... was located on the second floor of the house."

- "apartment" !
- "room" !
- "balcony" ?

Basic intuition about distributional semantics

- "Her friend's ..... was located on the second floor of the house."

- "apartment" !
- "room" !
- "balcony" ?
- "cat" ??

Basic intuition about distributional semantics

- "Her friend's ..... was located on the second floor of the house."

- "apartment" !
- "room" !
- "balcony" ?
- "cat" ??
- "shark" ???

What does this example tell us?

- We are able to rank the likelihood of these words in the given context
- We use world knowledge, but also linguistic competency, for this
- Computers can learn this too, based on cooccurrence patterns

What does this example tell us?

- We are able to rank the likelihood of these words in the given context
- We use world knowledge, but also linguistic competency, for this
- Computers can learn this too, based on cooccurrence patterns
- That's how distributional semantics works!

Basic idea

- The meaning of words depends on their context
  "You shall know a word by the company it keeps" (Firth, 1957)

Basic idea

- The meaning of words depends on their context
  "You shall know a word by the company it keeps" (Firth, 1957)
- Words frequently appearing in similar contexts have similar meanings
- Words that can appear in very similar, specific contexts have similar grammatical functions

Two applications of this idea

- Topic Modeling
- Word Embeddings

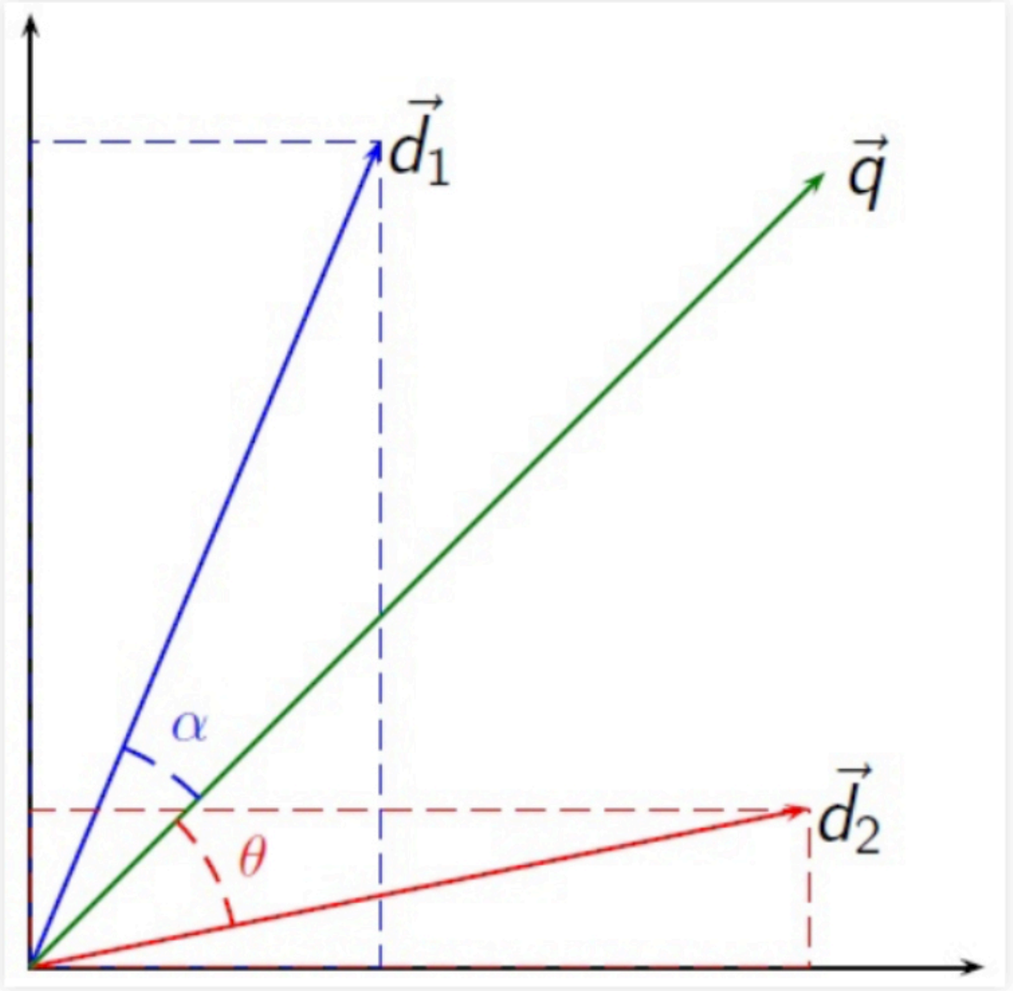
What are Word Embeddings?

Information Retrieval: Vector Space Model

Information Retrieval: Vector Space Model

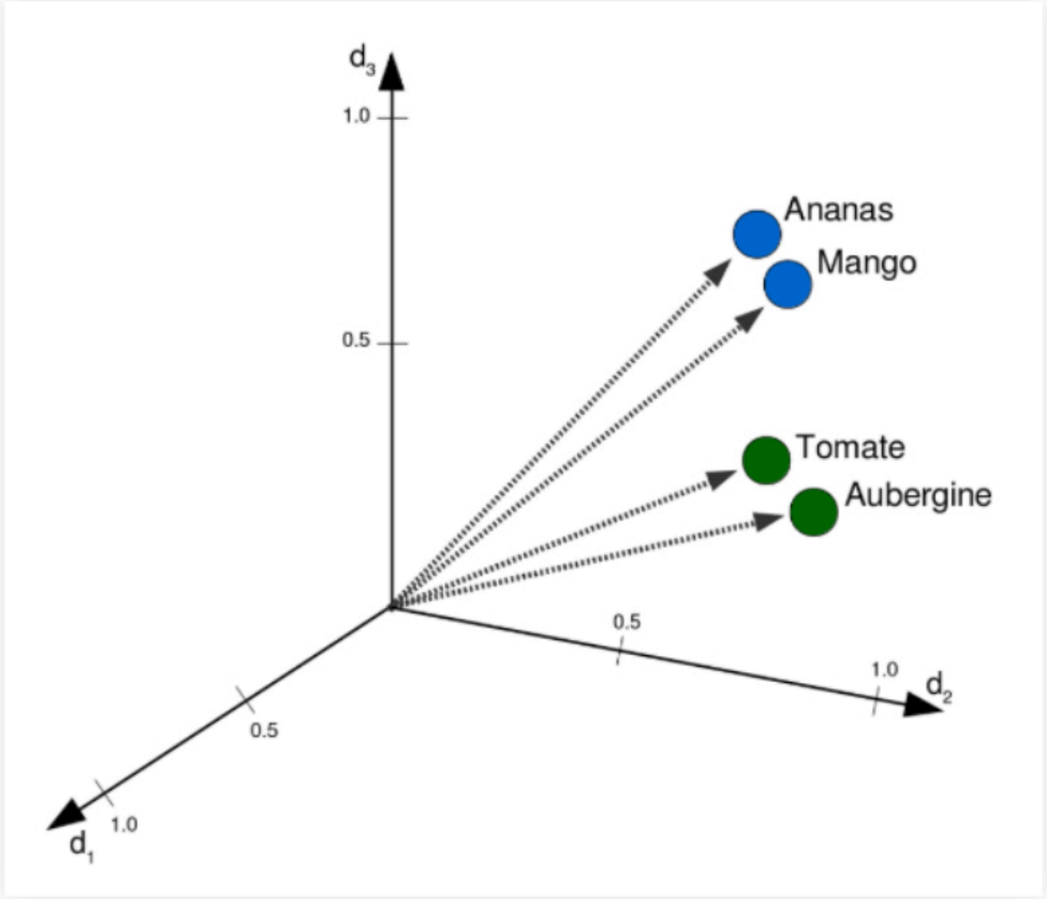
- Each document has a certain place in a vector space
- That place is determined by the keywords that appear in the document
- Each word is a dimension in the vector space
- Documents with shared vocabulary end up in the same area of the vector space

Information Retrieval: Vector Space Model



(Image Credit: Riclas, Wikipedia, Creative Commons Atttribution 3.0)

Words in vector space



(Artificial data. Image credit: Christof Schöch, 2019, Creative Commons Attribution 4.0 International)

Example: French Wikipedia Model

- 1.8 million articles, 750 million words
- transform term-document-matrix into dense matrix
- "low-dimensional", dense representation
- skip-gram model, 300 dimensions
- vector semantics: geometric relations = semantic relations

Similar Words Query

```
Query:   ['poésie_nom', 10]
Result:  poétique_adj      0.841
         poème_nom         0.790
         prose_nom         0.733
         littérature_nom   0.715
         poète_nom         0.704
         poétique_nom      0.701
         poésie_nam        0.700
         anthologie_nom    0.695
         littéraire_adj    0.655
         sonnet_nom        0.651
```

(authentic data, Wikipedia model)

Similarity Query

```
Query: ['prose_nom', 'littérature_nom']
Result: 0.511518681366

Query: ['poésie_nom', 'littérature_nom']
Result: 0.714615326722
```

(authentic data, Wikipedia model)

Evaluation

- Method: Using a "find-the-wrong word"-task

Evaluation

- Method: Using a "find-the-wrong word"-task
- Lists of similar words:
  - vert, bleu, jaune, rouge, orange
  - billet, monnaie, portemonnaie, payement

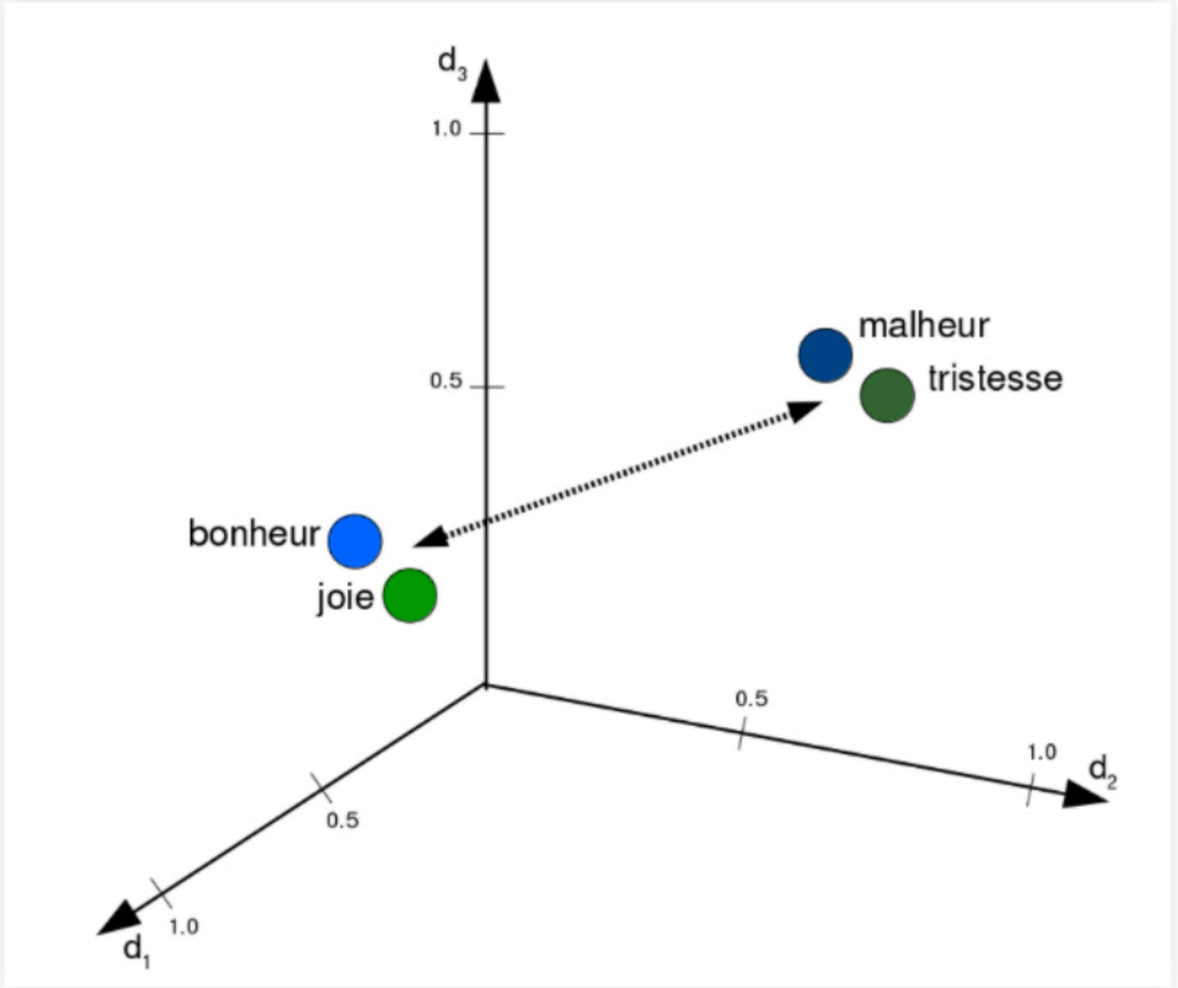
Evaluation

- Method: Using a "find-the-wrong word"-task
- Lists of similar words:
  - vert, bleu, jaune, rouge, orange
  - billet, monnaie, portemonnaie, payement
- Generate lists with an error
  - vert, bleu, monnaie, jaune, rouge

