Probabilistic Topic Models

Carlos Badenes-Olmedo
Oscar Corcho

Ontology Engineering Group (OEG)
Universidad Politécnica de Madrid (UPM)

K-CAP 2017
Knowledge Capture
December 4th-6th, 2017
Austin, Texas, United States

ocorcho@fi.upm.es
@ocorcho
oeg-upm.net
github.com/librairy

get ready to play ..

```
$ git clone git@github.com:librairy/tutorial.git .

$ docker run --name test -v "$PWD":/src librairy/compiler
```

Outline

1. **Artifacts**
2. Parameters
3. Training
4. Evaluation and Interpretation
5. Inference
6. Trends
7. Domains
8. Topic-based Similarity

Documents

Words

section

datum

extraction

dataset

knowledge

ontology

instance

science

term

page

example

query

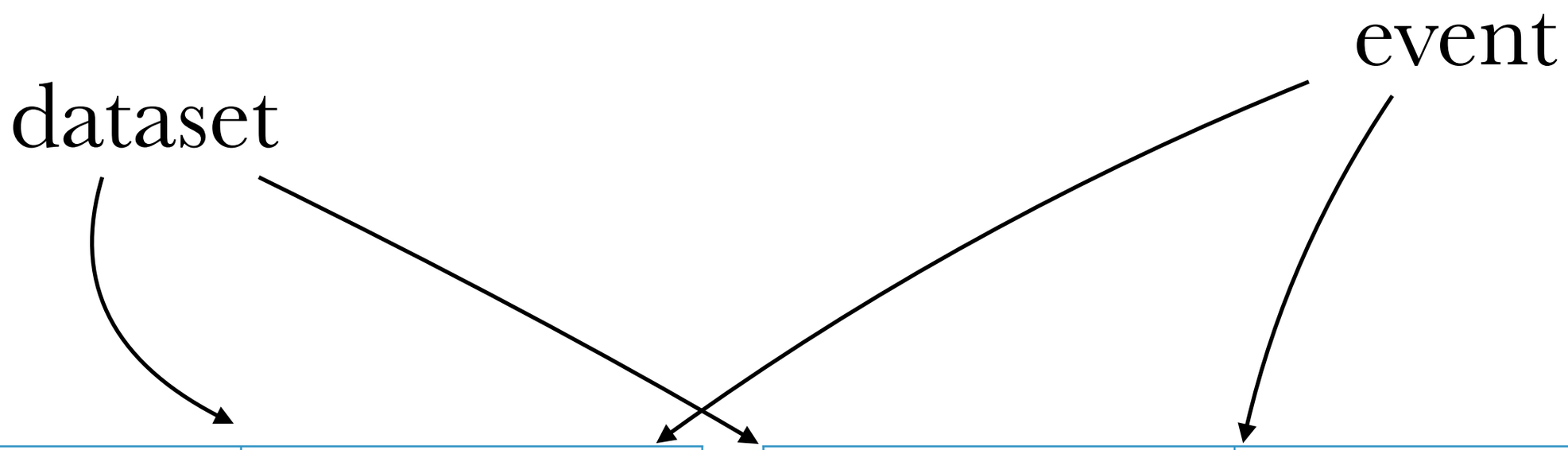
measure

class

concept

event

Topics



ontology	0.011696399528999426
query	0.0084018938429745 2
datum	0.007648923027408832
example	0.007628095111872131
knowledge	0.006779803572327723
page	0.006705717576295264
event	0.006642360848678050 5
case	0.006638825784874849
dataset	**0.006361954166489025**
concept	0.005970016686794867
..	..

class	0.0002888599713777317
extraction	0.0002747438325633771
datum	0.0002738706330554531
section	0.0002686989206165 26
example	0.0002680552015363467
event	0.0002648849674904949
dataset	**0.0002602128572589013**
case	0.0002595933993586539
result	0.0002585102532654629
entity	0.0002571243653838122
..	..

Documents

dataset

event

ontology	0.011696399528999 42
query	0.00840189384297452
datum	0.00764892302740883
example	0.00762809511187213
knowledge	0.00677980357232772
page	0.00670571757629526
event	0.00664236084867805
case	0.00663882578487484
dataset	**0.00636195416648902**
concept	0.00597001668679486
..	..

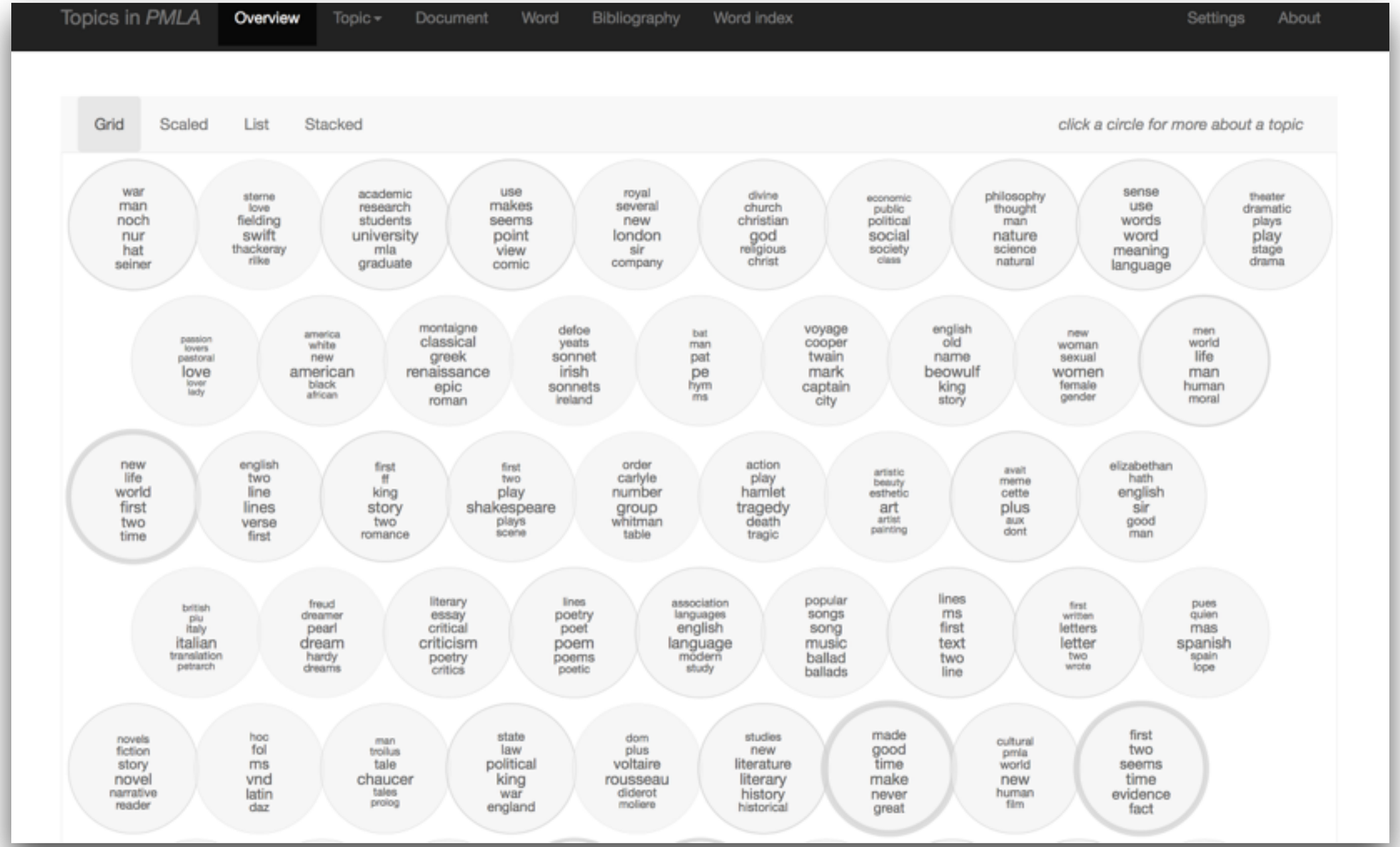
class	0.00028885997137773
extraction	0.00027474383256337
datum	0.00027387063305545
section	0.00026869892061652
example	0.00026805520153634
event	0.00026488496749049
dataset	**0.00026021285725890**
case	0.00025959339935865
result	0.00025851025326546
entity	0.00025712436538381
..	..

Representational Model

Topic Model

[T0, T1, T2]	TOPICS
	DOCUMENTS
[0.6, 0.3, 0.1]	
[0.6, 0.3, 0.1]	
[0.6, 0.3, 0.1]	
[0.6, 0.3, 0.1]	
[0.6, 0.3, 0.1]	
[0.6, 0.3, 0.1]	
	WORDS
[0.6, 0.3, 0.1]	dataset
[0.2, 0.2, 0.6]	event
[0.4, 0.3, 0.3]	term
[0.6, 0.3, 0.1]	datum
[0.6, 0.3, 0.1]	concept
[0.6, 0.3, 0.1]	knowledge

Visualization



- *PMLA* items from JSTOR's Data for Research service
- restricted to items categorized as "full-length articles" with more than 2000 words
- 5605 articles out of the 9200 items from the years 1889–2007
- LDA, 64 topics

1. Artifacts
2. **Parameters**
3. Training
4. Evaluation and Interpretation
5. Inference
6. Trends
7. Domains
8. Topic-based Similarity

