Day 2: Lecture 1

John Paisley

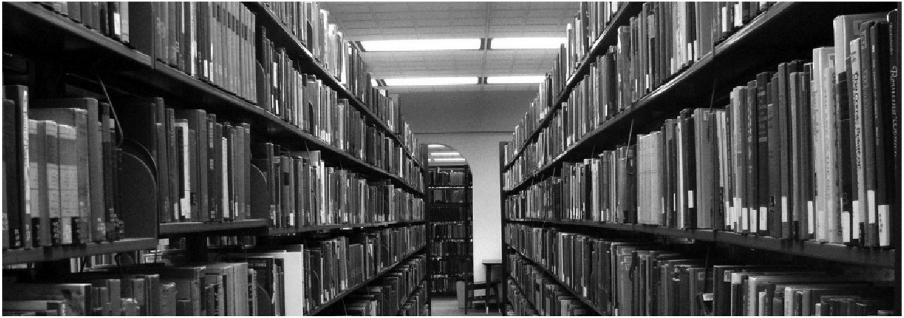
Columbia University

Duke-Tsinghua Machine Learning Summer School

July 26, 2017

Given text we may want to:

- Organize
- Visualize
- Summarize
- Search
- Predict
- Understand



Topic modeling provides one approach to these tasks.

1. Discovers the thematic structure in text.
2. Annotates the documents according to themes.
3. Use annotations to visualize, organize, summarize, etc.

The New York Times | http://nyti.ms/WO91dx

BUSINESS DAY

A Digital Shift on Health Data Swells Profits in an Industry

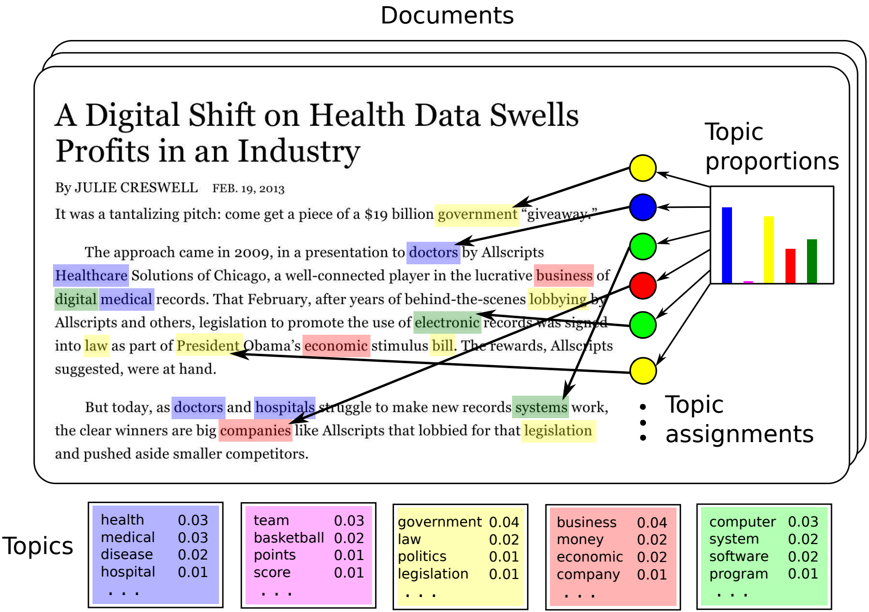
By JULIE CRESWELL FEB. 19, 2013

It was a tantalizing pitch: come get a piece of a $19 billion government “giveaway.”