Evaluation

- Method: Using a "find-the-wrong word"-task
- Lists of similar words:
  - vert, bleu, jaune, rouge, orange
  - billet, monnaie, portemonnaie, payement
- Generate lists with an error
  - vert, bleu, monnaie, jaune, rouge
- Wikipedia model: 90% accuracy in finding the error

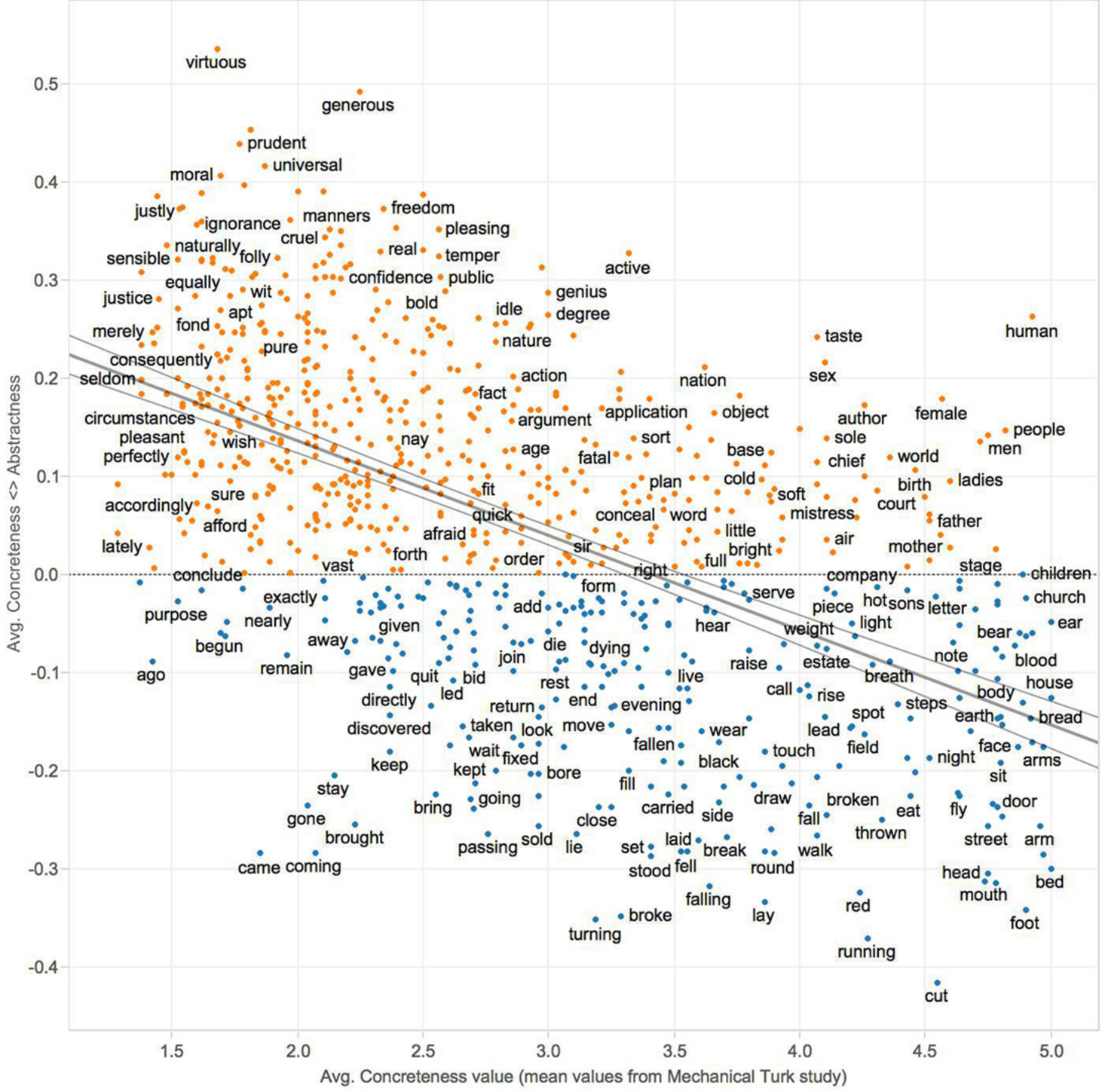
Axes of meaning



(Artificial data. Image credit: Christof Schöch, 2019, Creative Commons Attribution 4.0 International)

Axes of meaning (Ryan Heuser)

(Concrete vs. abstract. Image credit: Ryan Heuser, http://ryanheuser.org/word-vectors-3, 2015)


Avg. Concreteness <> Abstractness
Avg. Concreteness value (mean values from Mechanical Turk study)
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
virtuous
generous
prudent
universal
moral
justly
ignorance
manners
freedom
pleasing
cruel
naturally
folly
real
temper
sensible
confidence
public
active
equally
wit
genius
justice
apt
bold
idle
degree
merely
fond
nature
pure
consequently
seldom
action
fact
nation
taste
sex
human
circumstances
argument
application
object
author
female
pleasant
wish
nay
sort
sole
people
perfectly
age
fatal
base
chief
world
men
plan
cold
birth
ladies
sure
fit
soft
court
accordingly
quick
conceal
word
mistress
father
afford
afraid
little
air
lately
sir
bright
mother
forth
order
full
stage
vast
right
company
children
conclude
form
serve
exactly
add
piece
hot
sons
letter
church
purpose
nearly
given
hear
weight
light
ear
begun
away
die
dying
bear
join
raise
estate
note
blood
ago
remain
gave
quit
bid
rest
live
breath
house
directly
led
call
rise
body
return
end
evening
spot
steps
earth
bread
discovered
taken
look
move
wear
lead
fallen
touch
field
face
keep
wait
fixed
black
night
arms
kept
bore
sit
stay
fill
draw
broken
door
bring
going
carried
side
eat
fly
gone
close
fall
thrown
brought
passing
sold
lie
set
laid
break
walk
street
arm
came
coming
stood
fell
round
head
bed
falling
mouth
broke
lay
red
foot
turning
running
cut

Axis query

```
Axis: [["bonheur", "joie"],          # positive
       ["malheur", "tristesse"]]     # negative

Query:   ange
Result:  0.0875

Query:   monstre
Result   -0.1407
```

(authentic data)

Time for questions!

References

- Goldberg, Yoav, und Omer Levy. „word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method". arXiv.org, 2014. http://arxiv.org/abs/1402.3722.
- Heuser, Ryan. „Word Vectors in the Eighteenth Century". In Digital Humanities 2017: Conference Abstracts, 256–60. Montréal: McGill University & Université de Montréal, 2017.
- Mikolov, Tomas, Kai Chen, Greg Corrado, und Jeffrey Dean. „Efficient Estimation of Word Representations in Vector Space". arXiv.org, 2013. http://arxiv.org/abs/1301.3781.
- Pennington, Jeffrey, Richard Socher, und Christopher D. Manning. „Glove: Global vectors for word representation", 2014. doi:10.1.1.671.1743.
- Turney, Peter T., und Patrick Pantel. „From Frequency to Meaning: Vector Space Models of Semantics". Journal of Artificial Intelligence Research 37 (2010): 141–88. https://arxiv.org/abs/1003.1141.
- Widdows, Dominic. *Geometry and meaning*. CSLI lecture notes, no. 172. Stanford CA: CSLI Publications, 2004.

What is Topic Modeling?

(a) Some fundamentals

Topic Modeling: basic idea

- Works on the basis of (large) collections of documents

Topic Modeling: basic idea

- Works on the basis of (large) collections of documents
- Each document is understood as a mixture of topics

Topic Modeling: basic idea

- Works on the basis of (large) collections of documents
- Each document is understood as a mixture of topics
- The purpose is to discover thematic trends and patterns

Topic Modeling: basic idea

- Works on the basis of (large) collections of documents
- Each document is understood as a mixture of topics
- The purpose is to discover thematic trends and patterns
- Discovered through generative probabilistic modeling

Usage scenarios

Usage scenarios

- Information Retrieval: Search not for individual terms, but themes / semantic fields

Usage scenarios

- Information Retrieval: Search not for individual terms, but themes / semantic fields
- Recommender Systems: Recommend similar journal articles etc. to users

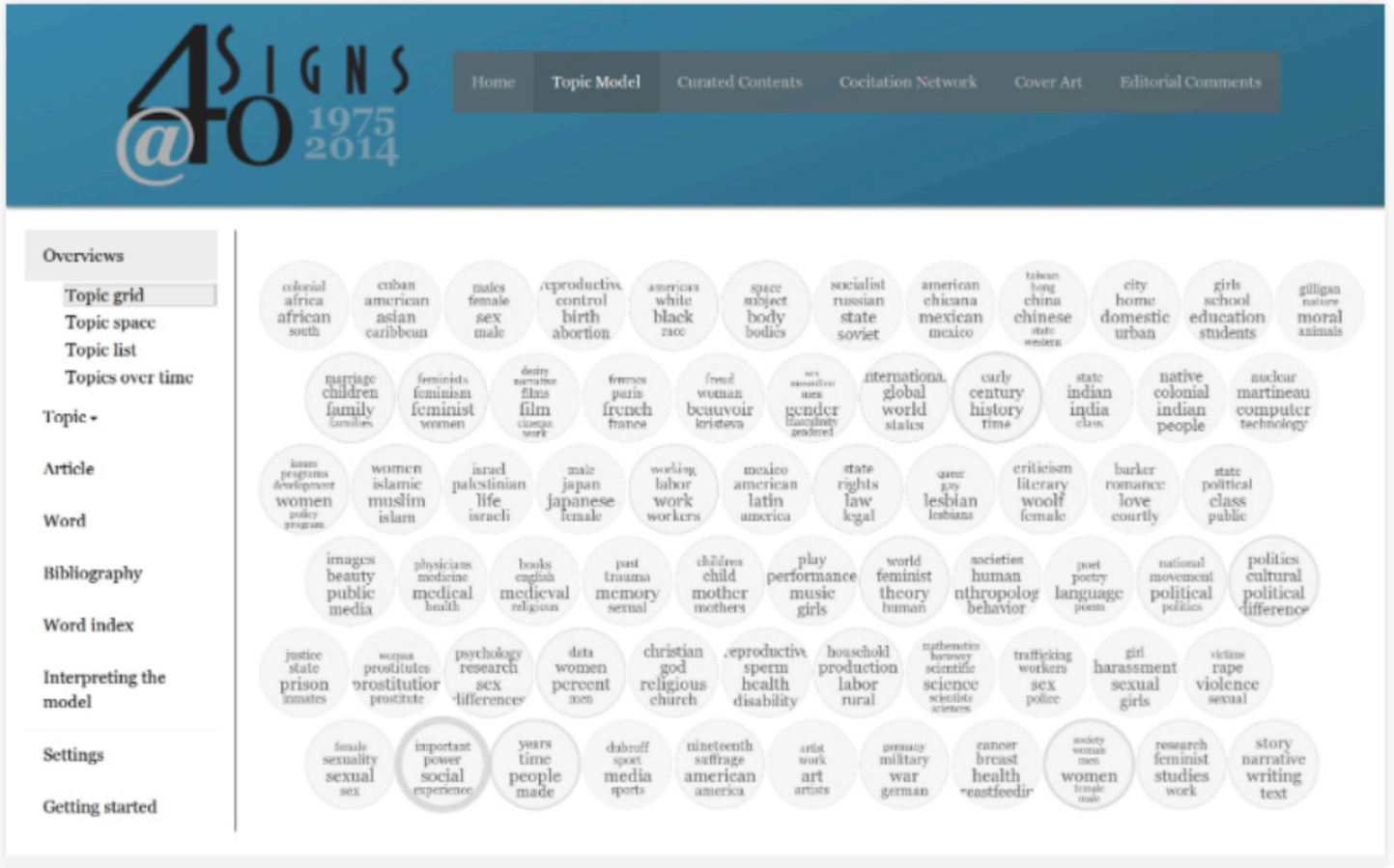
Usage scenarios

- Information Retrieval: Search not for individual terms, but themes / semantic fields
- Recommender Systems: Recommend similar journal articles etc. to users
- Exploration of text collections: what is an email or newspaper corpus about?