Probabilistic Topic Models

Topics

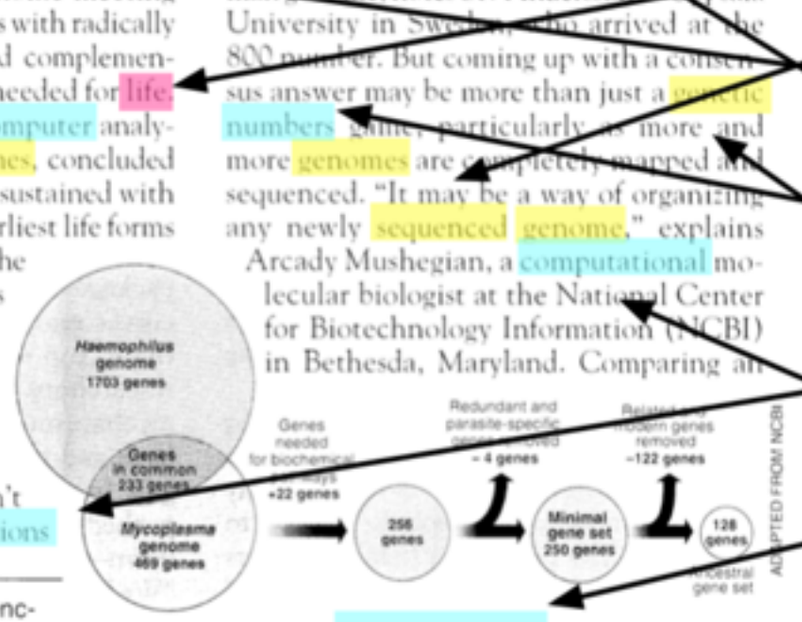
```
gene      0.04
dna       0.02
genetic   0.01
...

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



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

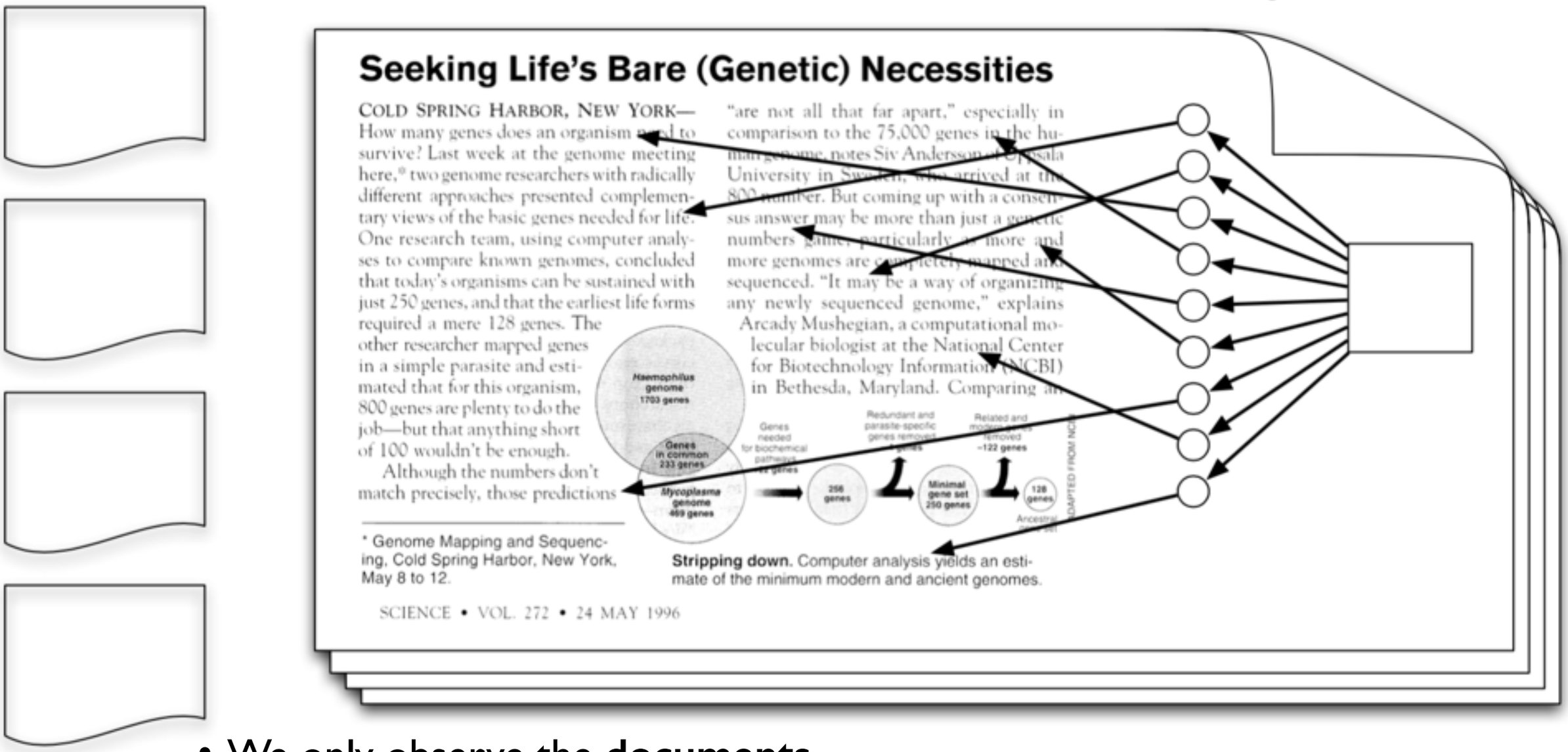
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments

- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

Probabilistic Topic Models



- We only observe the **documents**
- The other items in this structure are **hidden variables**.
- Our goal is to **infer** the hidden variables by computing their distribution conditioned on the documents:

$$p(\text{topics}, \text{proportions}, \text{assignments} \mid \text{documents})$$

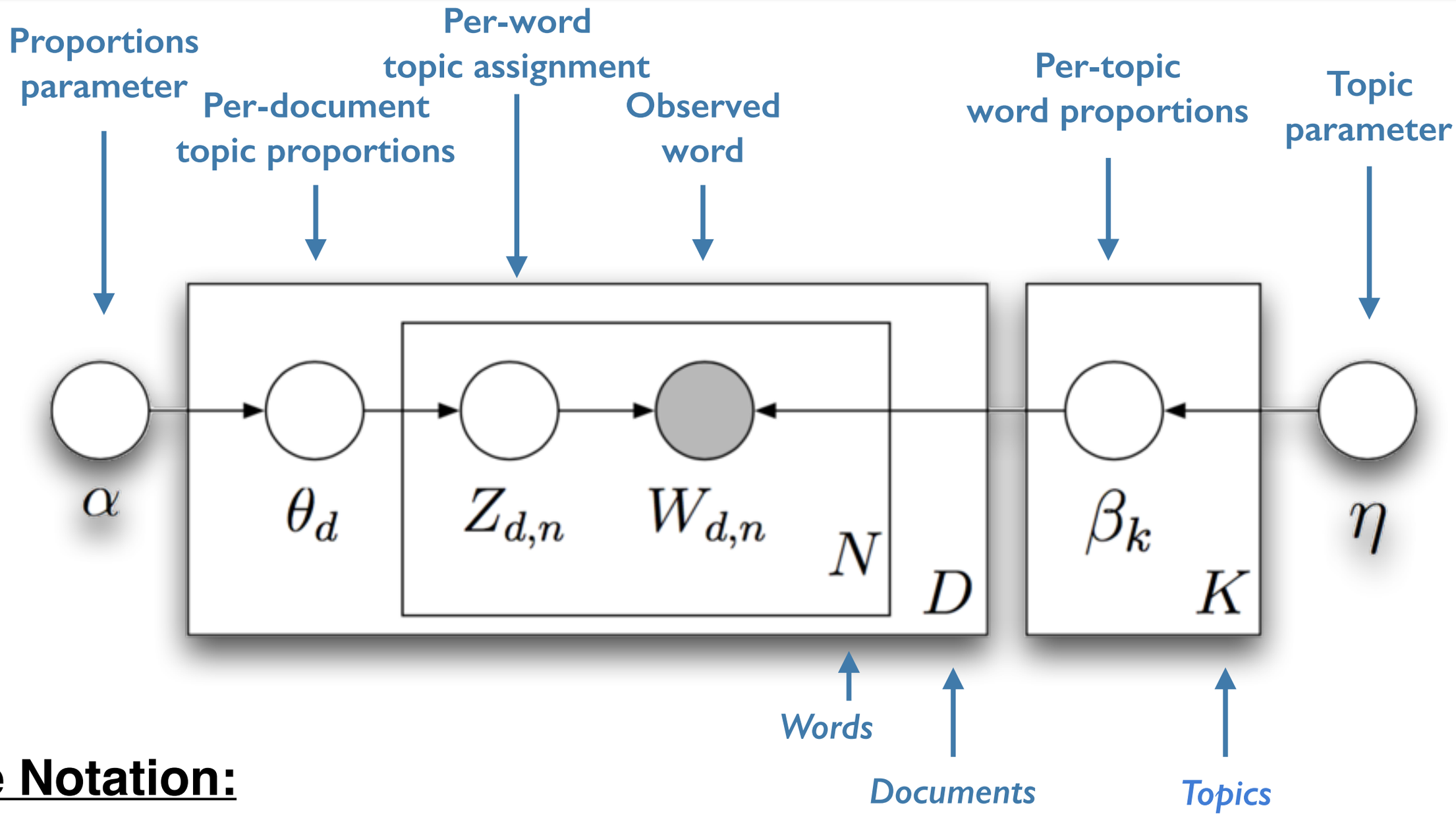
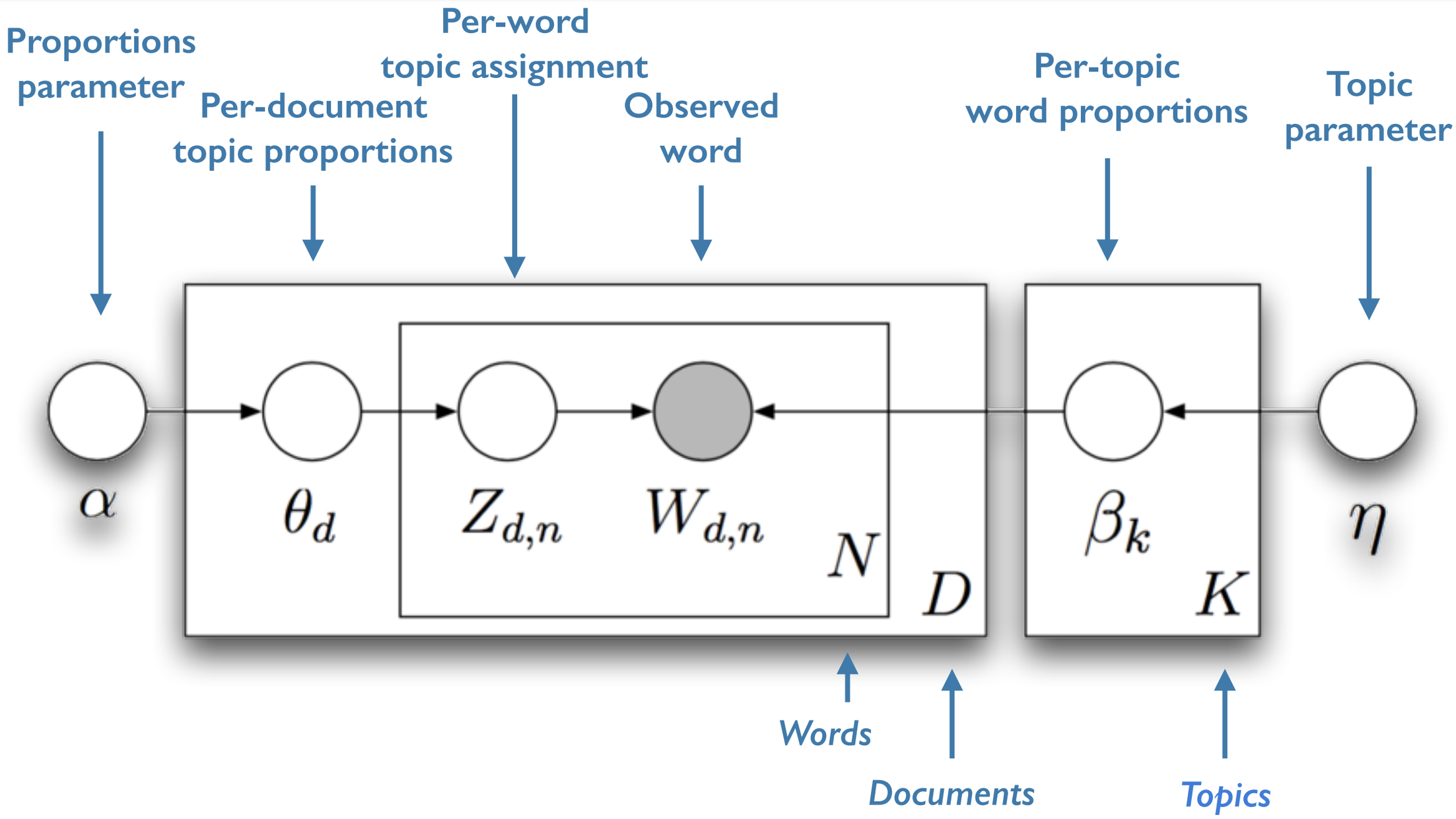


Plate Notation:

- Encodes **assumptions**
- Defines a **factorization** of the joint distribution
- Connects to **algorithms** for computing with data

- **Nodes** are random variables
- **Edges** indicate dependence
- **Shaded** nodes are observed
- **Plates** indicate replicated variables

Latent Dirichlet Allocation (LDA) [Blei et al, 2003]



$$\prod_{i=1}^{K} p(\beta_i \mid \eta) \prod_{d=1}^{D} p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right)$$

Probabilistic Topic Models as Graphical Models

$$\prod_{i=1}^{K} p(\beta_i \mid \eta) \prod_{d=1}^{D} p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right)$$

From a collection of documents, infer:

- Per-word topic assignment *Zd,n*
- Per-document topic proportions θ*d*
- Per-corpus topic distributions β*k*

Approximate posterior inference algorithms:

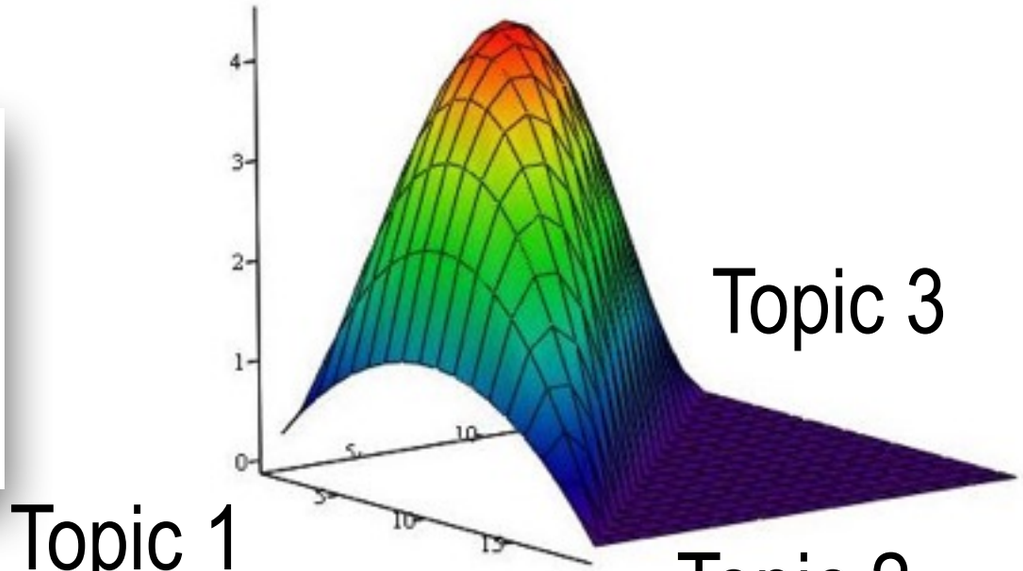
- Mean Field Variational methods [*Blei et al., 2001,2003*]
- Expectation Propagation [*Minka and Lafferty, 2002*]
- Collapsed Gibbs Sampling [*Griggiths and Steyvers, 2002*]
- Collapsed Variational Inference [*Teh et al., 2006*]
- Online Variational Inference [*Hoffman et al., 2010*]

Dirichlet Distribution

- Exponential family distribution over the simplex,
  i.e. ***positive vectors that sum to one***

- The parameter α controls the mean shape and sparsity of θ

$$p(\theta \mid \vec{\alpha}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}.$$



probability distribution of documents over topics (dt)