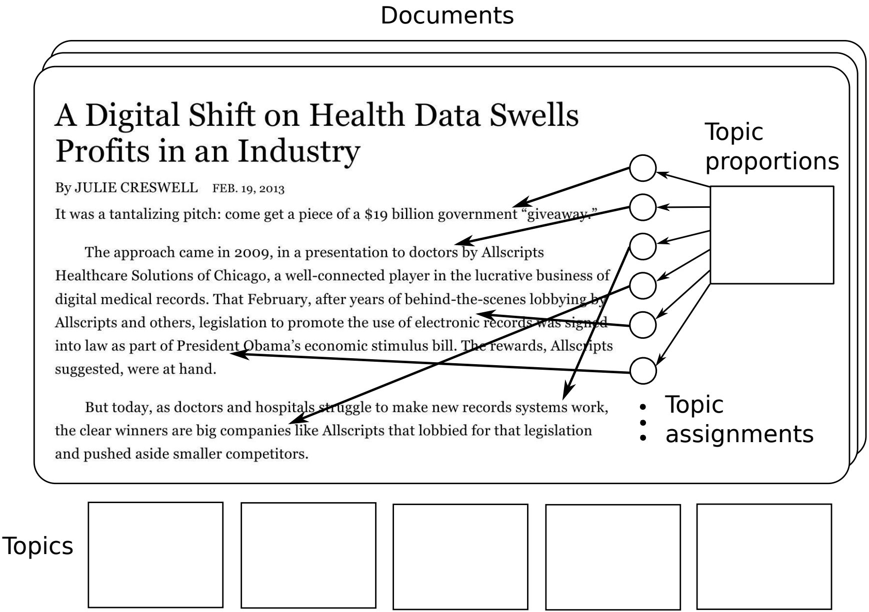
The approach came in 2009, in a presentation to doctors by Allscripts Healthcare Solutions of Chicago, a well-connected player in the lucrative business of digital medical records. That February, after years of behind-the-scenes lobbying by Allscripts and others, legislation to promote the use of electronic records was signed into law as part of President Obama’s economic stimulus bill. The rewards, Allscripts suggested, were at hand.

But today, as doctors and hospitals struggle to make new records systems work, the clear winners are big companies like Allscripts that lobbied for that legislation and pushed aside smaller competitors.

Documents exhibit multiple topics



Topic Modeling



Topic Modeling

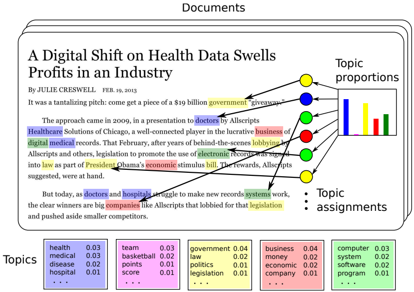
There are three key ingredients to any topic model.

1. **Topics:** Probability distributions on vocabulary.
2. **Topic proportions:** Probability distributions on the topics.
3. **Topic assignments:** Assigns each observed word to a topic.

Topics are **global** variables. All documents share the same topics.

Topic proportions are **local** variables. They change with each document.

Topic assignments are also local and help us learn the first two.



Most topic models are **bag-of-words** models. This means that *which* words are contained in the document matters, but their *order* does not.

The New York Times | http://nyti.ms/WO91dx

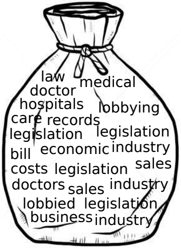
BUSINESS DAY

A Digital Shift on Health Data Swells Profits in an Industry

By JULIE CRESWELL FEB. 19, 2013

It was a tantalizing pitch: come get a piece of a $19 billion government "giveaway."

The approach came in 2009, in a presentation to doctors by Allscripts Healthcare Solutions of Chicago, a well-connected player in the lucrative business of digital medical records. That February, after years of behind-the-scenes lobbying by Allscripts and others, legislation to promote the use of electronic records was signed into law as part of President Obama's economic stimulus bill. The rewards, Allscripts suggested, were at hand.