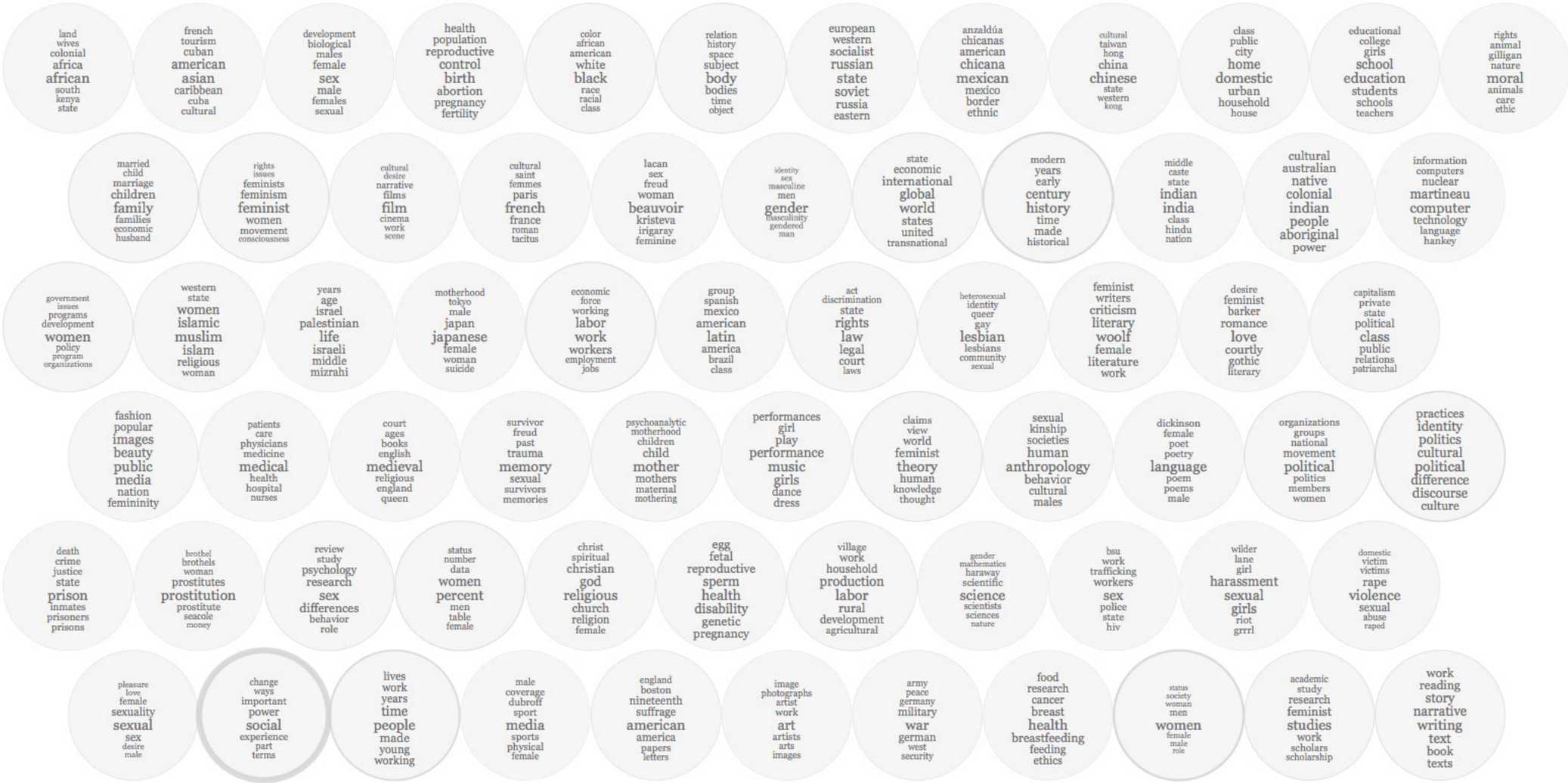
Usage scenarios

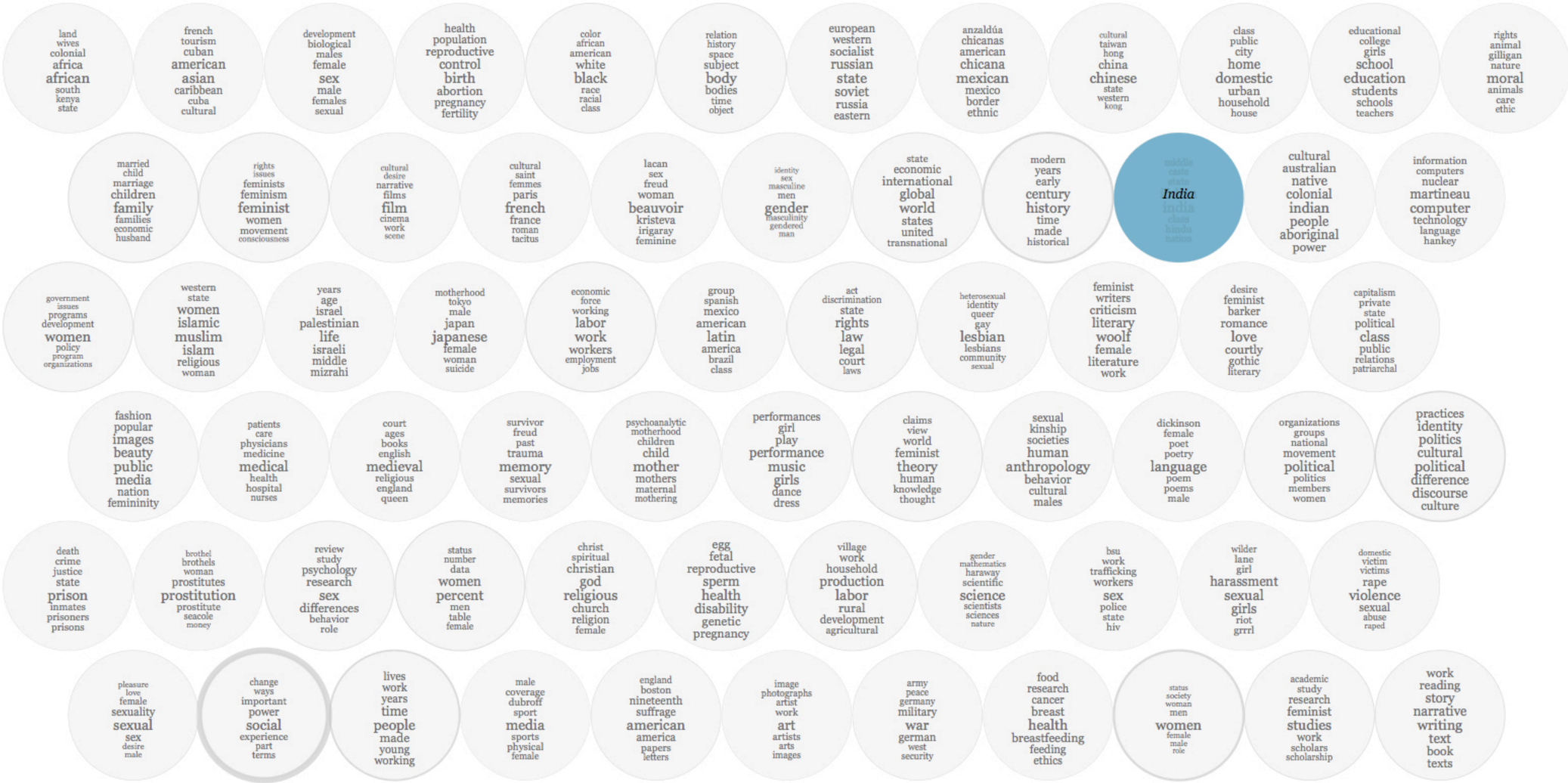
- Information Retrieval: Search not for individual terms, but themes / semantic fields
- Recommender Systems: Recommend similar journal articles etc. to users
- Exploration of text collections: what is an email or newspaper corpus about?
- Research questions from literary studies, cultural studies, history of ideas: topics across authors, genres, time periods

Explorative Visualization



Signs at 40: http://signsat40.signsjournal.org/topic-model/#/model/grid





Overviews
Topic
Article
Word
Bibliography
Word index
Interpreting the model
Settings
Getting started

India

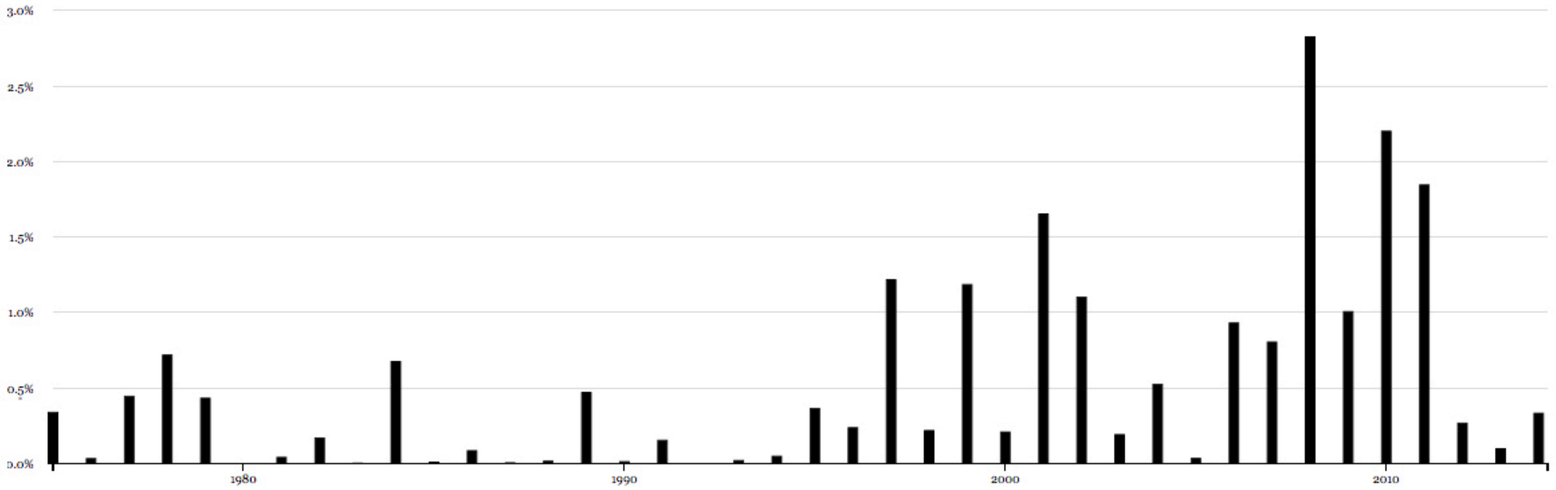
Top words

Word	Weight
india	
indian	
class	
state	
hindu	
caste	
nation	
middle	
village	
south	
english	
spivak	
delhi	
colonial	
devi	
world	
postcolonial	
kamasutra	
modernity	
western	
modern	
national	
west	
british	
tradition	
phoolan	
radha	
bengal	
literacy	

Yearly proportion of words in topic

clear selected year

0.5% of corpus in total. *Click a bar to limit articles to that year.*



Top articles

Article	%	Tokens
Goswami, Namita. "Autophagia and Queer Transnationality: Compulsory Heteroimperial Masculinity in Deepa Mehta's *Fire*." *Signs* 33, no. 2 (Winter 2008): 343–369.	37.0%	1836
Puri, Jyoti. "Concerning *Kamasutras*: Challenging Narratives of History and Sexuality." *Signs* 27, no. 3 (Spring 2002): 603–639.	33.3%	2293
Patil, Vrushali. "Reproducing-Resisting Race and Gender Difference: Examining India's Online Tourism Campaign from a Transnational Feminist Perspective." *Signs* 37, no. 1 (Autumn 2011): 185–210.	32.9%	1369
Bhatt, Amy, Madhavi Murty, and Priti Ramamurthy. "Hegemonic Developments: The New Indian Middle Class, Gendered Subalterns, and Diasporic Returnees in the Event of Neoliberalism." *Signs* 36, no. 1 (Autumn 2010): 127–152.	31.9%	1451
Fernandes, Leela. "Reading 'India's Bandit Queen': A Trans/national Feminist Perspective on the Discrepancies of Representation." *Signs* 25, no. 1 (Autumn 1999): 123–152.	30.4%	1968
Rajan, Gita. "Constructing–Contesting Masculinities: Trends in South Asian Cinema." *Signs* 31, no. 4 (Summer 2006): 1099–1124.	28.6%	1456
Oza, Rupal. "Showcasing India: Gender, Geography, and Globalization." *Signs* 26, no. 4 (Summer 2001): 1067–1095.	28.2%	1603

Overviews
Topic
Article
Word
Bibliography
Word index
Interpreting the model
Settings
Getting started

Goswami, Namita. "Autophagia and Queer Transnationality: Compulsory Heteroimperial Masculinity in Deepa Mehta's *Fire*." *Signs* 33, no. 2 (Winter 2008): 343–369.