$\longrightarrow$ *[dt1, dt2, dt3]*

dt1+dt2+dt3 = 1

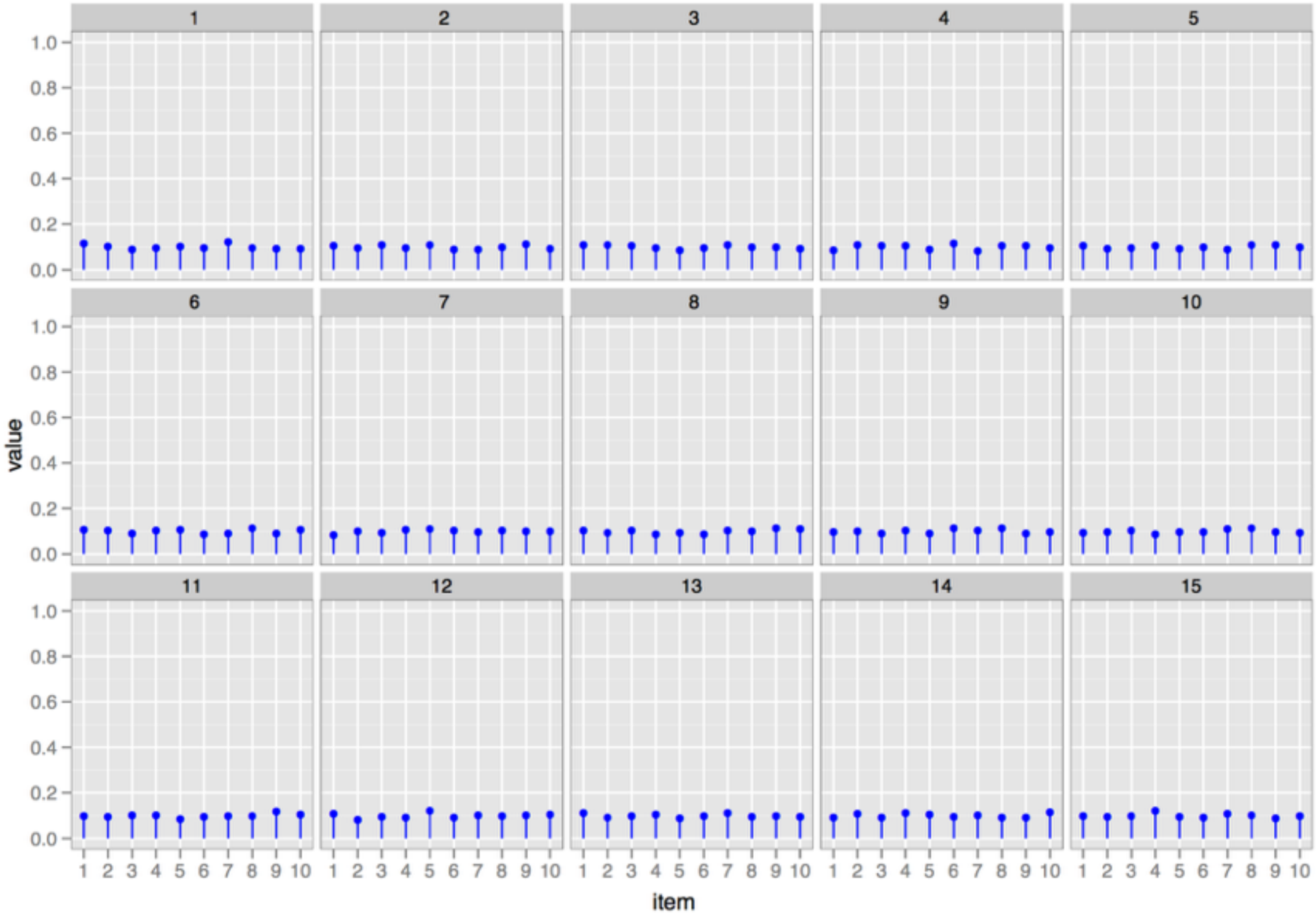
probability distribution of words over topics (wt)

Word $\rightarrow$ *[wt1, wt2, wt3]*

wt1+wt2+wt3 = 1

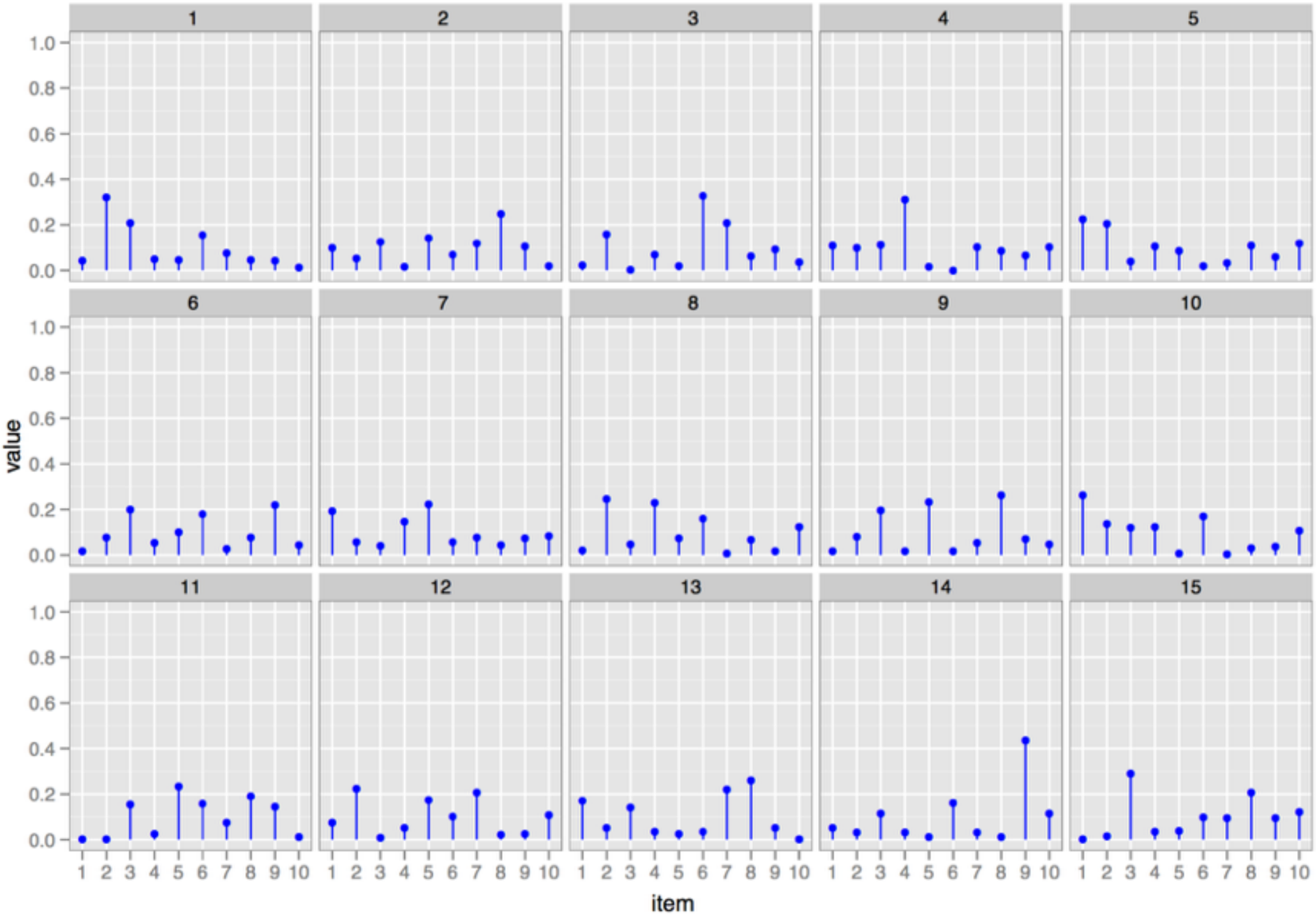
Latent Dirichlet Allocation (LDA) [Blei et al, 2003]

α=100



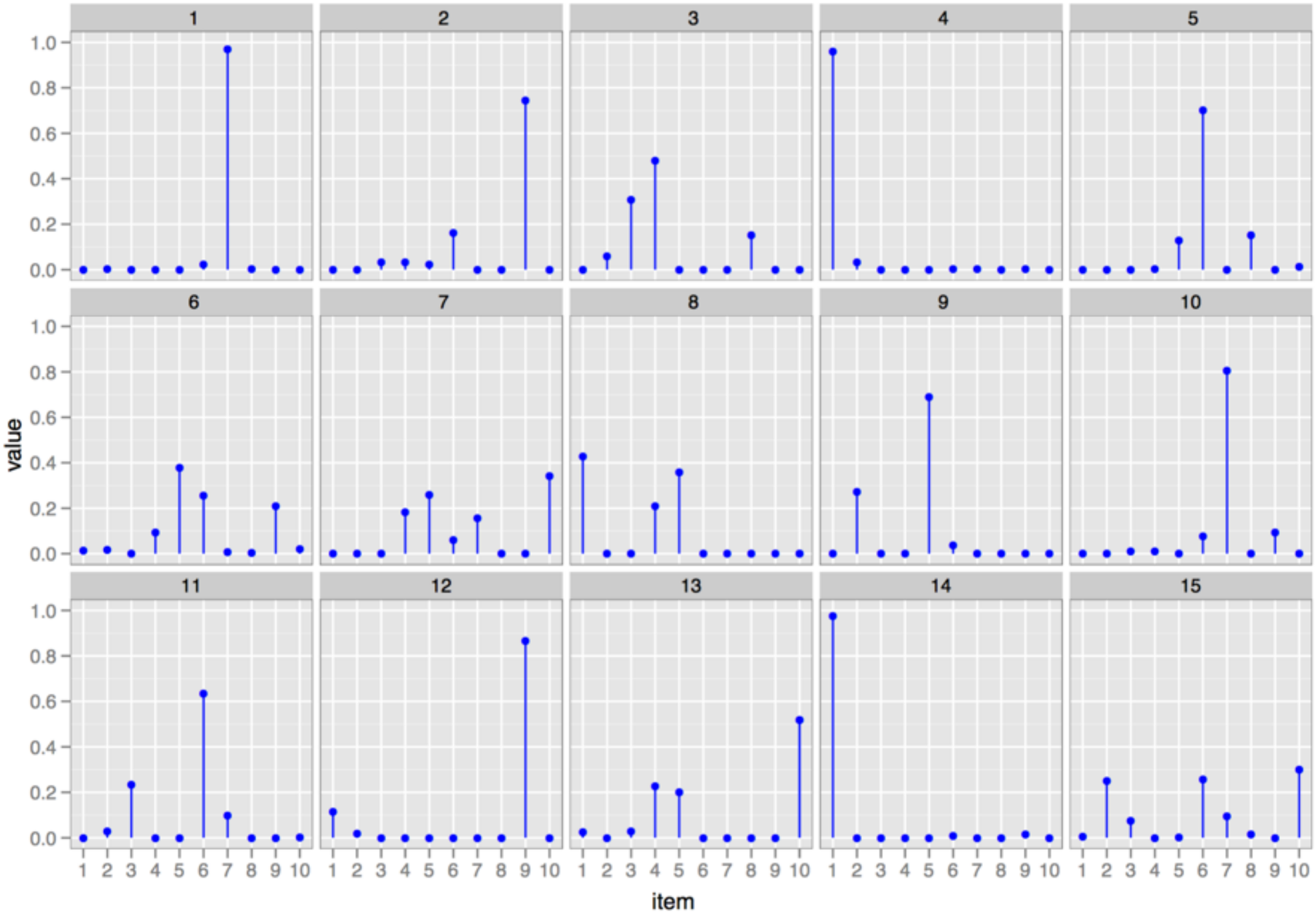
Latent Dirichlet Allocation (LDA) [Blei et al, 2003]

α=1



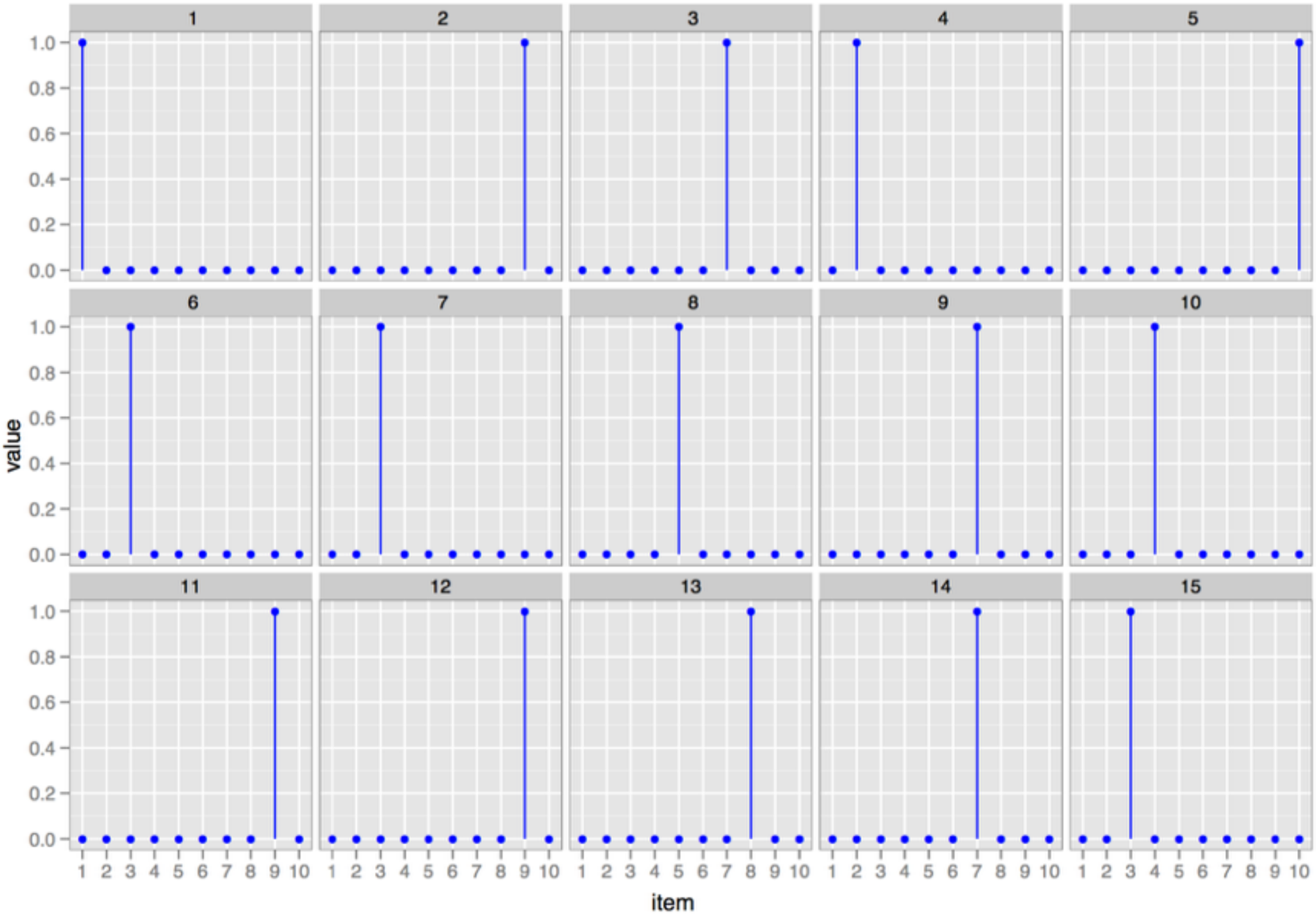
Latent Dirichlet Allocation (LDA) [Blei et al, 2003]

α=0.1



Latent Dirichlet Allocation (LDA) [Blei et al, 2003]

α=0.001



1. Artifacts
2. Parameters
3. **Training**
4. Evaluation and Interpretation
5. Inference
6. Trends
7. Domains
8. Topic-based Similarity

librAIry

http://librairy.linkeddata.es/api

user: oeg **password**: kcap2017

- **Distributing Text Mining tasks with librAIry.**
  Badenes-Olmedo, C.; Redondo-Garcia, J.; Corcho, O.
  In *Proceedings of the 2017 ACM Symposium on Document Engineering* (DocEng '17). ACM, 63-66. DOI: https://doi.org/10.1145/3103010.3121040

CORPUS

- **kcap**: <librairy-api>/domains/kcap

- **kcap2015**: <librairy-api>/domains/kcap2015

- **kcap2017**: <librairy-api>/domains/kcap2017

DOCUMENTS

- **Documents per Corpus**:
  <librairy-api>/domains/kcap/items

- **Document Content**:
  <librairy-api>/items/3707eb49b81fb67e76bf1d40da842275?content=true

- **Document Annotations**:
  <librairy-api>/items/3707eb49b81fb67e76bf1d40da842275/annotations

- **Document *(noun)* Lemmas**:
  <librairy-api>/items/3707eb49b81fb67e76bf1d40da842275/annotations/lemma

librAIry

TOPICS

- **Topics per Corpus**: <librairy-api>/domains/kcap/topics?words=10
- **Topic Details**: <librairy-api>/domains/kcap/topics/0
- **Most Representative Documents in a Topic:** <librairy-api>/domains/kcap/topics/0/items
- **Topics per Document**: <librairy-api>/domains/kcap/items/d8933e1de1888ccbdf4e0df42c404d7e/topics?words=5

librAIry

let's create an **LDA model**

TOPICS

- **Parameters**:

[GET] <librairy-api>/domains/{domainId}/parameters

- *lda.alpha = 0.1*
- *lda.beta = 0.01*
- *lda.topics = 6*
- *lda.vocabulary.size = 10000*
- *lda.max.iterations = 100*
- *lda.optimizer = manual (basic, nsga)*
- *lda.stopwords =example,service*

[POST]<librairy-api>/domains/{domainId}/parameters