The New York Times | http://nyti.ms/2ado9QP

PRO BASKETBALL

N.B.A. to Move All-Star Game From North Carolina

By SCOTT CACCIOLA and ALAN BLINDER JULY 21, 2016

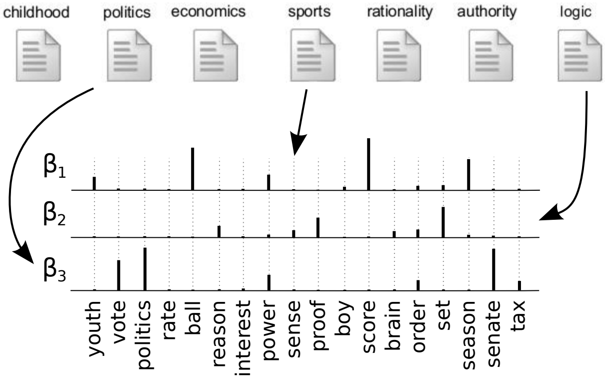
The National Basketball Association on Thursday dealt a blow to the economy and prestige of North Carolina by pulling next February's All-Star Game from Charlotte to protest a state law that eliminated anti-discrimination protections for lesbian, gay, bisexual and transgender people.

Part I

Latent Dirichlet Allocation

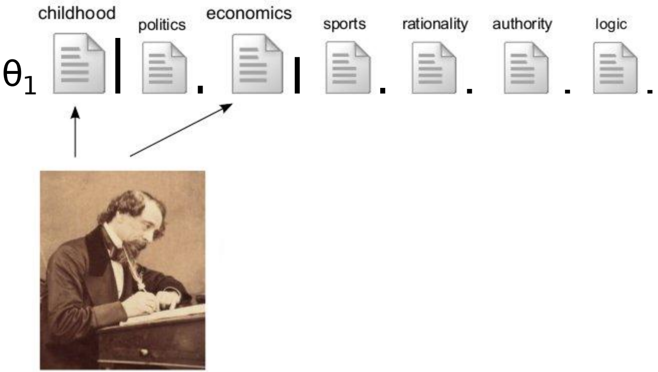
As with other topic models, LDA has

1. A collection of distributions on words called topics.
2. A distribution on topics for each document.



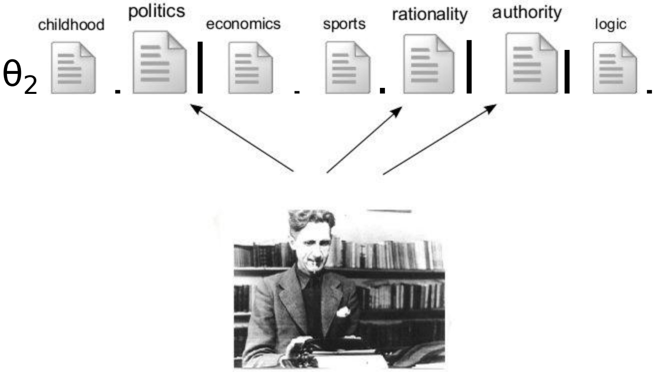
As with other topic models, LDA has

1. A collection of distributions on words called topics.
2. A distribution on topics for each document.



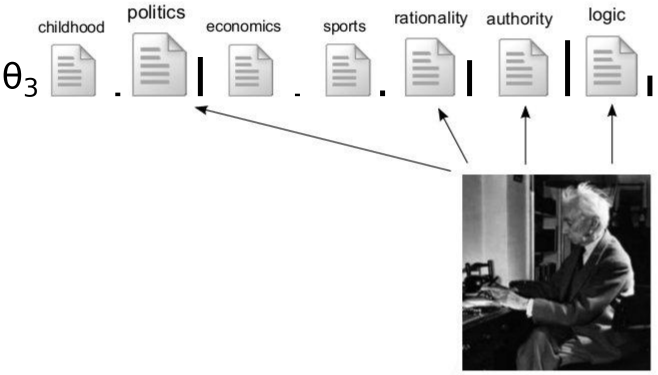
As with other topic models, LDA has

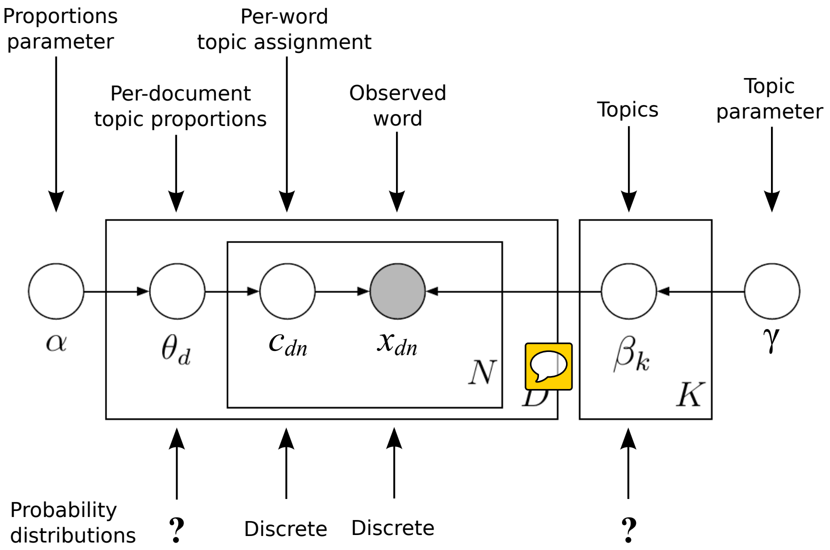
1. A collection of distributions on words called topics.
2. A distribution on topics for each document.



As with other topic models, LDA has

1. A collection of distributions on words called topics.