4068 tokens. (view on JSTOR)

Topic	%	Tokens
India: india indian class state hindu caste nation middle village south english spivak delhi colonial devi	45.1%	1836
Bodies: body subject bodies space time history object relation place desire butler logic future symbolic bodily	11.1%	450
The social: social power experience important part ways terms change process based life sense order individual form	10.6%	431
Film: film films cinema narrative work desire scene cultural images visual fantasy image theory documentary identification	7.5%	305
Gender theory: gender men masculinity masculine gendered sex man identity male feminine femininity biological trans natural masculinities	5.2%	211
Media images: public beauty media images nation popular femininity fashion image feminine figure consumer magazine american sphere	4.5%	185
Lesbian, gay, queer: lesbian gay lesbians queer community identity sexual heterosexual lesbianism relationships heterosexuality sexuality class theory homosexuality	3.4%	140
Women's roles: women men female woman male society role status culture traditional roles recent position world lives	3.3%	135
China: chinese china state hong western taiwan kong cultural li yu party young foot wang revolution	1.8%	74
Globalization: world global states international united economic transnational state rights countries local globalization human national gendered	1.8%	74
Asian American / Caribbean: asian american caribbean cuban cuba tourism cultural french white immigrant filipino sheldon identity colonial tourists	1.5%	62
Sexuality: sexual sexuality sex female desire love male pleasure marriage pornography erotic behavior homosexuality heterosexual relationship	1.1%	45
Family, poverty, welfare: family children families marriage economic child husband married wife support mothers home poor care welfare	1.1%	43
Performance: music performance girls play dance girl dress performances men theater queen clothing audience songs boys	0.7%	27
Law: law rights legal state court discrimination laws act equality legislation courts case civil protection equal	0.4%	17
Political movements: political movement politics national members groups women organizations public activists leaders party government organization union	0.4%	17
Rural economies: labor production rural household development work agricultural village land women farm agriculture economic households market	0.2%	10
Feminist movements: feminist feminism women feminists movement issues consciousness rights political movements radical collective liberation struggle issue	0.1%	6

Overviews
Topic
Article
Word
Bibliography
Word index
Interpreting the model
Settings
Getting started

Film

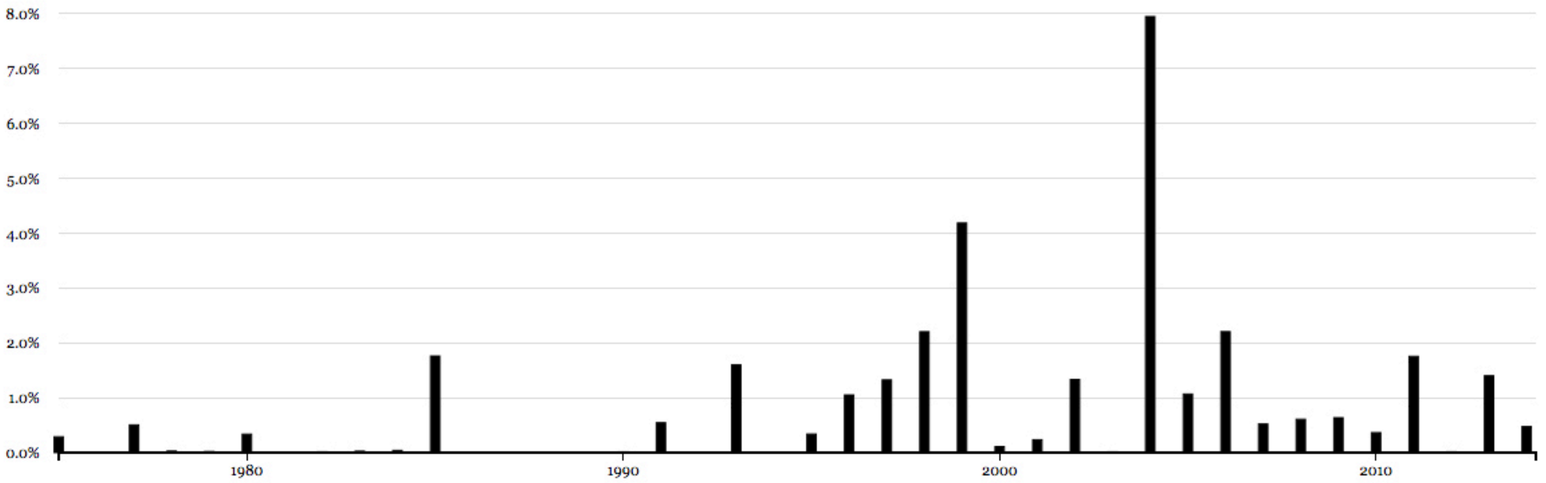
Top words

Word	Weight
film	
films	
cinema	
narrative	
work	
desire	
scene	
cultural	
images	
visual	
fantasy	
image	
theory	
documentary	
identification	
camera	
hollywood	
screen	
cinematic	
gaze	
media	
malcolm	
pleasure	

Yearly proportion of words in topic

clear selected year

0.9% of corpus in total. *Click a bar to limit articles to that year.*



Top articles

Article	%	Tokens
Gaines, Jane M. "First Fictions." *Signs* 30, no. 1 (Autumn 2004): 1293–1317.	46.1%	1751
Lauretis, Teresa de. "Popular Culture, Public and Private Fantasies: Femininity and Fetishism in David Cronenberg's 'M. Butterfly.'" *Signs* 24, no. 2 (Winter 1999): 303–334.	46.0%	3045
Mottahedeh, Negar. "'Life Is Color!' Toward a Transnational Feminist Analysis of Mohsen Makhmalbaf's *Gabbeh*." *Signs* 30, no. 1 (Autumn 2004): 1403–1424.	44.6%	1909
Mayne, Judith. "Marlene, Dolls, and Fetishism." *Signs* 30, no. 1 (Autumn 2004): 1257–000.	43.8%	565
Stevens, Maurice E. "Dis/Identification: Subject to Countermemory: Disavowal and Black Manhood in Spike Lee's *Malcolm X*." *Signs* 28, no. 1 (Autumn 2002): 277–301.	42.4%	1966
Mayne, Judith. "Feminist Film Theory and Criticism." *Signs* 11, no. 1 (Autumn 1985): 81–100.	41.7%	1749

malcolm
pleasure
video
characters
viewer
audience
shot
western
popular
culture
song
goldman
representation
spectator
butterfly
movies
girl
reading
contemporary
vision
filmmakers
role
movie
mulvey
narratives
view
benning
character
scenes

Manhood in Spike Lee's *Malcolm X*." *Signs* 28, no. 1 (Autumn 2002): 277–301.

Mayne, Judith. "Feminist Film Theory and Criticism." *Signs* 11, no. 1 (Autumn 1985): 81–100. 41.7% 1749

Johnson, Liza. "Perverse Angle: Feminist Film, Queer Film, Shame." *Signs* 30, no. 1 (Autumn 2004): 1361–1384. 40.8% 1681

McCormick, Richard W. "From 'Caligari' to Dietrich: Sexual, Social, and Cinematic Discourses in Weimar Film." *Signs* 18, no. 3 (Spring 1993): 640–668. 40.2% 2309

Fleetwood, Nicole R. "Visible Seams: Gender, Race, Technology, and the Media Art of Fatimah Tuggar." *Signs* 30, no. 1 (Autumn 2004): 1429–1452. 36.7% 1903

Jayamanne, Laleen. "Pursuing Micromovements in Room 202." *Signs* 30, no. 1 (Autumn 2004): 1248–000. 35.7% 506

Ryan, Judylyn S. "Outing the Black Feminist Filmmaker in Julie Dash's *Illusions*." *Signs* 30, no. 1 (Autumn 2004): 1319–1344. 35.7% 1641

Brody, Jennifer DeVere. "The Returns of 'Cleopatra Jones.'" *Signs* 25, no. 1 (Autumn 1999): 91–121. 34.9% 1763

Carter, Mia. "The Politics of Pleasure: Cross-Cultural Autobiographic Performance in the Video Works of Sadie Benning." *Signs* 23, no. 3 (Spring 1998): 745–769. 34.3% 1943

Oishi, Eve. "Visual Perversions: Race, Sex, and Cinematic Pleasure." *Signs* 31, no. 3 (Spring 2006): 641–674. 33.4% 2164

Williams, Linda. "Why I Did Not Want to Write This Essay." *Signs* 30, no. 1 (Autumn 2004): 1264–000. 32.6% 433

Stacey, Jackie. "Masculinity, Masquerade, and Genetic Impersonation: *Gattaca's* Queer Visions." *Signs* 30, no. 3 (Spring 2005): 1851–1877. 32.1% 1412

Shimizu, Celine Parreñas, and Helen Lee. "Sex Acts: Two Meditations on Race and Sexuality." *Signs* 30, no. 1 (Autumn 2004): 1385–1402. 32.0% 1075

Warren, Shilyh. "Recognition on the Surface of Madeline Anderson's *I Am Somebody*." *Signs* 38, no. 2 (Winter 2013): 353–378. 31.1% 1336

Kosta, Barbara, and Richard W. McCormick. "Interview with Jutta Brückner." *Signs* 21, no. 2 (Winter 1996): 343–373. 30.0% 1806

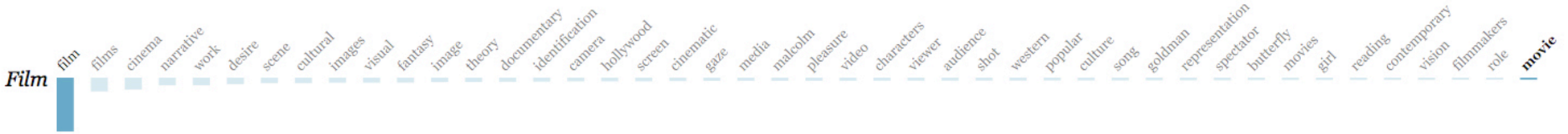
Aufderheide, Pat, and Debra Zimmerman. "From A to Z: A Conversation on Women's Filmmaking." *Signs* 30, no. 1 (Autumn 2004): 1455–1472. 29.8% 846

Overviews
Topic
Article
Word
Bibliography
Word index
Interpreting the model
Settings
Getting started

Enter a word | List topics

Prominent topics for *movie*

Click row labels to go to the corresponding topic page; click a word to show the topic list for that word.