```
{
  "name": "lda.beta",
  "value": "0.01"
}
```

- **(Re)Train a model**: [PUT] <librairy-api>/domains/kcap/topics

librAIry

TOPICS

- **Move** to the ‘lda’ stage

```
$ git checkout lda
```

- **Adjust** *parameter.properties*

```
$ nano parameters.properties
```

- **Try**

```
$ docker start -i test
```

- **Results** in ‘*output/models/lda' folder*

librAIry

TOPICS

- **See** '*output/models/lda/kcap' folder*:

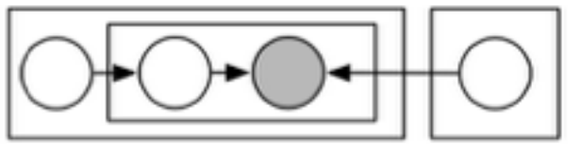
```
$ dataset.csv.gz         # texts
$ model.documents.txt    # topics per document
$ model.parameters.txt   # model settings
$ model.topics.txt       # topic words
$ model.uris.txt         # document URI
$ model.vocabulary.txt   # list of words
$ model.words.txt        # topics per word in document
```

librAIry

1. Artifacts
2. Parameters
3. Training
4. **Evaluation and Interpretation**
5. Inference
6. Trends
7. Domains
8. Topic-based Similarity

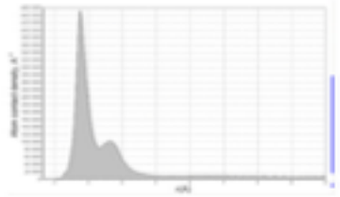
Evaluation and Interpretation

Make assumptions

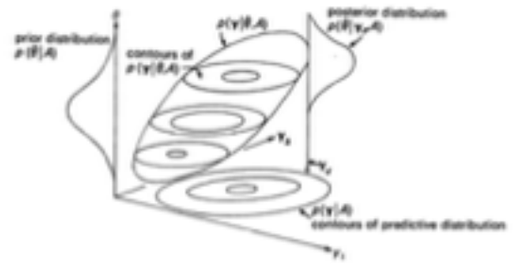


Collect data

Infer the posterior

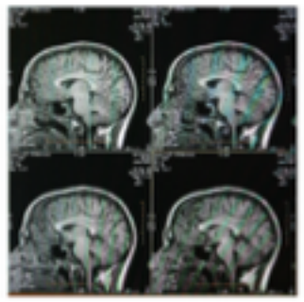


Check

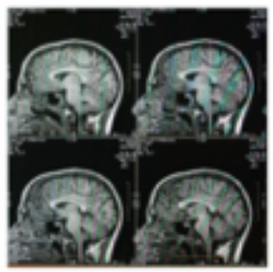


Predict

Explore



Explore



How interpretable is the model by a human

- **Word Intrusion**: how semantically 'cohesive' the topics inferred by a model are and tests whether topics correspond to natural groupings for humans (***Topic Coherence***)

{ dog, cat, horse, apple, pig, cow}

{ car, teacher, platypus, agile, blue, Zaire}

- **Topic Intrusion**: how well a topic model's decomposition of a document as a mixture of topics agrees with human associations of topics with a document

Chang, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. Advances in Neural Information Processing Systems 22, 288–296

Outline

1. Artifacts
2. Parameters
3. Training
4. Evaluation and Interpretation
5. **Inference**
6. Trends
7. Domains
8. Topic-based Similarity

INFERENCE TOPIC DISTRIBUTIONS

- **Create a Document**
  [POST] <librairy-api>/items/{id}

```
{
  "content": "string",
  "language": "string",
  "name": “string"
}
```

- **Add it to a Domain**
  [POST] <librairy-api>/domains/{domainId}/items/{itemId}

- **Read Topics per Document**:
  [GET] <librairy-api>/domains/{domainId}/items/{id}/topics

librAIry

INFERENCE TOPIC DISTRIBUTIONS

- **Move** to the ‘inference’ stage

```
$ git checkout inference
```

- **Write** text in file.txt*:*

```
$ nano file.txt
```

- ***Execute:***

```
$ docker start -i test
```

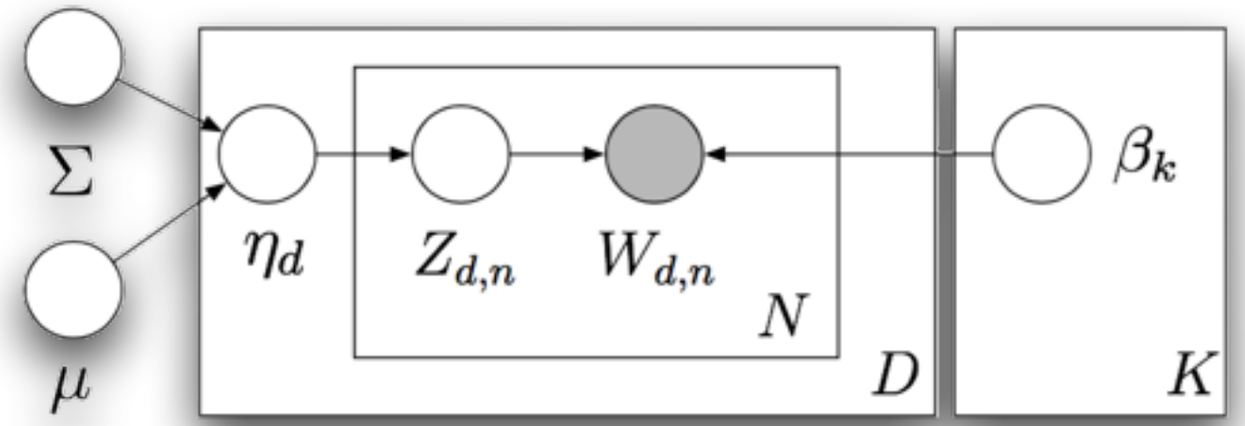
- **Results** in ‘*output/inferences/lda' folder*

librAIry

Outline

1. Artifacts
2. Parameters
3. Training
4. Evaluation and Interpretation
5. Inference
6. **Trends**
7. Domains
8. Topic-based Similarity

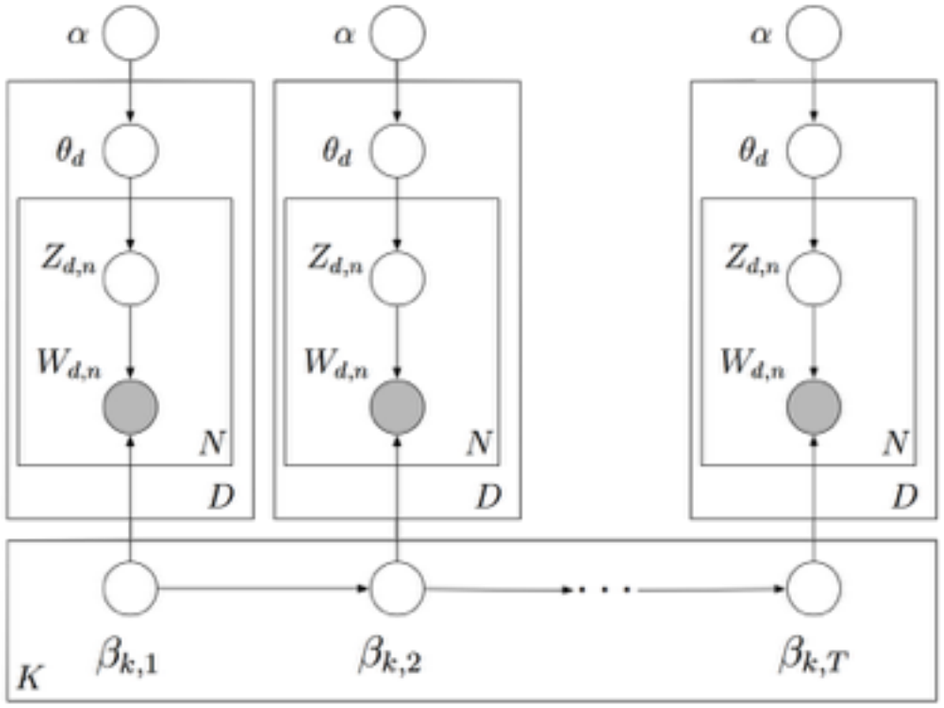
Correlated Topic Model (CTM) [Blei and Lafferty, 2006]

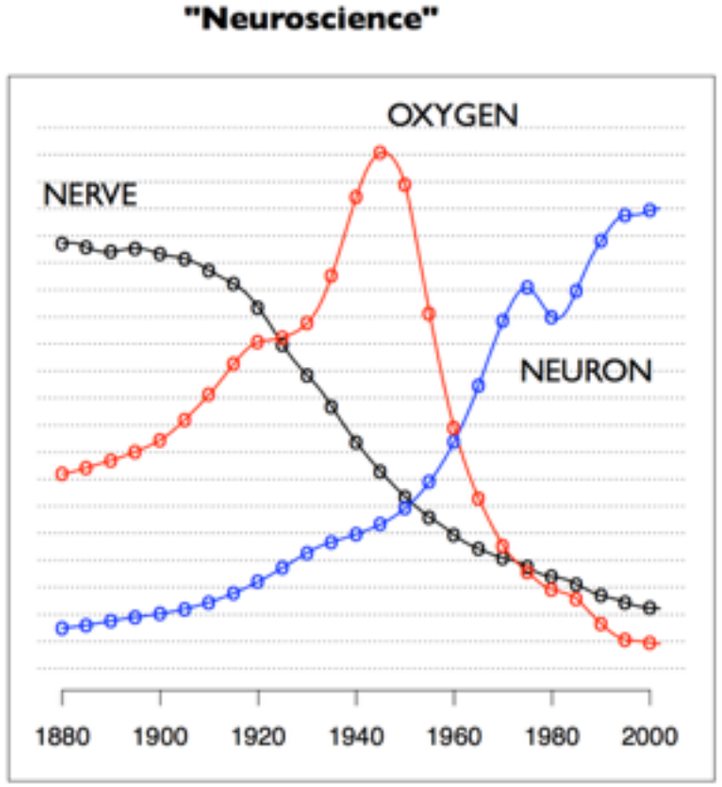


Correlated Topic Model (CTM)
uses a logistic normal distribution for the topic proportions instead of a Dirichlet distribution to allow representing correlations between topics

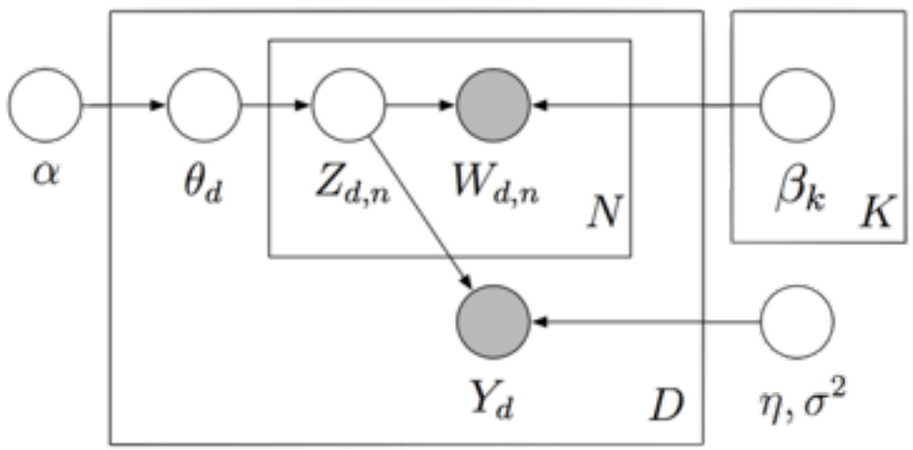
M, B. D., & D, L. J. (2006). Correlated Topic Models. Advances in Neural Information Processing Systems 18, 147–154

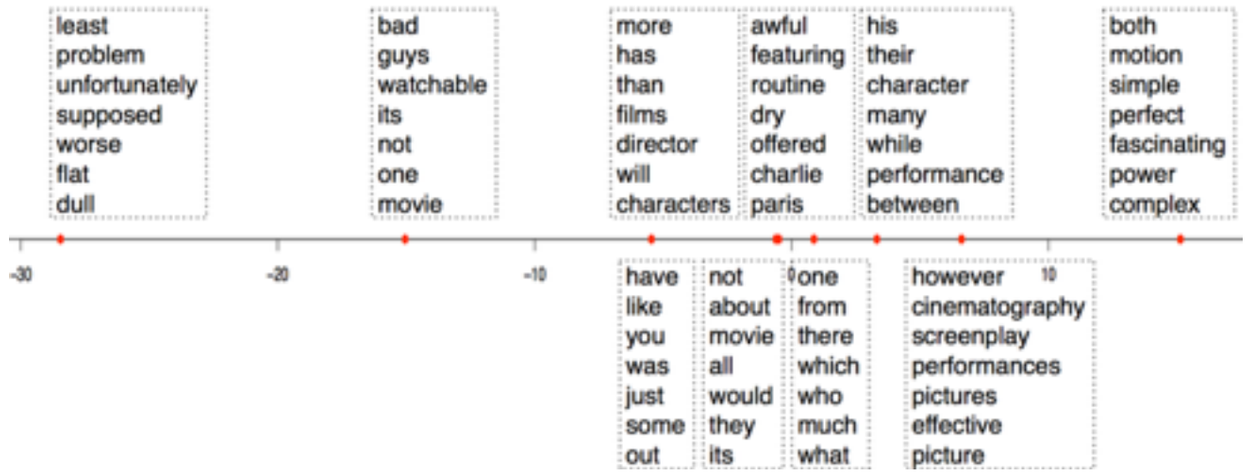
- *A **Dynamic Topic Model (DTM)** uses a logistic normal in a linear dynamic model to capture how topics change over time:*



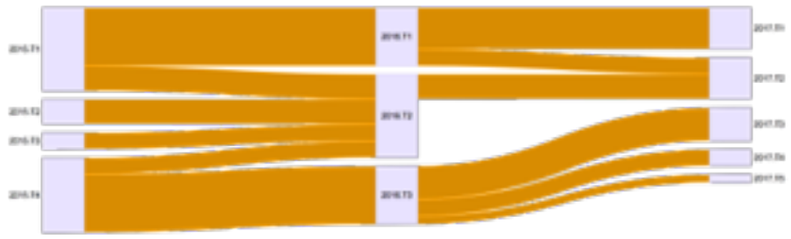


- ***Supervised LDA** are topic models of documents and responses, fit to find topics predictive of the response*





COMPARE TOPICS



- **Move** to the '*trends*' stage