2. A distribution on topics for each document.



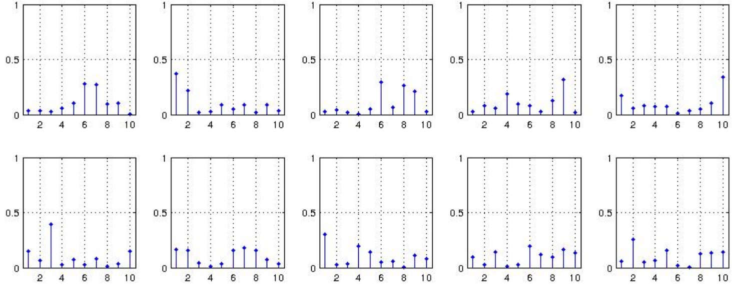

Proportions parameter
Per-word topic assignment
Per-document topic proportions
Observed word
Topics
Topic parameter
α
θ_d
c_{dn}
x_{dn}
N
D
β_k
K
γ
Probability distributions
?
Discrete
Discrete
?

Dirichlet distribution: A continuous distribution on discrete probability vectors. Let β_k be a probability vector and γ a positive parameter vector,

$$p(\beta_k|\gamma) = \frac{\Gamma(\sum_v \gamma_v)}{\prod_{v=1}^{V} \Gamma(\gamma_v)} \prod_{v=1}^{V} \beta_{k,v}^{\gamma_v - 1}$$

This defines the Dirichlet distribution. Some examples of β_k generated from this distribution for a constant value of γ and $V = 10$ are given below.

$\gamma = 1$

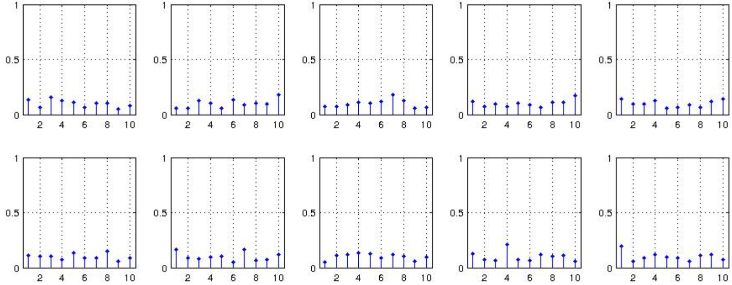


Dirichlet distribution: A continuous distribution on discrete probability vectors. Let β_k be a probability vector and γ a positive parameter vector,

$$p(\beta_k|\gamma) = \frac{\Gamma(\sum_v \gamma_v)}{\prod_{v=1}^{V} \Gamma(\gamma_v)} \prod_{v=1}^{V} \beta_{k,v}^{\gamma_v - 1}$$

This defines the Dirichlet distribution. Some examples of β_k generated from this distribution for a constant value of γ and $V = 10$ are given below.