Existing Studies

Existing Studies

- Cameron Blevins: "Topic Modeling Martha Ballard's Diary" (2010): diary

Existing Studies

- Cameron Blevins: "Topic Modeling Martha Ballard's Diary" (2010): diary
- Ted Underwood und Andrew Goldstone (2012): "What can topic models of PMLA teach us...": history of a discipline

Existing Studies

- Cameron Blevins: "Topic Modeling Martha Ballard's Diary" (2010): diary
- Ted Underwood und Andrew Goldstone (2012): "What can topic models of PMLA teach us...": history of a discipline
- Lisa Rhody, "Topic Modeling and Figurative Language" (2012): ekphrasis in poetry

Existing Studies

- Cameron Blevins: "Topic Modeling Martha Ballard's Diary" (2010): diary
- Ted Underwood und Andrew Goldstone (2012): "What can topic models of PMLA teach us...": history of a discipline
- Lisa Rhody, "Topic Modeling and Figurative Language" (2012): ekphrasis in poetry
- Matthew Jockers, Macroanalysis (2013): novel, nationality, gender

Existing Studies

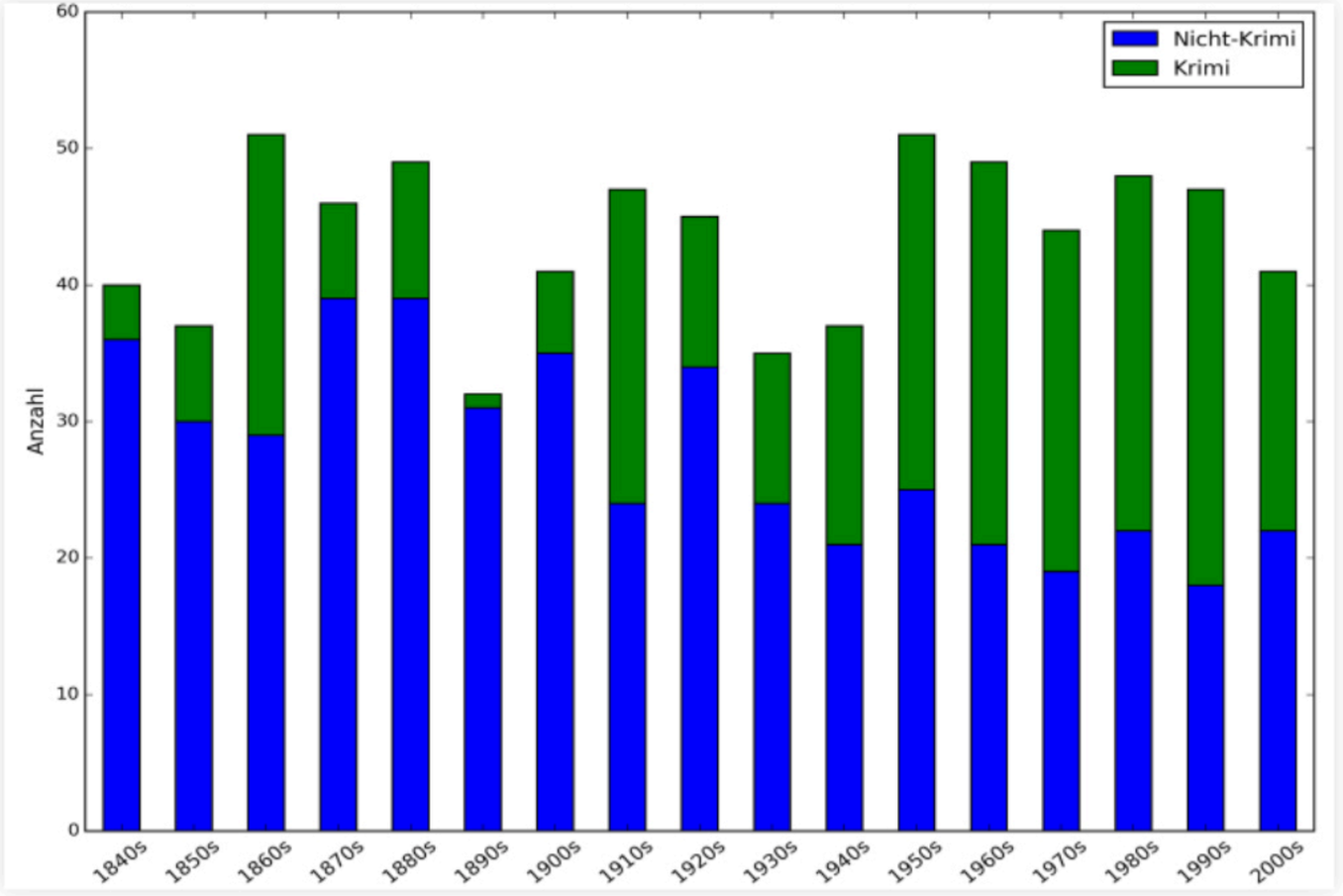
- Cameron Blevins: "Topic Modeling Martha Ballard's Diary" (2010): diary
- Ted Underwood und Andrew Goldstone (2012): "What can topic models of PMLA teach us...": history of a discipline
- Lisa Rhody, "Topic Modeling and Figurative Language" (2012): ekphrasis in poetry
- Matthew Jockers, Macroanalysis (2013): novel, nationality, gender
- Ben Schmidt: "Typical TV episodes" (2014): TV shows; temporal development

Existing Studies

- Cameron Blevins: "Topic Modeling Martha Ballard's Diary" (2010): diary
- Ted Underwood und Andrew Goldstone (2012): "What can topic models of PMLA teach us...": history of a discipline
- Lisa Rhody, "Topic Modeling and Figurative Language" (2012): ekphrasis in poetry
- Matthew Jockers, Macroanalysis (2013): novel, nationality, gender
- Ben Schmidt: "Typical TV episodes" (2014): TV shows; temporal development
- Christof Schöch, "Topic Modeling Genre" (2017): drama, subgenres

(b) A topic model for French crime fiction

Text collection: 840 French Novels



(Image credit: Christof Schöch, 2019, CC-BY 4.0 Int'l)

Crime fiction (prototypical)

Crime fiction (prototypical)

- Long, narrative, fictional prose (=novel)

Crime fiction (prototypical)

- Long, narrative, fictional prose (=novel)
- Character inventory: investigators, criminals, suspects, witnesses, victims

Crime fiction (prototypical)

- Long, narrative, fictional prose (=novel)
- Character inventory: investigators, criminals, suspects, witnesses, victims
- Plot: violent crime, rational elucidation

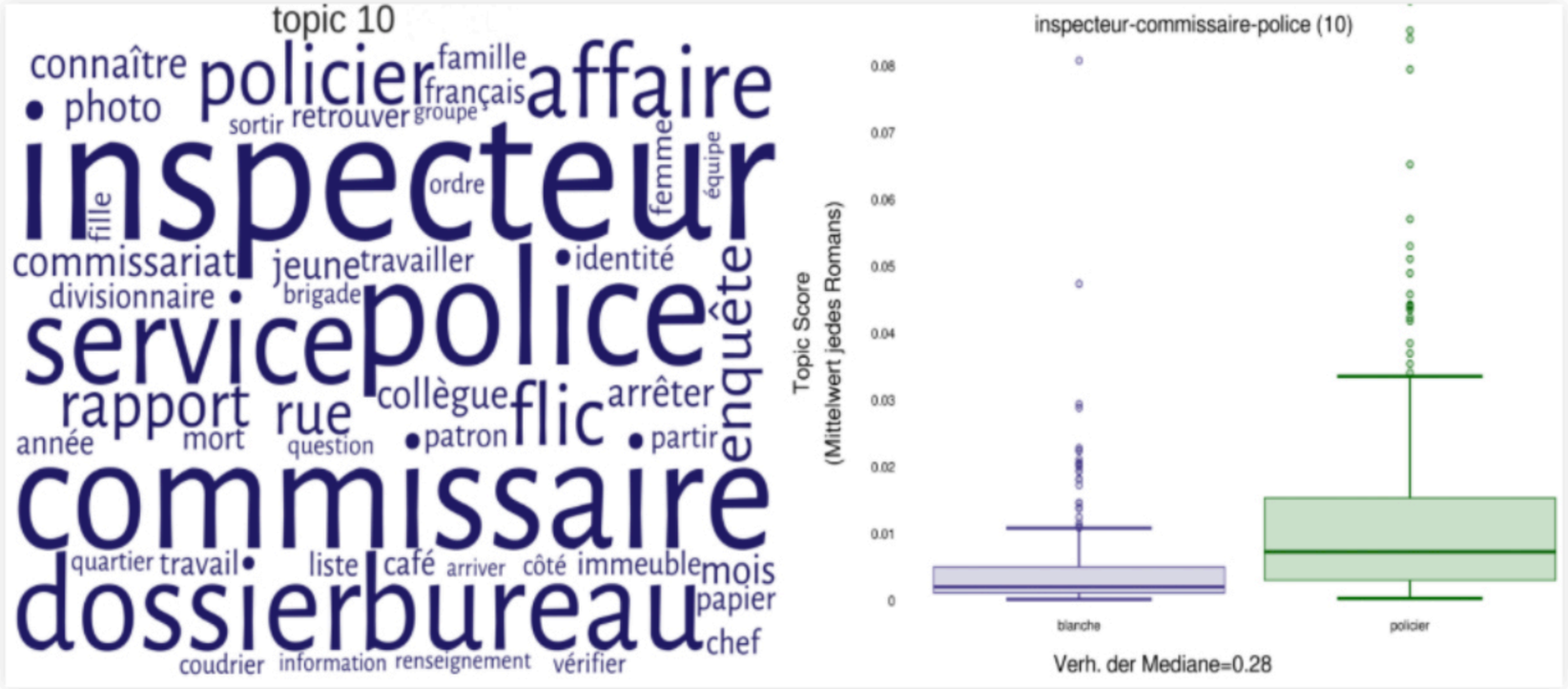
Crime fiction (prototypical)

- Long, narrative, fictional prose (=novel)
- Character inventory: investigators, criminals, suspects, witnesses, victims
- Plot: violent crime, rational elucidation
- Setting: urban space

Crime fiction (prototypical)

- Long, narrative, fictional prose (=novel)
- Character inventory: investigators, criminals, suspects, witnesses, victims
- Plot: violent crime, rational elucidation
- Setting: urban space
- => Hypotheses regarding possible topics

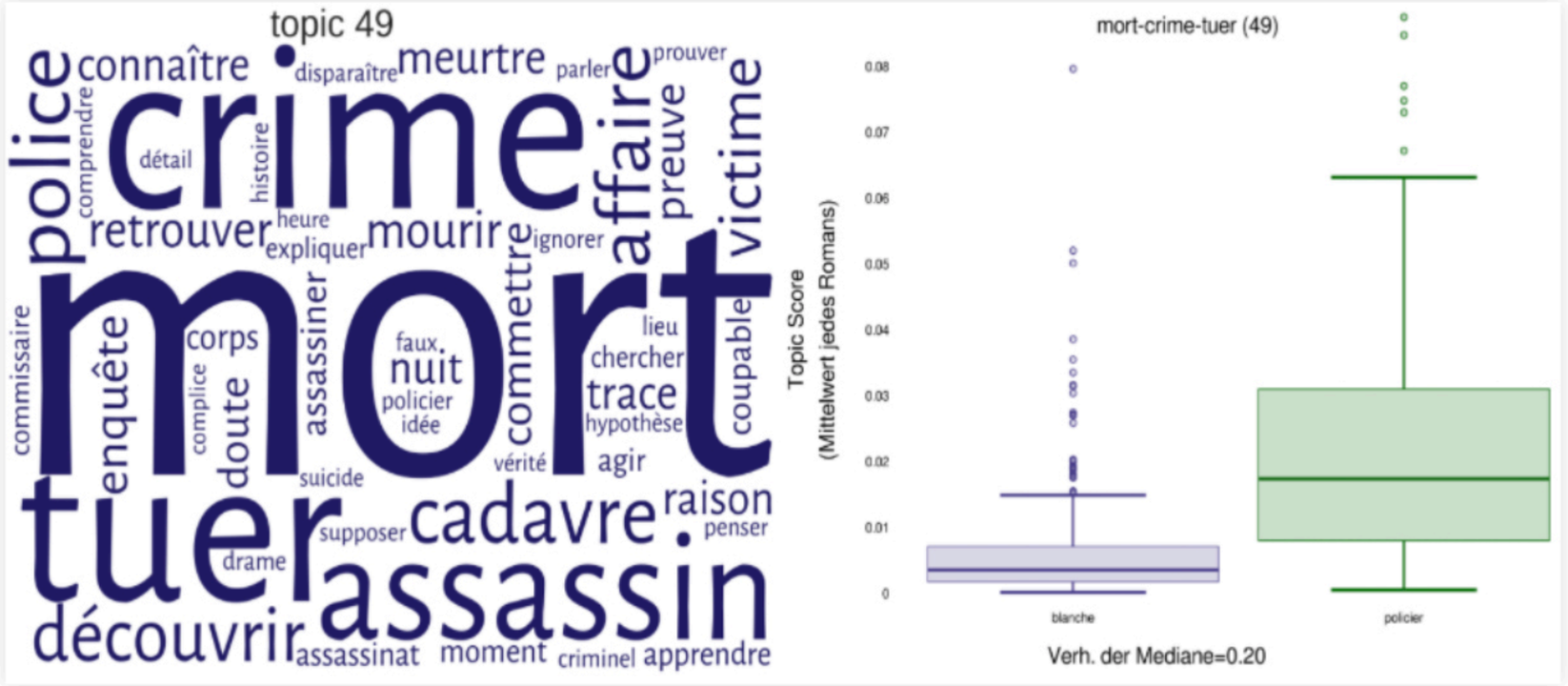
Topic and subgenre



Topic 10: detective, inspector, police

Distinctive of crime fiction (content & statistics) ($p < \alpha=0.01$)
(Image credit: Christof Schöch, 2019, CC-BY 4.0 Int'l)