```
$ git checkout trends
```

- **Adjust** *parameter.properties*

```
$ nano parameters.properties
```

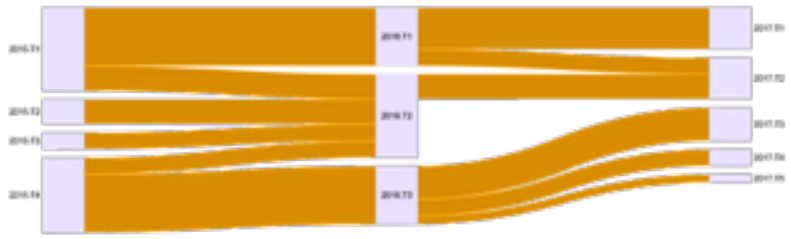
- **Execute**

```
$ docker start -i test
```

- **Results** in '*output/similarities/topics' folder*

librAIry

COMPARE TOPICS



- **See** the '*output/similarities/topics/ folder*:

```
$ similarities.csv    # pairwise topic similarities
$ topics.{domain}.txt # topic words in domain
```

Kumar, R., & Vassilvitskii, S. (2010). **Generalized distances between rankings**. Proceedings of the 19th International Conference on World Wide Web - WWW '10,(3), 571. http://doi.org/10.1145/1772690.1772749

librAIry

Outline

1. Artifacts
2. Parameters
3. Training
4. Evaluation and Interpretation
5. Inference
6. Trends
7. **Domains**
8. Topic-based Similarity

Topic Models in Historical Documents

- Finding themes in document that reflect **temporal trends**

- Looking not just for the topical of each time period, but how those topics shift in concentration as they are **influenced by historical events**

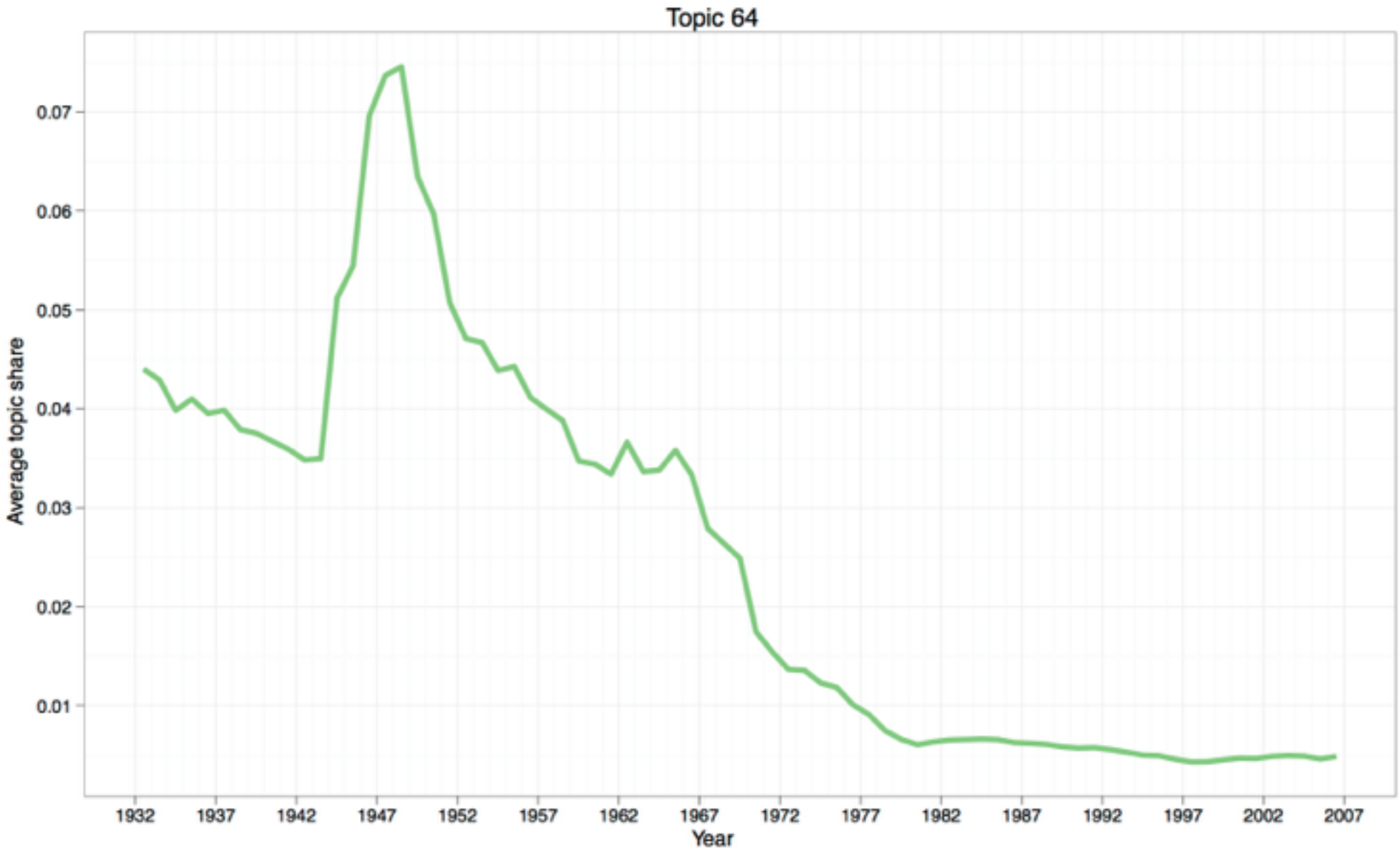
- Discovering how **events** are reflected in writing and how **ideas** and **emotions** emerge in response to changing events.

Historical Domain

Allen Beye Riddell. ***How to Read 22,198 Journal Articles: Studying the History of German Studies with Topic Models,*** *pages 91–113. Camden House, 2012.*

A LDA model trained with 150 topics on 22.198 articles from JSTOR in the 20th century, removing numbers and articles having fewer than 200 words

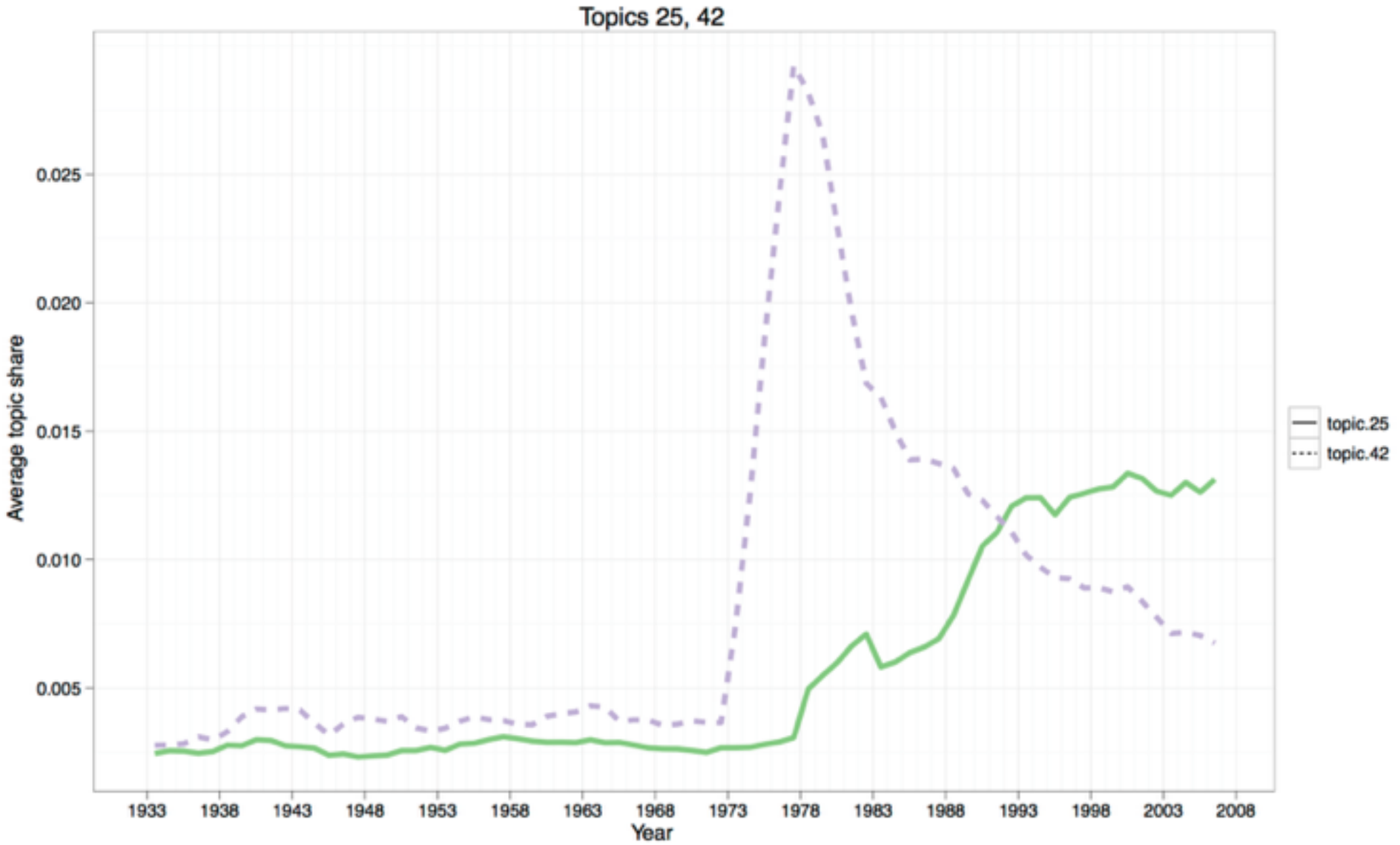
students language german student reading course class time teacher teaching read foreign method college material



Allen Beye Riddell. ***How to Read 22,198 Journal Articles: Studying the History of German Studies with Topic Models,*** *pages 91–113. Camden House, 2012.*

Topic 25: women female woman male feminist gender sexual feminine social role patriarchal movement sex roles masculine

Topic 42: social bourgeois class political critique society theory historical capitalist production marxist marx revolutionary capitalism economic



Conclusions

- the chronological trajectory of a topic is not the same thing as the trajectory of the individual words that compose it

- individual topics always need to be interpreted in the context of the larger model

- it becomes essential validate the description provided by a topic model by reference to something other than the topic model itself

Topic Models in Scientific Publications

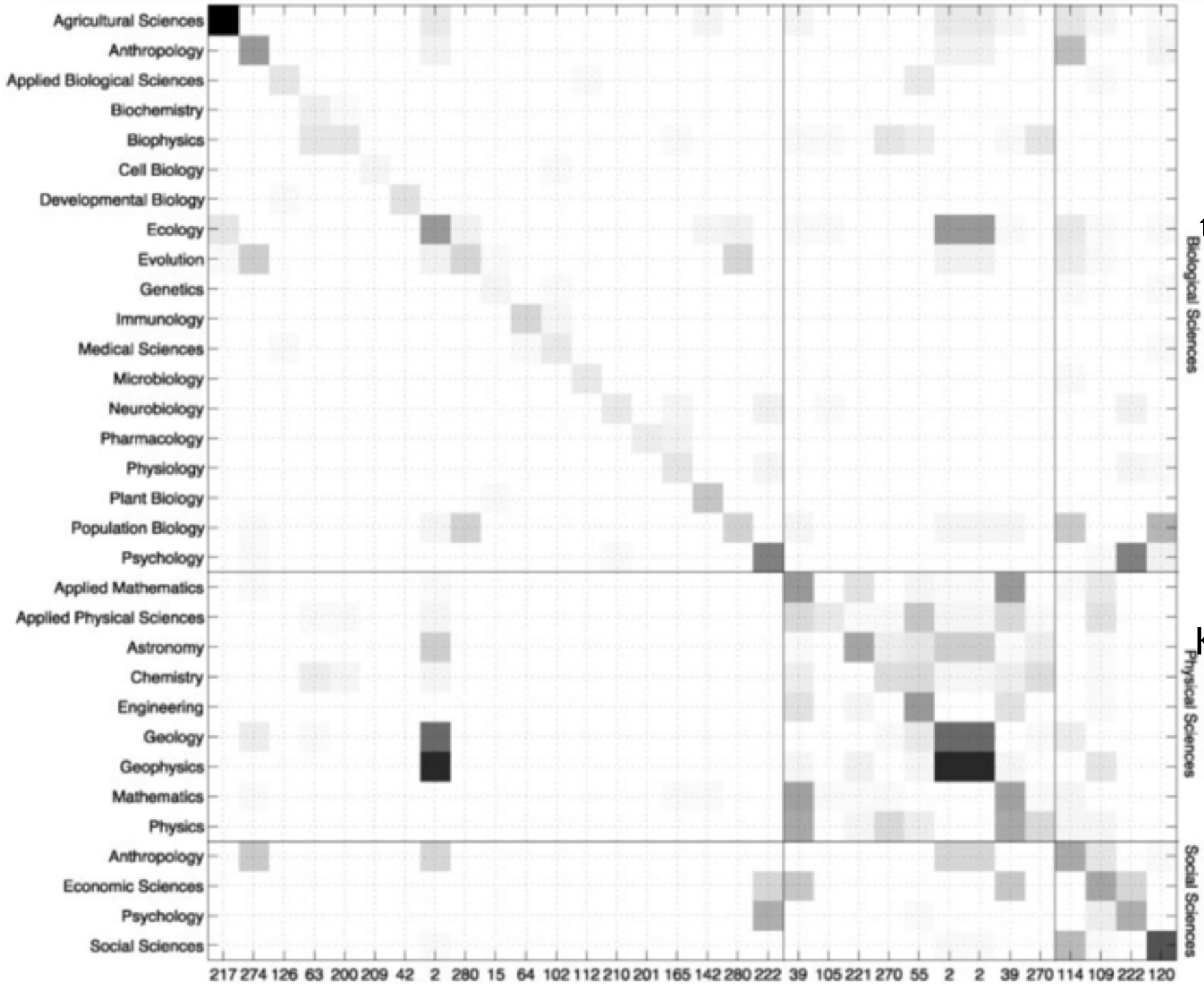
- **Specialized vocabularies** define fields of study

- Scientific documents **innovate**: unlike other domains, scientific documents are not just reports of news or events; they are news. Innovation is hard to detect and hard to attribute

- Science and **Policy**: understanding scientific publications is important for funding agencies, lawmakers, and the public.

Scientific Domain

Thomas L. Griffths and Mark Steyvers. ***Finding scientific topics.*** *Proceedings of the National Academy of Sciences, 101(Suppl 1):5228–5235, 2004*