$\gamma = 10$

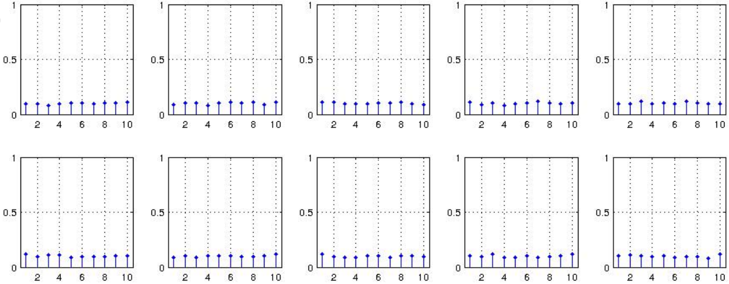


Dirichlet distribution: A continuous distribution on discrete probability vectors. Let β_k be a probability vector and γ a positive parameter vector,

$$p(\beta_k|\gamma) = \frac{\Gamma(\sum_v \gamma_v)}{\prod_{v=1}^{V} \Gamma(\gamma_v)} \prod_{v=1}^{V} \beta_{k,v}^{\gamma_v - 1}$$

This defines the Dirichlet distribution. Some examples of β_k generated from this distribution for a constant value of γ and $V = 10$ are given below.

$\gamma = 100$

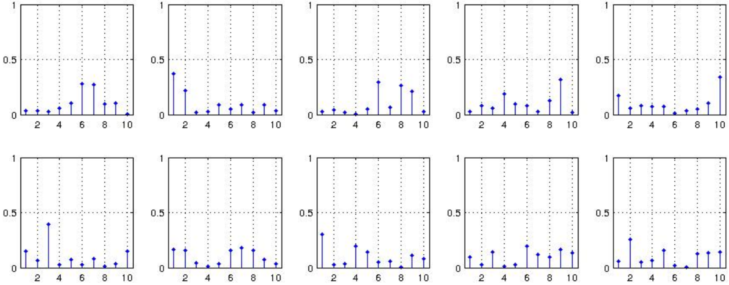


Dirichlet distribution: A continuous distribution on discrete probability vectors. Let β_k be a probability vector and γ a positive parameter vector,

$$p(\beta_k|\gamma) = \frac{\Gamma(\sum_v \gamma_v)}{\prod_{v=1}^{V} \Gamma(\gamma_v)} \prod_{v=1}^{V} \beta_{k,v}^{\gamma_v - 1}$$

This defines the Dirichlet distribution. Some examples of β_k generated from this distribution for a constant value of γ and $V = 10$ are given below.

$\gamma = 1$

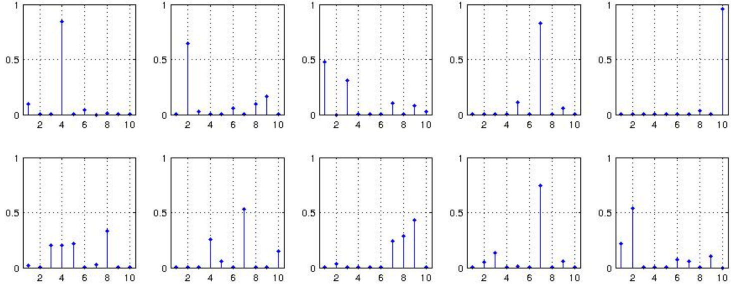


Dirichlet distribution: A continuous distribution on discrete probability vectors. Let β_k be a probability vector and γ a positive parameter vector,

$$p(\beta_k|\gamma) = \frac{\Gamma(\sum_v \gamma_v)}{\prod_{v=1}^{V} \Gamma(\gamma_v)} \prod_{v=1}^{V} \beta_{k,v}^{\gamma_v - 1}$$

This defines the Dirichlet distribution. Some examples of β_k generated from this distribution for a constant value of γ and $V = 10$ are given below.