Topic and subgenre

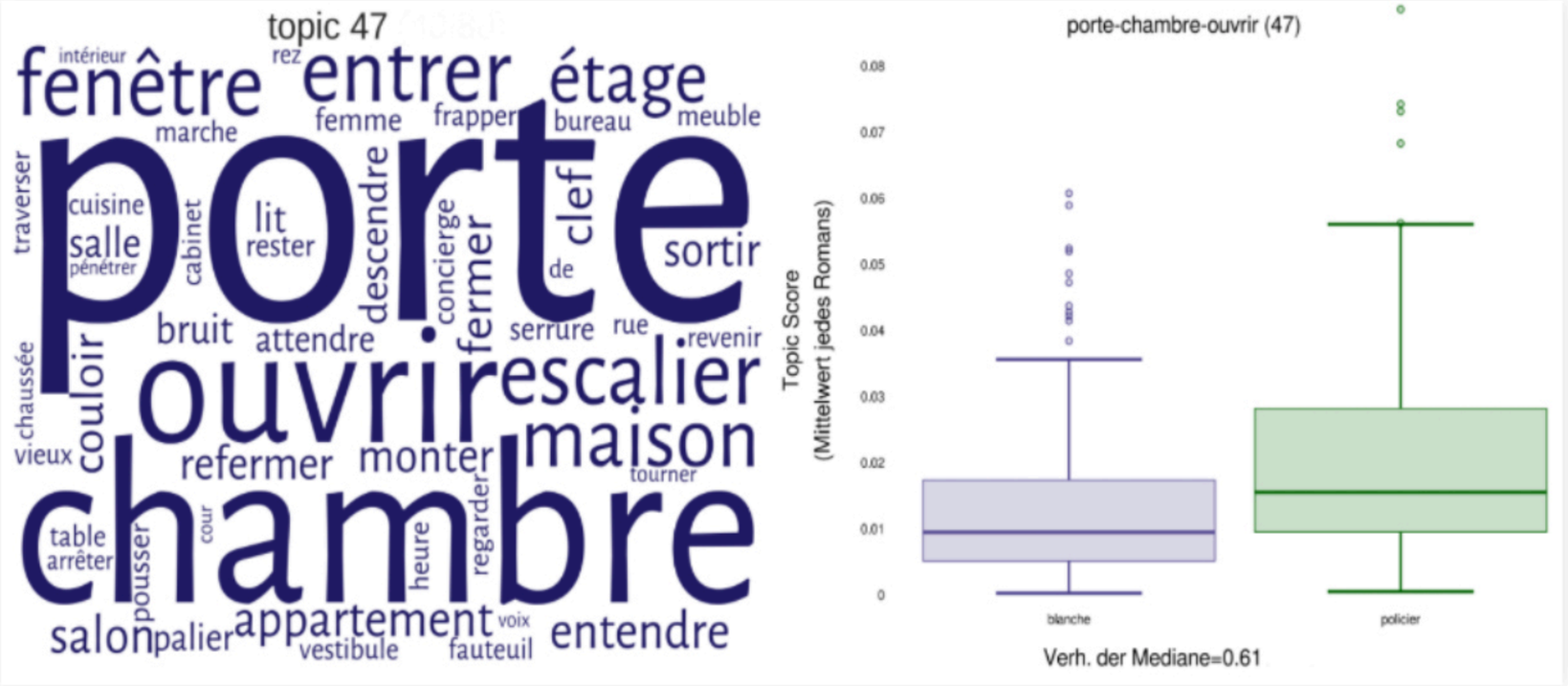


Topic 49: death, crime, to kill

Distinctive of crime fiction (content & statistics) ($p < \alpha=0.01$)

(Image credit: Christof Schöch, 2019, CC-BY 4.0 Int'l)

Topic and subgenre

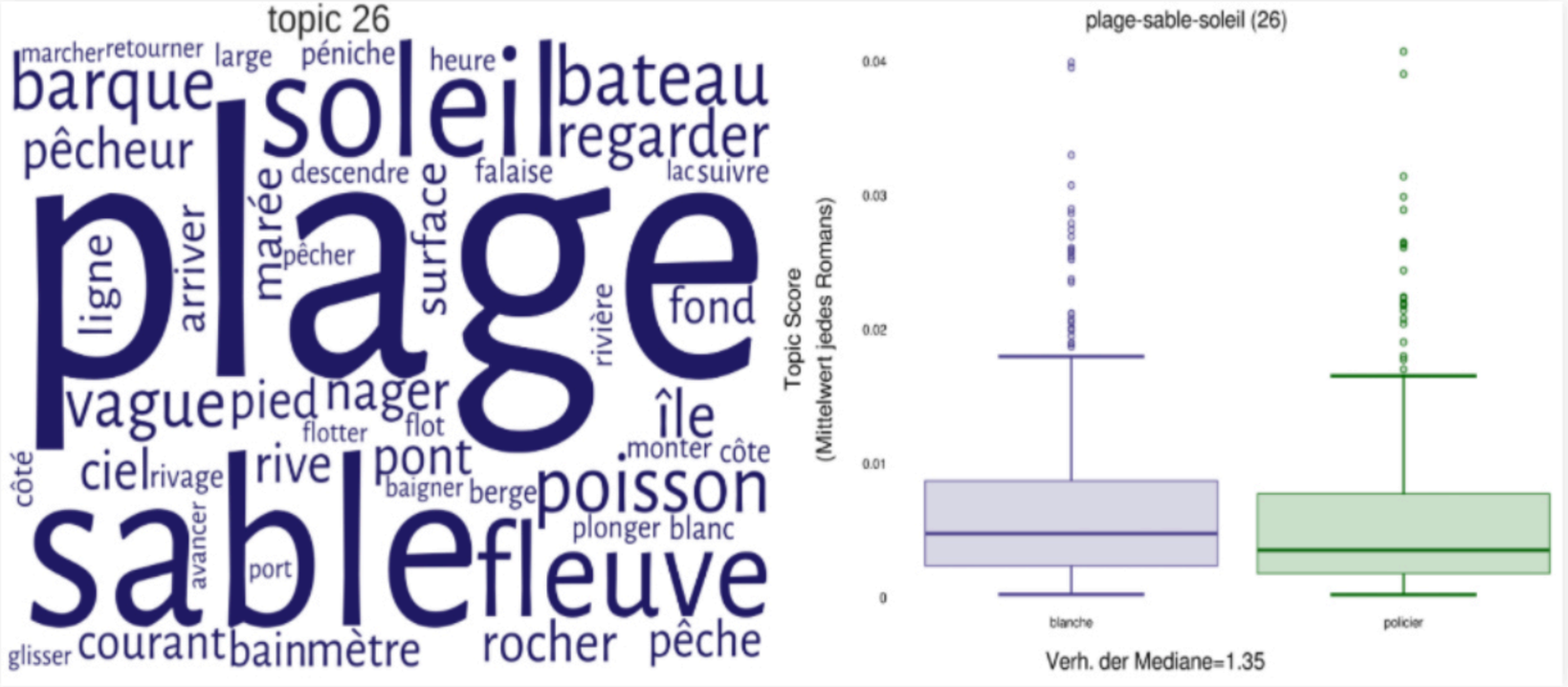


Topic 47: door, room, to open

Statistically distinctive ($p < \alpha=0.01$); but content-wise?

(Image credit: Christof Schöch, 2019, CC-BY 4.0 Int'l)

Topic and subgenre

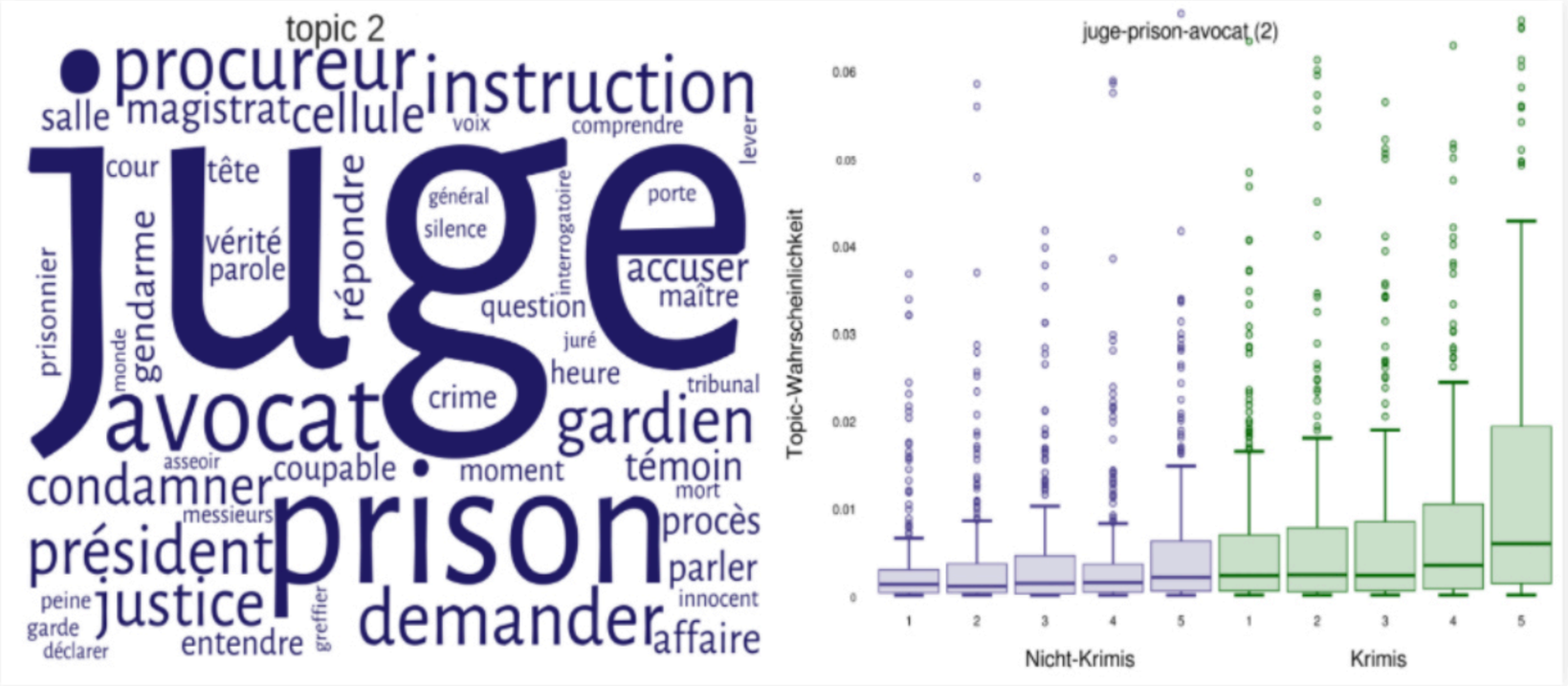


Topic 26: beach, sand, sun

Distinctive of non-crime fiction ($p < \alpha=0.001$)

(Image credit: Christof Schöch, 2019, CC-BY 4.0 Int'l)

Topics over text segments

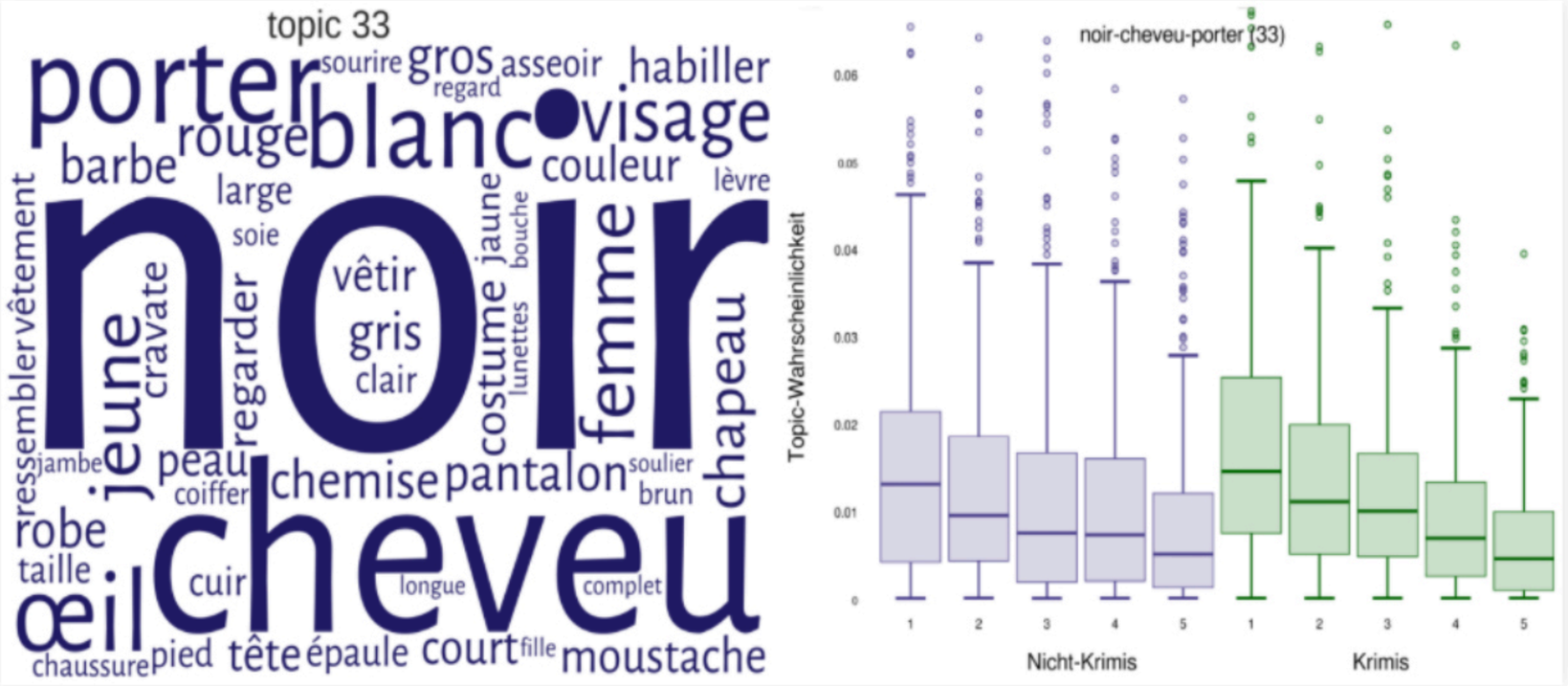


Topic 2: judge, prison, lawyer/attorney

Statistically significant (crime fiction): (1,4), (4,5) etc.