Reconstruct the official *Proceedings of the National Academy of Sciences* (PNAS) topics automatically using topic models

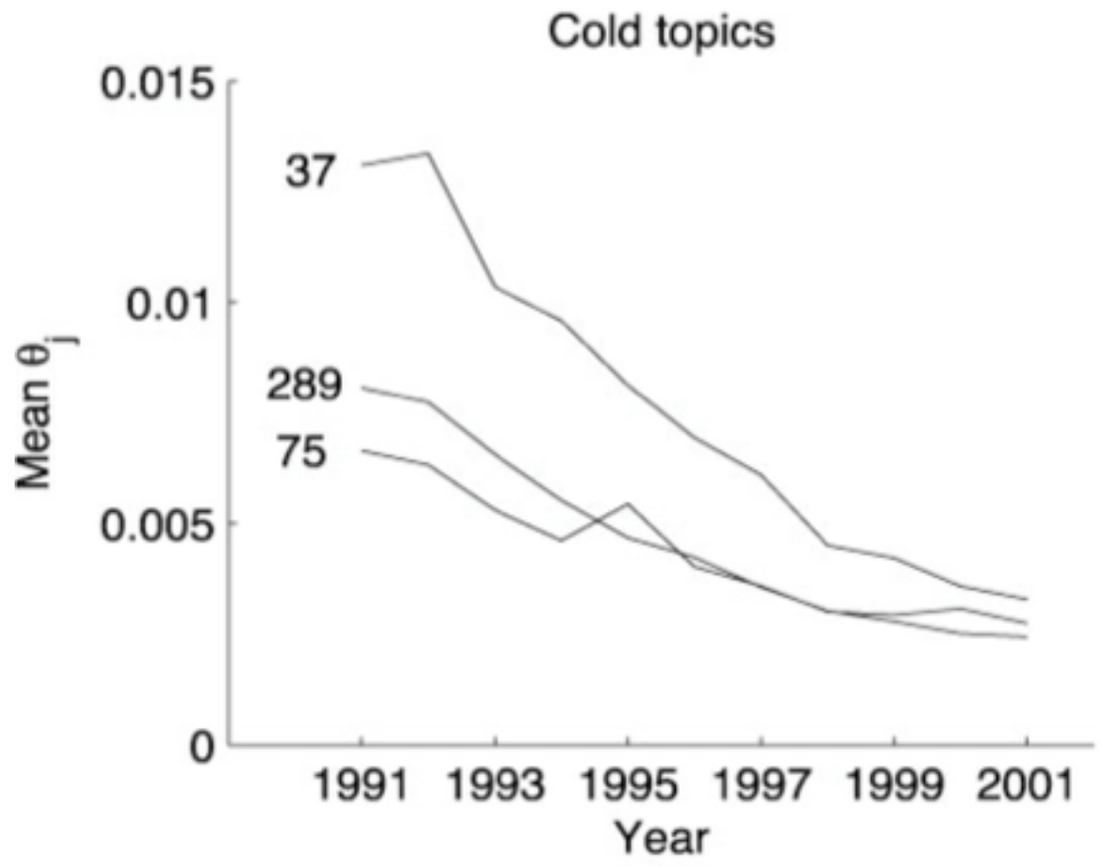
28.154 abstract published in PNAS from 1991 to 2001

LDA, beta=0.1, alpha=50/K, K=numTopics=50/100/200../1000

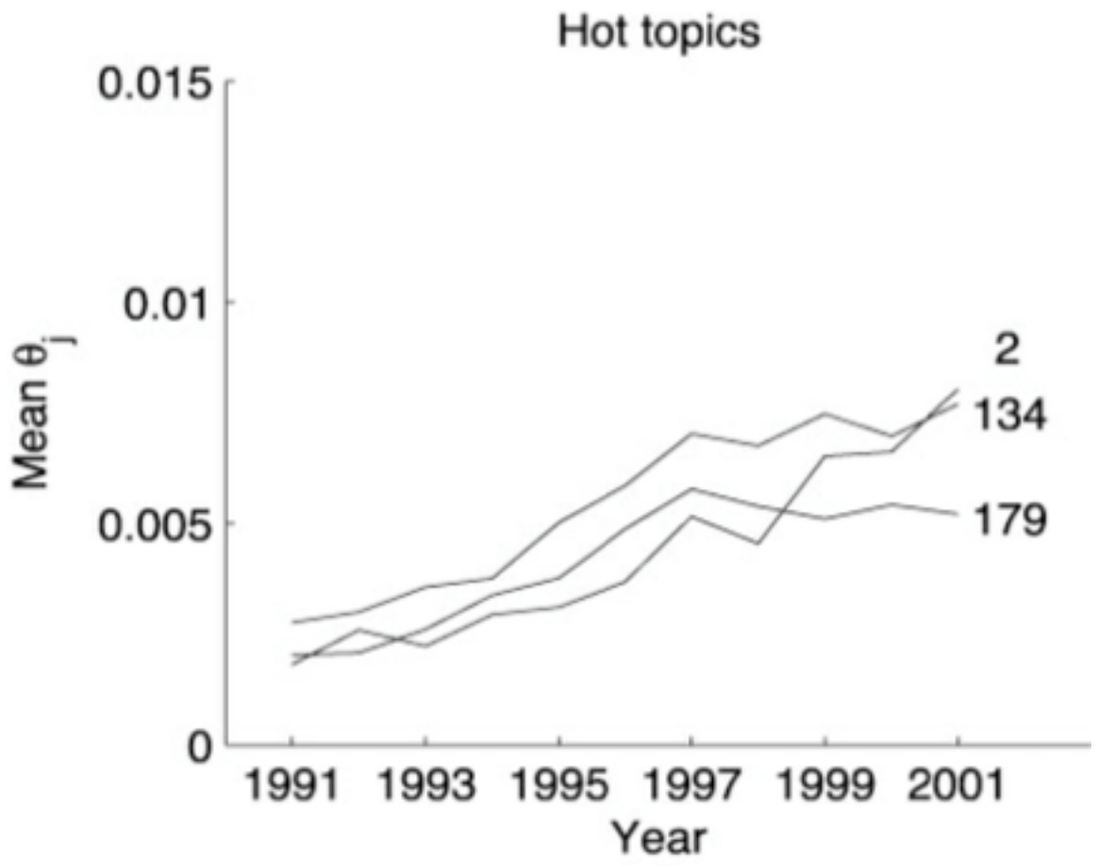
217
INSECT
MYB
PHEROMONE
LENS
LARVAE

274
SPECIES
PHYLOGENETIC
EVOLUTION
EVOLUTIONARY
SEQUENCES

Thomas L. Griffths and Mark Steyvers. ***Finding scientific topics.*** *Proceedings of the National Academy of Sciences, 101(Suppl 1):5228–5235, 2004*



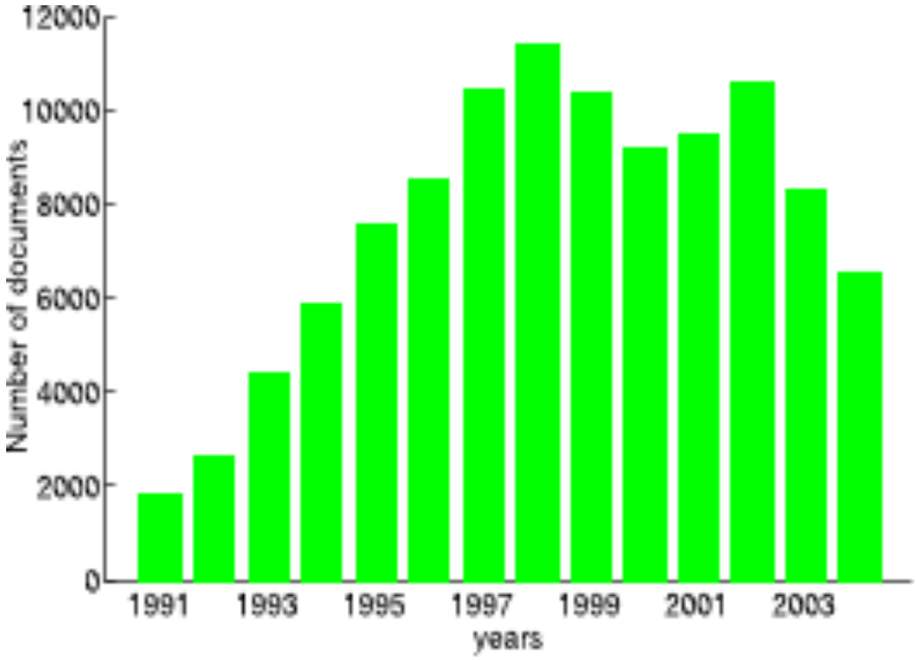
37	289	75
CDNA	KDA	ANTIBODY
AMINO	PROTEIN	ANTIBODIES
SEQUENCE	PURIFIED	MONOCLONAL
ACID	MOLECULAR	ANTIGEN
PROTEIN	MASS	IGG
ISOLATED	CHROMATOGRAPHY	MAB
ENCODING	POLYPEPTIDE	SPECIFIC
CLONED	GEL	EPITOPE
ACIDS	SDS	HUMAN
IDENTITY	BAND	MABS
CLONE	APPARENT	RECOGNIZED
EXPRESSED	LABELED	SERA



2	134	179
SPECIES	MICE	APOPTOSIS
GLOBAL	DEFICIENT	DEATH
CLIMATE	NORMAL	CELL
CO2	GENE	INDUCED
WATER	NULL	BCL
ENVIRONMENTAL	MOUSE	CELLS
YEARS	TYPE	APOPTOTIC
MARINE	HOMOZYGOUS	CASPASE
CARBON	ROLE	FAS
DIVERSITY	KNOCKOUT	SURVIVAL
OCEAN	DEVELOPMENT	PROGRAMMED
EXTINCTION	GENERATED	MEDIATED

Zhou, Ding, Xiang Ji, Hongyuan Zha and C. Lee Giles. ***Topic evolution and social interactions: how authors effect research****. CIKM (2006).*

given a seemingly new topic, from where does this topic evolve?
a newly emergent topic is truly new or rather a variation of an old topic?

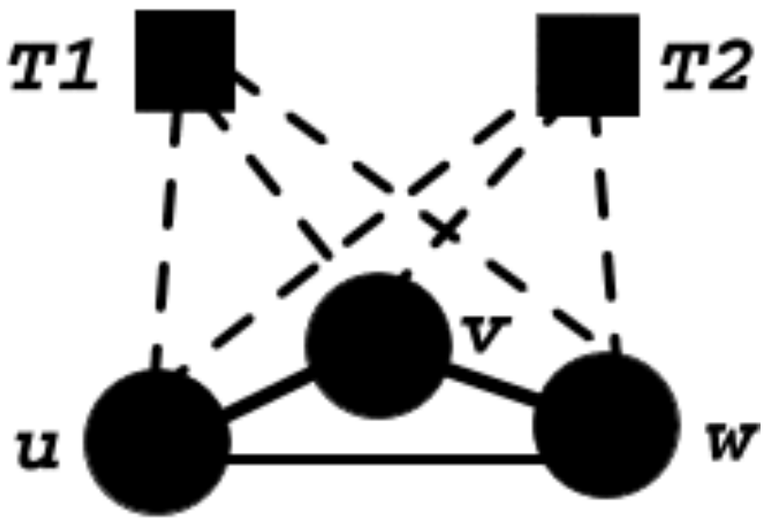


(a) Number of documents versus year acquired

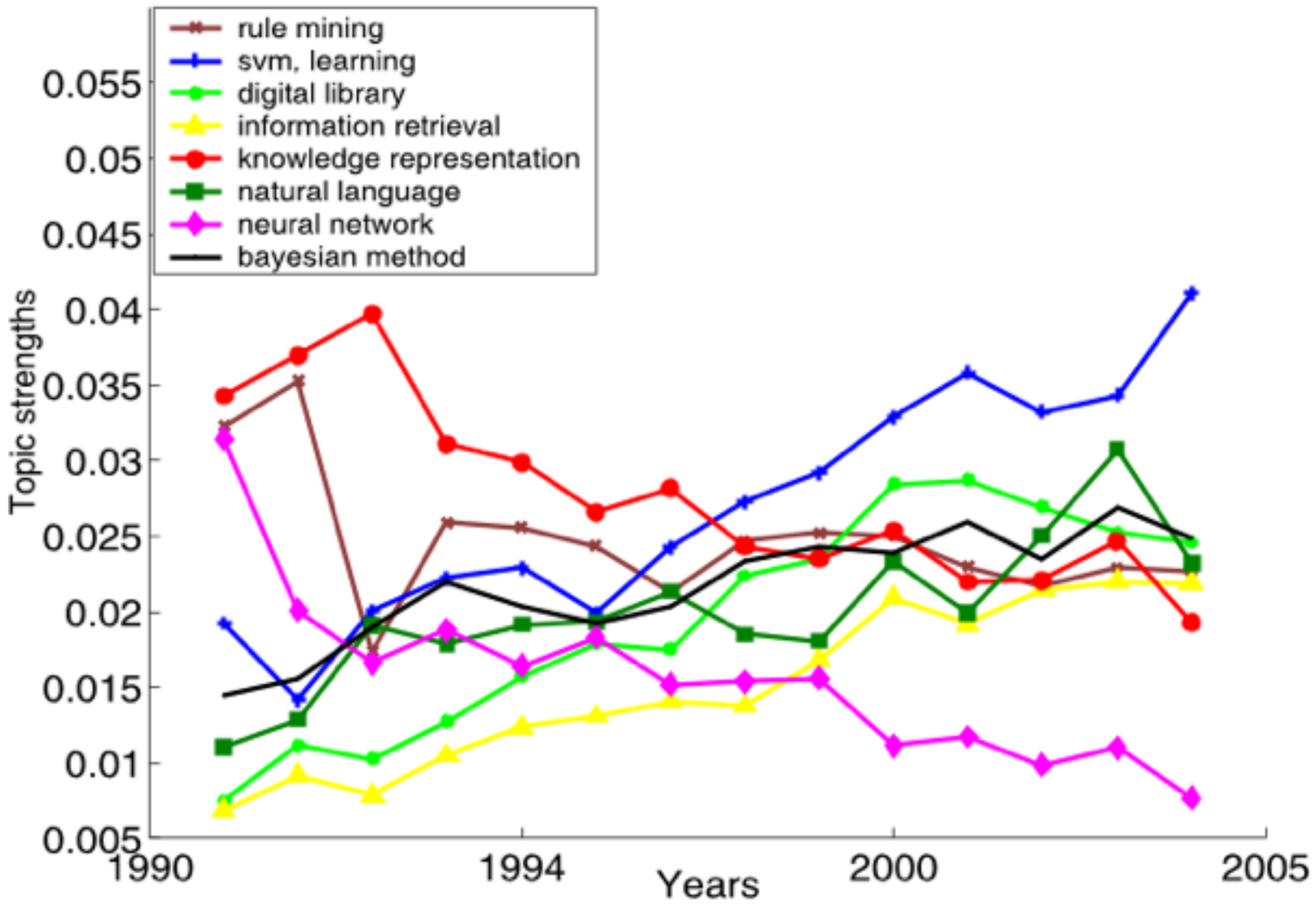
CiteSeer Dataset

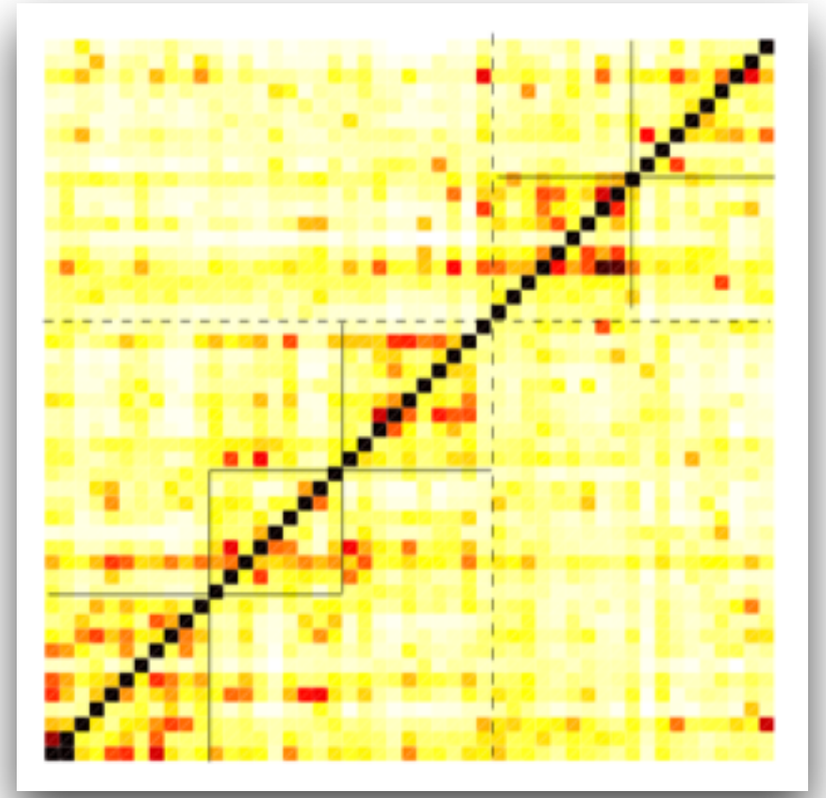
hypothesis

- *"one topic evolves into another topic when the corresponding social actors interact with other actors with different topics in the latent social network"*



Zhou, Ding, Xiang Ji, Hongyuan Zha and C. Lee Giles. ***Topic evolution and social interactions: how authors effect research.*** *CIKM (2006).*





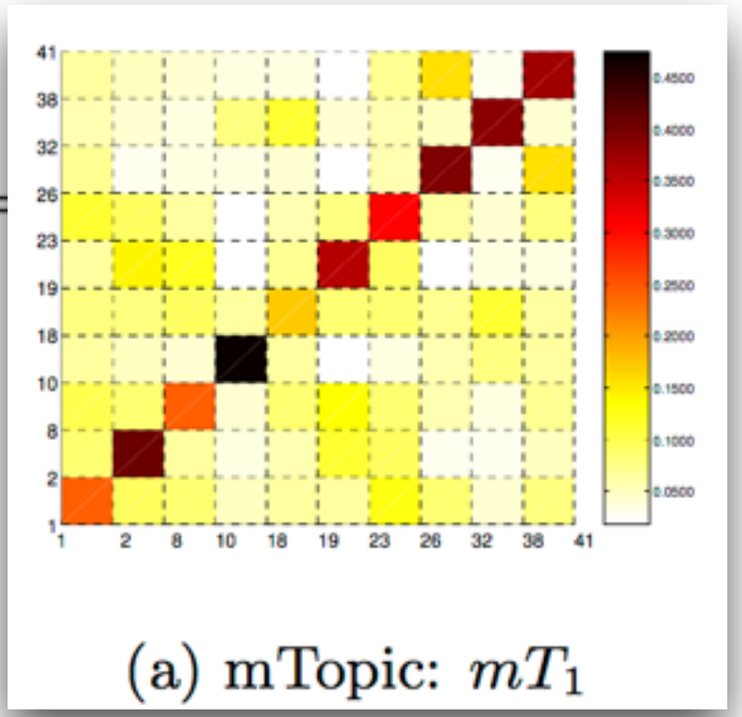
Topics with heavy methodology requirements (e.g. np problem, linear system) and/or popular topics (e.g. mobile computing, net- work) are more likely to remain stable. By contrast, topics closely related to applications are more likely to have higher transition probabilities than other topics (e.g. data mining in database, security) all things being equal

Zhou, Ding, Xiang Ji, Hongyuan Zha and C. Lee Giles. ***Topic evolution and social interactions: how authors effect research.*** *CIKM (2006).*

Table 2: Discovery of mTopics via block diagonal Markov transition matrix.

#	Topic IDs										
mT_1	1	2	8	10	18	19	23	26	32	38	41
mT_2	0	5	7	20	21	22	27	28	35	43	45
mT_3	6	25	30	36	37	40	44				
mT_4	13	14	15	17	24	31	33	39	42	47	48
mT_5	3	4	9	11	12	16	29	34	46	49	

mT_1	data management, data mining
mT_2	system, programming language, and architecture
mT_3	network and communication
mT_4	numerical analysis, machine learning
mT_5	statistical methods and applications



(a) mTopic: mT_1

Conclusions

- Understanding science communication allows us to see how our **understanding** of nature, technology, and engineering have advanced over the years

- Topic models can capture how these fields have **changed** and have **gained** additional knowledge with each new discovery

- As the scientific enterprise becomes **more distributed** and **faster** moving, these tools are important for scientists hoping to understand trends and development and for policy makers who seek to guide innovation

Topic Models in Literature

- **What is a document?**: Treating novels as a single bag of words does not work. Topics resulting from this corpus treatment are overly vague and lack thematic **coherence**

- **People and Places**: Because most works of fiction are set in imaginary worlds that do not exist outside the work itself, they have words such as character names that are extremely frequent locally but never occur elsewhere

- **Beyond the Literal:** Documents that are valued not just for their information content but for their **artistic expression**

- Topic models complement close reading in two ways, as a **survey** method and as a means for **tracing** and **comparing** large-scale patterns

Beyond the Literal

Rhody [2012] demonstrates on a corpus of poetry that although topics do not represent symbolic meanings, they are a good way of detecting the concrete language associated with repeated **metaphors**.

In a corpus of 4600 poems and a sixty-topic model, one of the topics discovered:

{night, light, moon, stars, day, dark, sun, sleep, sky, wind, time, eyes, star, darkness, bright }

Analyzing the context of the topic (i.e. poems), they are using a metaphor relating night and sleep to death.

Thus, the topic can be considered as the death as metaphor.

Lia M. Rhody. ***Topic modeling and figurative language****. Journal of Digital Humanities, 2(1), 2012*

Conclusions

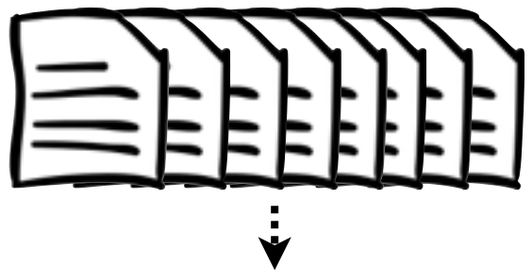
- Topic models cannot by themselves study literature, but they are useful tools for scholars studying literature

- Literary concepts are complicated, but they often have surprisingly strong statistical signatures

- Models can still be useful in identifying areas of potential interest, even if they don't "understand" what they are finding

1. Artifacts
2. Parameters
3. Training
4. Evaluation and Interpretation
5. Inference
6. Trends
7. Domains
8. **Topic-based Similarity**

Topic-based Similarity



Topic Model

..

DOCUMENT 'A'

DOCUMENT 'B'

topic0: 0.520,
topic1: 0.327
topic2: 0.081
…
topic122: 0.182

Similarity Value

0.595

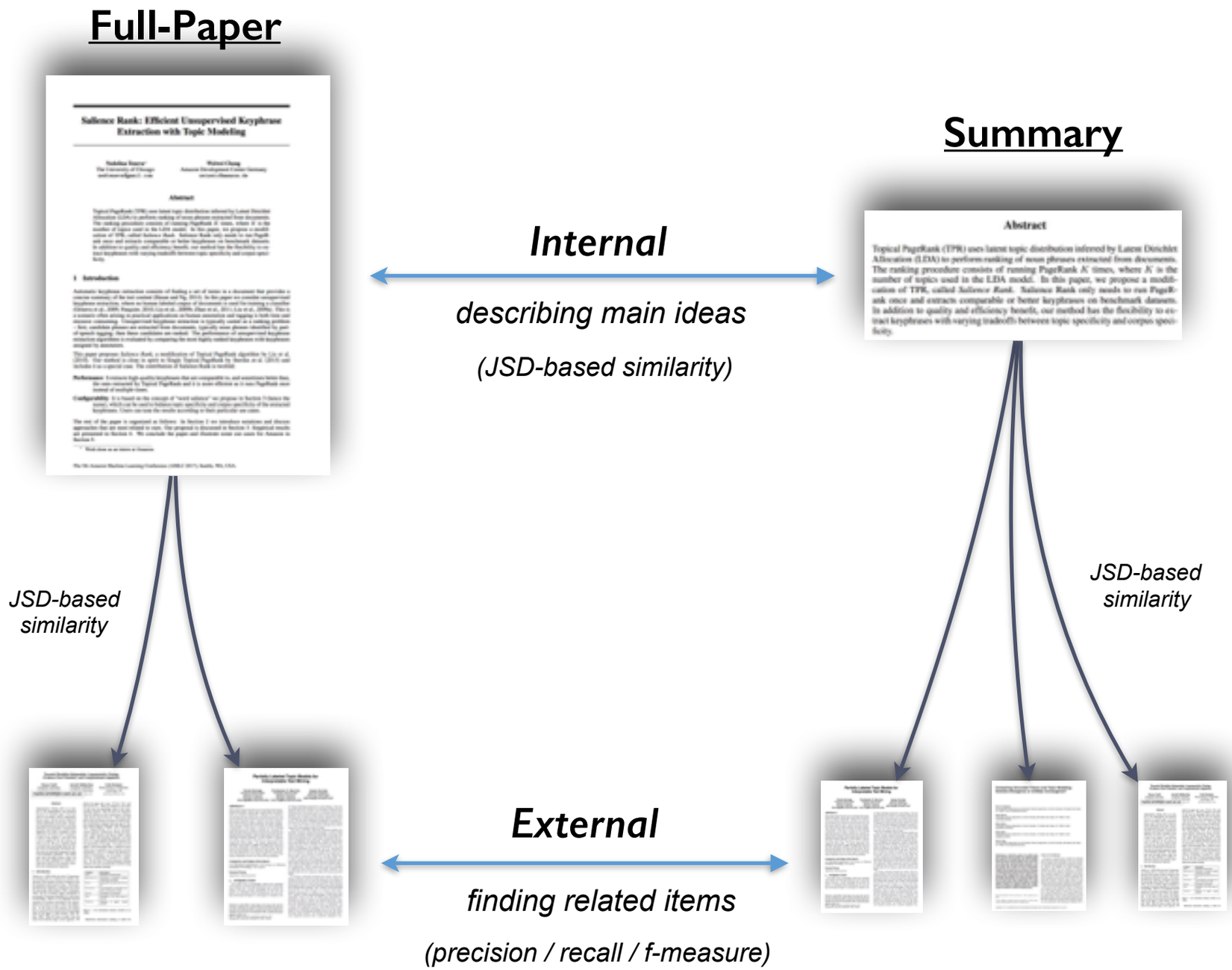
Jensen-Shannon Divergence

$$JSD(p,q) = \sum_{i=1}^{T} p_i * \log \frac{2 * p_i}{p_i + q_i} + \sum_{i=1}^{T} q_i * \log \frac{2 * q_i}{q_i + p_i}$$

topic1: 0.573,
topic2: 0.172
topic3: 0.136
…
topic122: 0.099

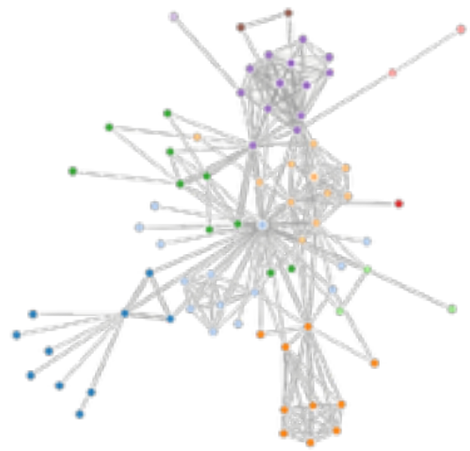
Representativeness

librAIry



Badenes-Olmedo, C., Redondo-Garcia, J. L., & Corcho, O. (2017). **An Initial Analysis of Topic-based Similarity among Scientific Documents based on their rhetorical discourse parts**. In Proceedings of the 1st SEMSCI workshop co-located with ISWC.

RELATIONS

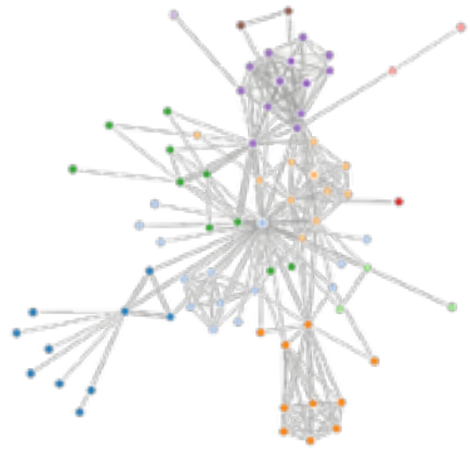


- **kcap**:

[GET] <librairy-api>/domains/{domainId}/items/{itemId}/relations

librAIry

RELATIONS



- **Move** *to the 'similarity' stage:*