(Image credit: Christof Schöch, 2019, CC-BY 4.0 Int'l)

Topics over text segments



Topic 33: black, hair, eyes, wear, eye, face

Statistically significant: crime fiction all but (2,3); non-crime fiction (1,3), (2,5)

(Image credit: Christof Schöch, 2019, CC-BY 4.0 Int'l)

Overall results

Overall results

- A large part of the topics is statistically distinctive: crime fiction (31/80) non-crime fiction (21/80)

Overall results

- A large part of the topics is statistically distinctive: crime fiction (31/80) non-crime fiction (21/80)
- Topics are not just themes, but also narrative motives, descriptive elements, character sets

Overall results

- A large part of the topics is statistically distinctive: crime fiction (31/80) non-crime fiction (21/80)
- Topics are not just themes, but also narrative motives, descriptive elements, character sets
- Textual progression: only a few topics have significant trends

Overall results

- A large part of the topics is statistically distinctive: crime fiction (31/80) non-crime fiction (21/80)
- Topics are not just themes, but also narrative motives, descriptive elements, character sets
- Textual progression: only a few topics have significant trends
- Overall: we can detect thematic trends in 840 novels without reading (all of) them!

Time for questions

Topic Modeling: Theory

(a) How does a topic model look like?

On a practical level

- A topic is a group of words with some (semantic) relation (e.g., common theme, motive, etc.)

On a practical level

- A topic is a group of words with some (semantic) relation (e.g., common theme, motive, etc.)
- Each topic is made up of words of varying importance and relevance to the topic

On a practical level

- A topic is a group of words with some (semantic) relation (e.g., common theme, motive, etc.)
- Each topic is made up of words of varying importance and relevance to the topic
- Each document is made up of several topics in various proportions

On a technical level

- A topic model is an abstract representation of all topics and documents in a collection

On a technical level

- A topic model is an abstract representation of all topics and documents in a collection
- A topic is a probability distribution over words

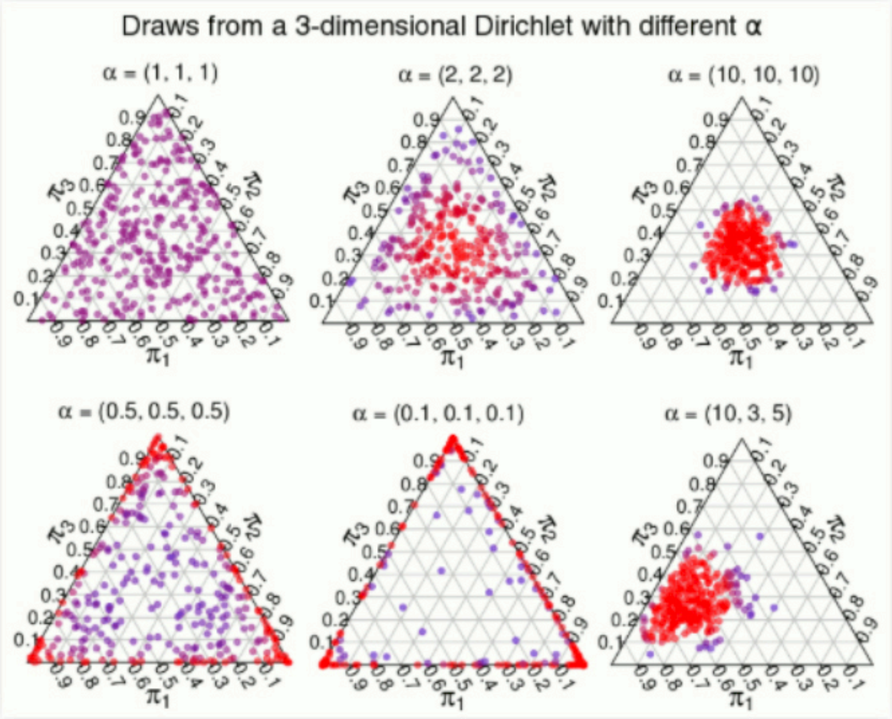
On a technical level

- A topic model is an abstract representation of all topics and documents in a collection
- A topic is a probability distribution over words
- A document is a probability distribution over topics

On a technical level

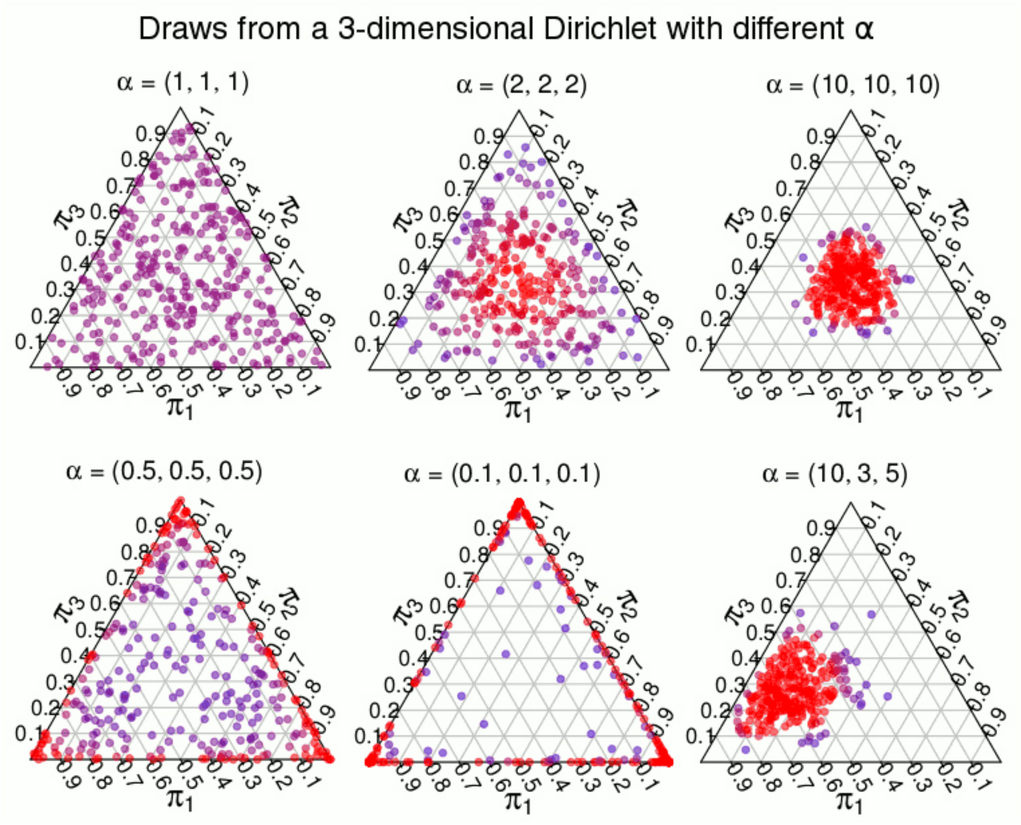
- A topic model is an abstract representation of all topics and documents in a collection
- A topic is a probability distribution over words
- A document is a probability distribution over topics
- The Dirichlet distribution (in LDA) describes the topic mixture distribution of the model

Dirichlet distributions



Describe the topic mixture distributions of the model.
Here several possible distributions with three topics.

(Image credit: Rasmus Bååth, "The Non-parametric Bootstrap as a Bayesian Model", R-Bloggers, 2015, https://www.r-bloggers.com/the-non-parametric-bootstrap-as-a-bayesian-model/) By


Draws from a 3-dimensional Dirichlet with different α
α = (1, 1, 1)
α = (2, 2, 2)
α = (10, 10, 10)
α = (0.5, 0.5, 0.5)
α = (0.1, 0.1, 0.1)
α = (10, 3, 5)
π1
π2
π3

Words in topic distribution

topic	word	rank	score
5	horse	1	169.0182333
5	road	2	120.0182333
5	carriage	3	68.01823332
5	sergeant	4	52.01823332
5	companion	5	40.01823332
…	…	…	…
5	bridle	50	9.018233317
…	…	…	…
5	charge	100	4.018233317
…	…	…	…
5	garden	500	0.018233317

(Each word has a score in each topic; here ordered by topic/rank)

Topics in document distribution

doc	doc-id	t1	t2	t3	t4	…
0	acd009§0005.txt	0.0009913736	0.0003037892	0.0002364825	0.0980390649	…
1	acd005§0068.txt	0.0015847905	0.0004856314	0.0070289298	0.1700251167	…
2	acd010§0026.txt	0.1416548293	0.2319451602	0.0001779831	0.0236859739	…
3	acd011§0020.txt	0.0007777962	0.0002383421	0.0001855357	0.0246910723	…
4	acd005§0048.txt	0.0813954255	0.0002265121	0.2049190768	0.1072239384	…
5	acd006§0001.txt	0.0007063237	0.0002164406	0.0001684866	0.1617409925	…
6	acd006§0070.txt	0.0250133258	0.0002216979	0.0001725791	0.0928001193	…
7	acd004§0054.txt	0.0047744552	0.0002815341	0.0002191581	0.3144879413	…
8	acd010§0036.txt	0.0761023277	0.2589537644	0.0001861433	0.0116724568	…
9	acd005§0031.txt	0.0484746437	0.0002322766	0.000180814	0.1767559653	…
10	acd006§0019.txt	0.0130915999	0.0002951922	0.0083153112	0.0993073862	…
…	…	…	…	…	…	…

(Each topic has a score in each document; ordered by document)

(b) How is a Topic Model created?