```
$ git checkout similarity
```

- **Adjust** *parameter.properties*

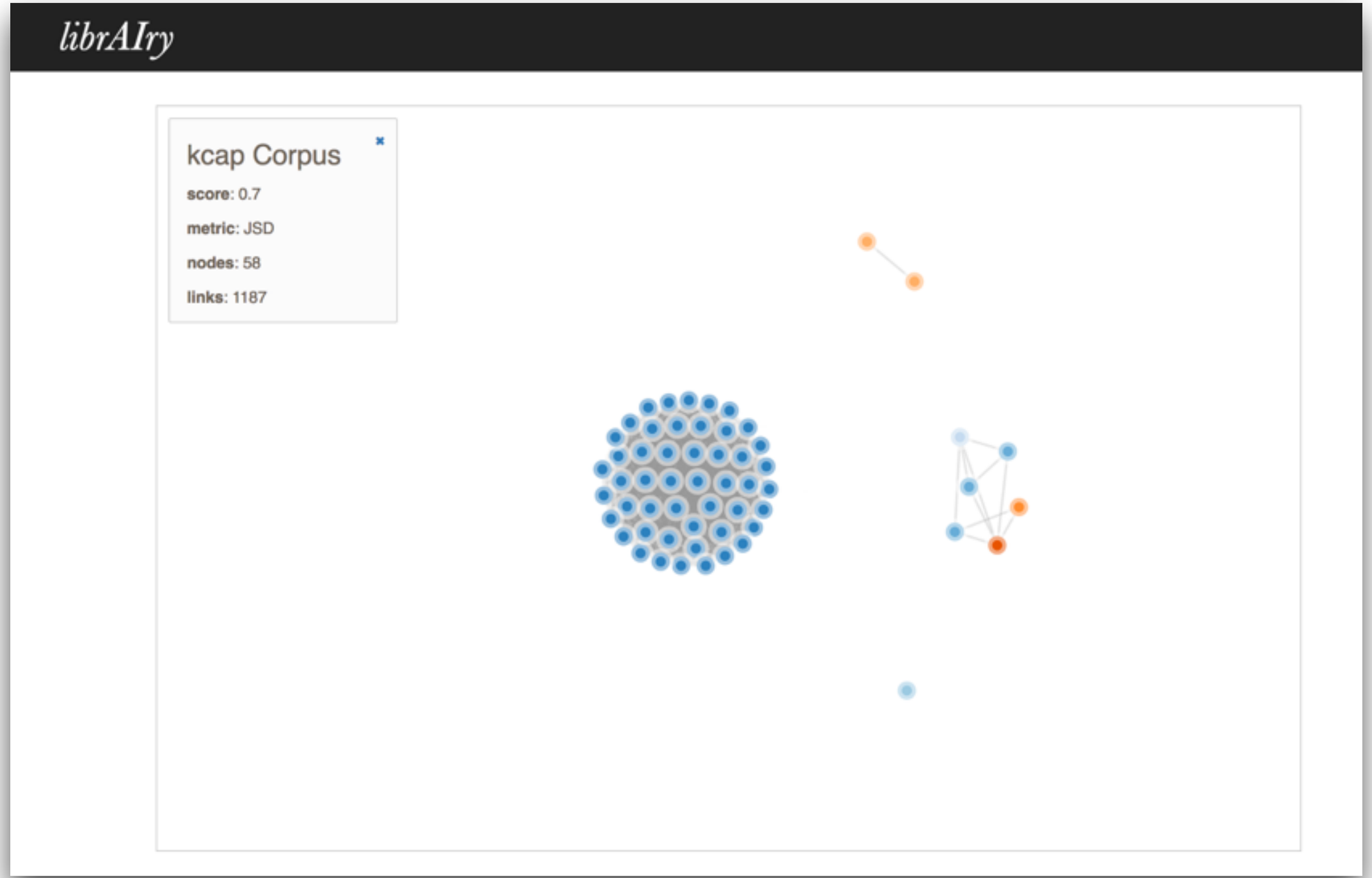
```
$ nano parameters.properties
```

- **Execute**

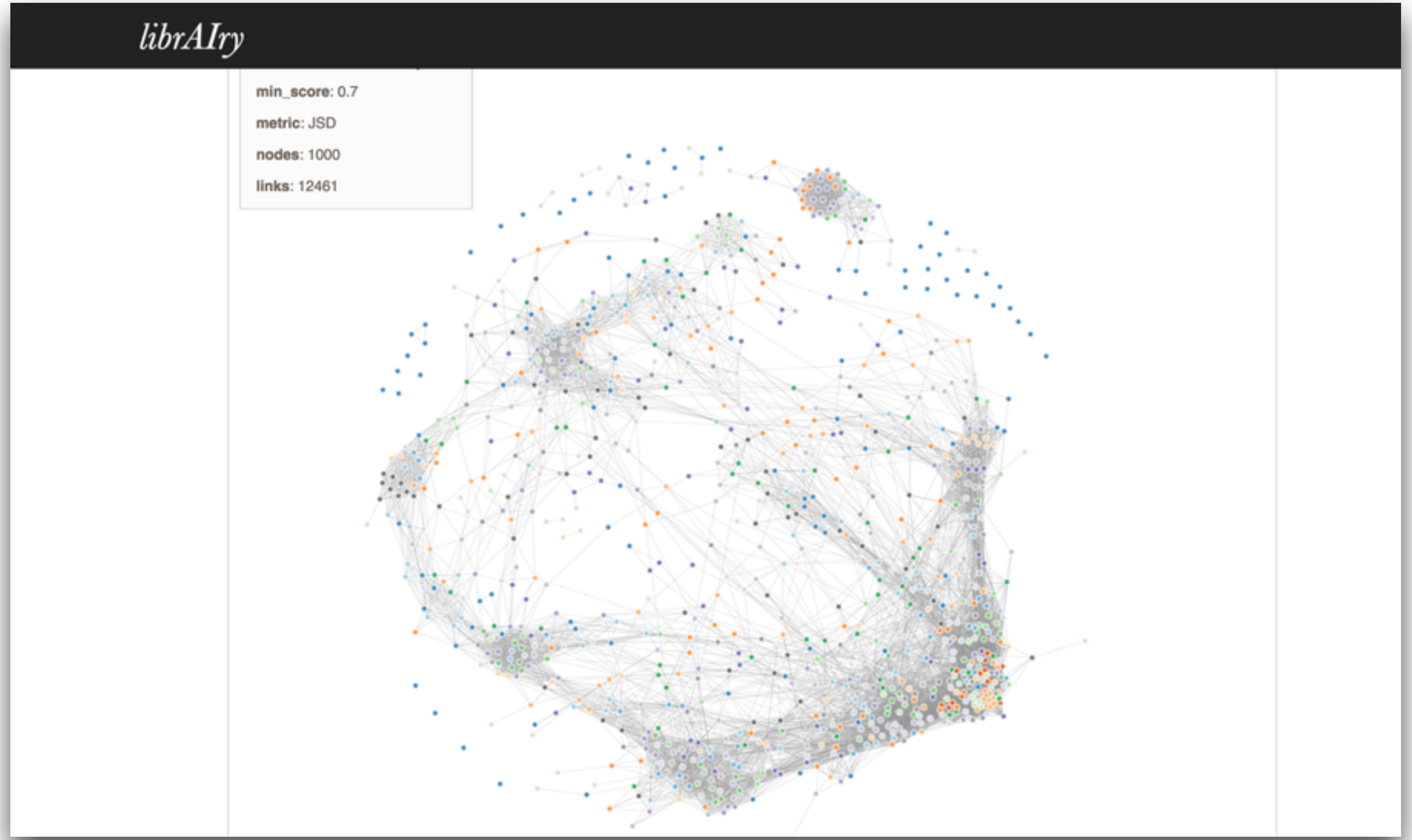
```
$ docker start -i test
```

- **Results** in '*output/models/similarities' folder*

- **Open** *'similarity-network.html' in a browser (Firefox)*



- **Try** with *input.domain=ecommerce (1000 nodes)*



References

- Jurafsky, D., & Martin, J. H. (2016). **Language Modeling with N- grams**. Speech and Language Processing, 1–28.
- Jordan Boyd-Graber, Yuening Hu and David Mimno (2017), "**Applications of Topic Models**", Foundations and Trends® in Information Retrieval: Vol. 11: No. 2-3, pp 143-296
- Blei, David M., Lawrence Carin and David B. Dunson. "**Probabilistic Topic Models**." IEEE Signal Processing Magazine 27 (2010): 55-65.
- Zhai, ChengXiang. "**Probabilistic Topic Models for Text Data Retrieval and Analysis**." SIGIR (2017).
- Wang, C., Blei, D., & Heckerman, D. (2008). **Continuous Time Dynamic Topic Models**. Proc of UAI, 579–586.
- Chang, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). **Reading Tea Leaves: How Humans Interpret Topic Models**. Advances in Neural Information Processing Systems 22, 288--296.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). **Latent Dirichlet Allocation**. Journal of Machine Learning Research, 3(4–5), 993–1022.
- Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). BTM: Topic Modeling over Short Texts. Knowledge and Data Engineering, IEEE Transactions
- Blei, D. M., & Lafferty, J. D. (2007). **A correlated topic model of Science**. The Annals of Applied Statistics, 1(1), 17–35.
- Griffiths, T. L., Jordan, M. I., Tenenbaum, J. B., & Blei, D. M. (2004). **Hierarchical Topic Models and the Nested Chinese Restaurant Process**. Advances in Neural Information Processing Systems, 17–24.
- Badenes-Olmedo, C.; Redondo-Garcia, J.; Corcho, O. (2017) **Distributing Text Mining tasks with librAIry.** In Proceedings of the 2017 ACM Symposium on Document Engineering (DocEng '17). ACM, 63-66.

Probabilistic Topic Models

THANKS!

Carlos Badenes-Olmedo
Oscar Corcho

Ontology Engineering Group (OEG)
Universidad Politécnica de Madrid (UPM)

K-CAP 2017
Knowledge Capture
December 4th-6th, 2017
Austin, Texas, United States

ocorcho@fi.upm.es
@ocorcho
oeg-upm.net
github.com/librairy