Some relevant ideas

- The most widespread implementation uses 'Latent Dirichlet Allocation'

Some relevant ideas

- The most widespread implementation uses 'Latent Dirichlet Allocation'
- Follows the "bag-of-words"-model: word order is irrelevant

Some relevant ideas

- The most widespread implementation uses 'Latent Dirichlet Allocation'
- Follows the "bag-of-words"-model: word order is irrelevant
- No semantic knowledge / dictionary / WordNet etc. is used; language-independent

Some relevant ideas

- The most widespread implementation uses 'Latent Dirichlet Allocation'
- Follows the "bag-of-words"-model: word order is irrelevant
- No semantic knowledge / dictionary / WordNet etc. is used; language-independent
- Based on distributional semantics: "a word is characterized by the company it keeps" (John Firth 1957)

Some relevant ideas

- The most widespread implementation uses 'Latent Dirichlet Allocation'
- Follows the "bag-of-words"-model: word order is irrelevant
- No semantic knowledge / dictionary / WordNet etc. is used; language-independent
- Based on distributional semantics: "a word is characterized by the company it keeps" (John Firth 1957)
- Discovers words which frequently occur together or in similar contexts (=topics)

Some relevant ideas

- The most widespread implementation uses 'Latent Dirichlet Allocation'
- Follows the "bag-of-words"-model: word order is irrelevant
- No semantic knowledge / dictionary / WordNet etc. is used; language-independent
- Based on distributional semantics: "a word is characterized by the company it keeps" (John Firth 1957)
- Discovers words which frequently occur together or in similar contexts (=topics)
- Infers how important each word is in each topic

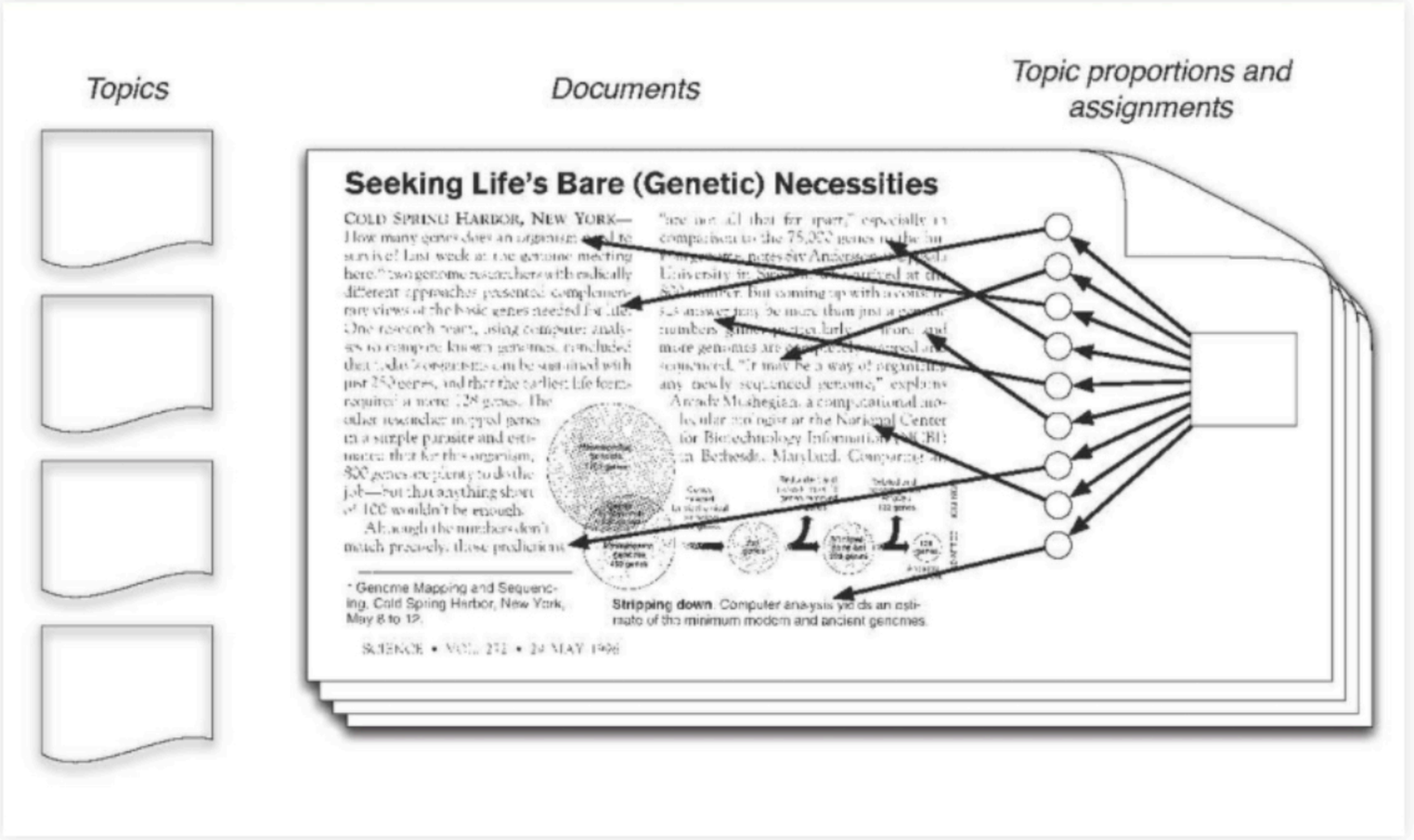
Some relevant ideas

- The most widespread implementation uses 'Latent Dirichlet Allocation'
- Follows the "bag-of-words"-model: word order is irrelevant
- No semantic knowledge / dictionary / WordNet etc. is used; language-independent
- Based on distributional semantics: "a word is characterized by the company it keeps" (John Firth 1957)
- Discovers words which frequently occur together or in similar contexts (=topics)
- Infers how important each word is in each topic
- Infers how important each topic is in each document

Generative, inverted, iterative

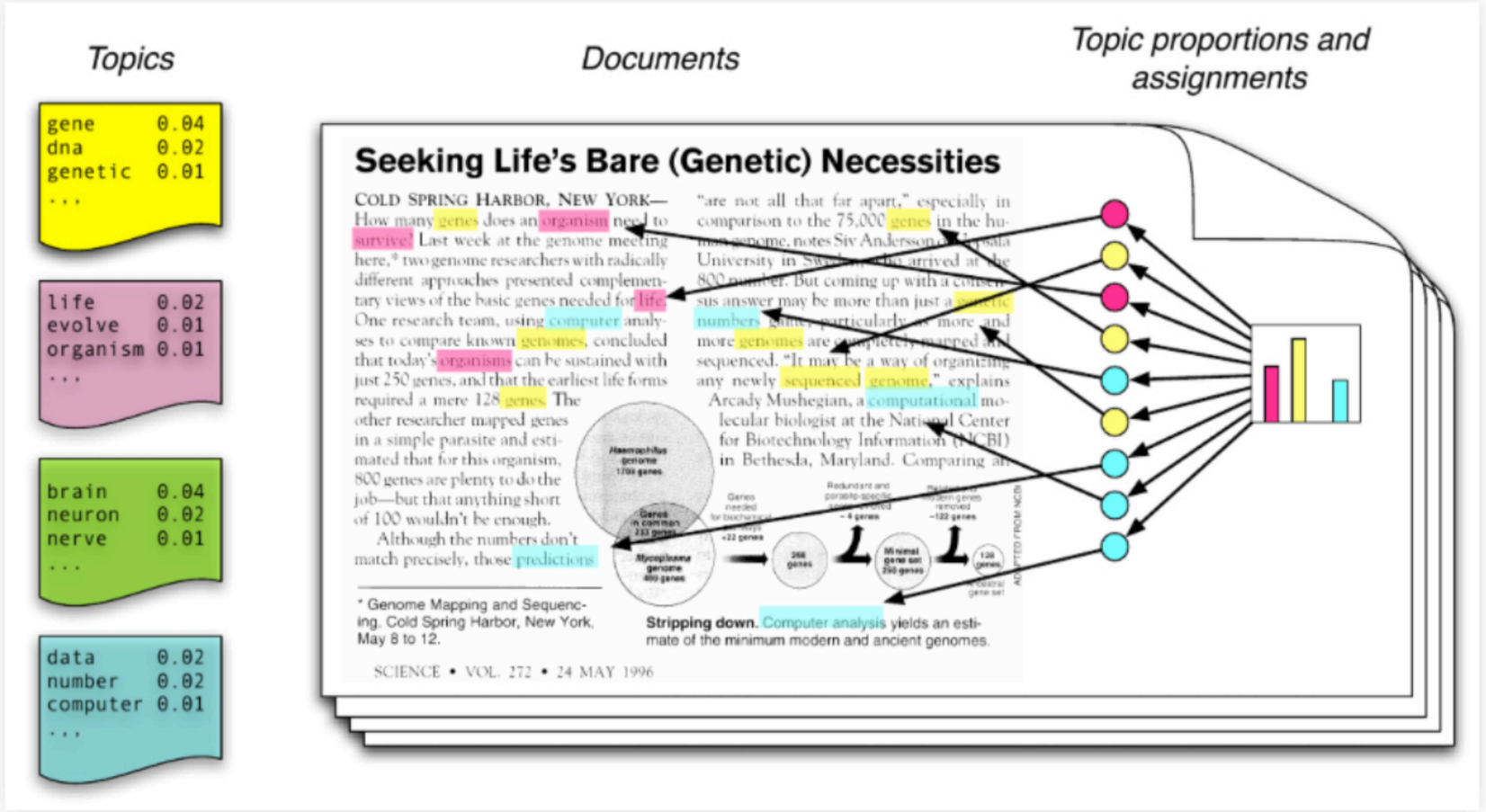
"A topic model is a generative model for documents: it specifies a simple probabilistic procedure by which documents can be generated. To make a new document, one chooses a distribution over topics. Then, for each word in that document, one chooses a topic at random according to this distribution, and draws a word from that topic. Standard statistical techniques can be used to invert this process, inferring the set of topics that were responsible for generating a collection of documents."
(Steyvers and Griffiths 2006)

Inference problem: observed data



(Image source: David Blei, "Topic Models", lecture slides, 2009, http://videolectures.net/mlss09uk_blei_tm/)

Inferred, latent model



(Image source: David Blei, "Topic Models", lecture slides, 2009, http://videolectures.net/mlss09uk_blei_tm/)

The starting point of LDA

- We have the documents with their words (e.g. as a word/document frequency matrix)

The starting point of LDA

- We have the documents with their words (e.g. as a word/document frequency matrix)
- We are looking for the word distributions per topic, the topic distributions per document, and the topic assignment of each word

The starting point of LDA

- We have the documents with their words (e.g. as a word/document frequency matrix)
- We are looking for the word distributions per topic, the topic distributions per document, and the topic assignment of each word
- Both distributions are dependent on each other (if a topic changes, the topic distributions change)

The starting point of LDA

- We have the documents with their words (e.g. as a word/document frequency matrix)
- We are looking for the word distributions per topic, the topic distributions per document, and the topic assignment of each word
- Both distributions are dependent on each other (if a topic changes, the topic distributions change)
- And both distributions need to fit with the original documents

The generative model behind LDA

- For each topic, there is a distribution over words
- For each document, there is a distribution over topics
- For each word in each document:
  - We sample a topic from the topic distribution of that document
  - We sample a word from the word distribution of that topic
- This can only work if we have the distributions; which we don't

Random initialization

- For each document, we generate a random distribution over topics

Random initialization

- For each document, we generate a random distribution over topics
- For each topic, we generate a random distribution over words

Random initialization

- For each document, we generate a random distribution over topics
- For each topic, we generate a random distribution over words
- For each word in each document:
  - Sample a topic from the topic distribution
  - Sample a word from the word distribution of that topic

Random initialization

- For each document, we generate a random distribution over topics
- For each topic, we generate a random distribution over words
- For each word in each document:
  - Sample a topic from the topic distribution
  - Sample a word from the word distribution of that topic
- Now we have a model; but we know it's most likely wrong (=low confidence)

Inference: iterative approximation

- Using the observed data and our (random/erroneous) model, we can improve the model

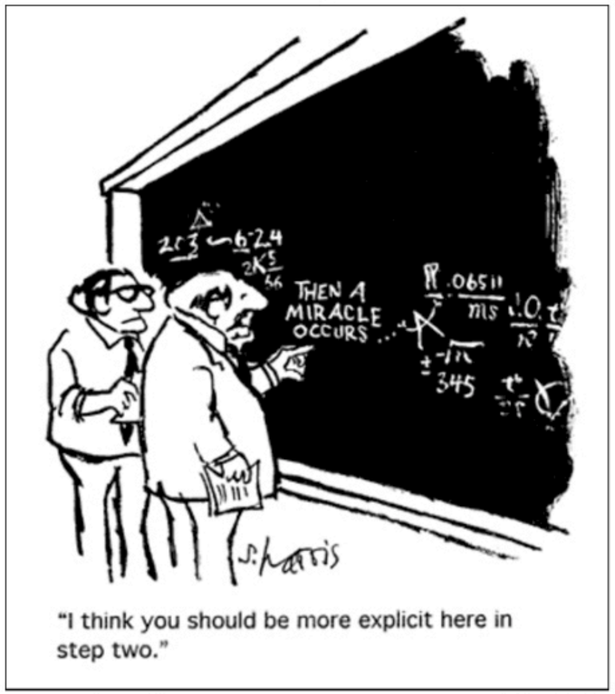
Inference: iterative approximation

- Using the observed data and our (random/erroneous) model, we can improve the model
- One among several methods: Gibbs sampling

Inference: iterative approximation

- Using the observed data and our (random/erroneous) model, we can improve the model
- One among several methods: Gibbs sampling
  - For one word in one document, remove the existing topic assignment
  - Based on the model and the other words in the document, assign a new topic to the word (cooccurrence!)
  - Update the overall model according to this assignment;
- Repeat until your time runs out or your evaluation task says it's ok to stop

How it works exactly, clearly explained



"I think you should be more explicit here in step two."

"I think you should be more explicit in step two..."

For reasons of copyright restrictions, please see here:
http://www.sciencecartoonsplus.com/gallery/math/index.php

Time for questions