$\gamma = 0.1$

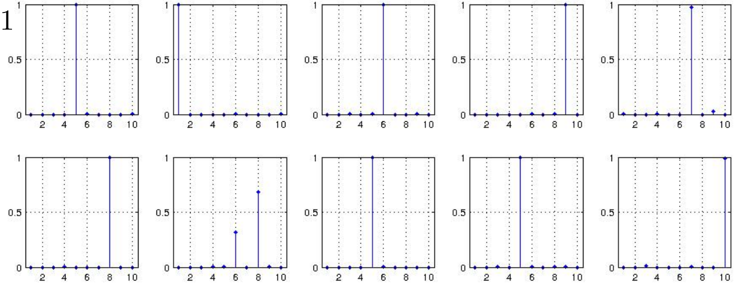


Dirichlet distribution: A continuous distribution on discrete probability vectors. Let β_k be a probability vector and γ a positive parameter vector,

$$p(\beta_k|\gamma) = \frac{\Gamma(\sum_v \gamma_v)}{\prod_{v=1}^{V} \Gamma(\gamma_v)} \prod_{v=1}^{V} \beta_{k,v}^{\gamma_v - 1}$$

This defines the Dirichlet distribution. Some examples of β_k generated from this distribution for a constant value of γ and $V = 10$ are given below.

$\gamma = 0.01$



As with any Bayesian model, we have to define the process for generating data and hidden model variables before we can learn them.

Generative process for LDA

1. Generate each topic — a distribution on words in a vocabulary

$$\beta_k \sim \text{Dirichlet}(\gamma), \quad k = 1, \ldots, K$$

2. For each document, generate a distribution on topics

$$\theta_d \sim \text{Dirichlet}(\alpha), \quad d = 1, \ldots, D$$

3. For the nth word in the dth document,
   a) Allocate the word to a topic, $c_{dn} \sim \text{Discrete}(\theta_d)$
   b) Generate the word from the selected topic, $x_{dn} \sim \text{Discrete}(\beta_{c_{dn}})$

How do we know what these are? All we have is the data.

Original documents

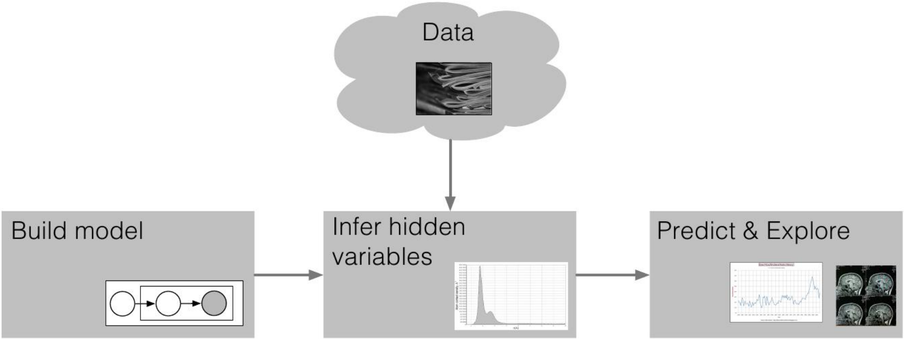
perspective identifying tumor suppressor genes in human...
letters global warming report leslie roberts article global....
research news a small revolution gets under way the 1990s....
a continuing series the reign of trial and error draws to a close...
making deep earthquakes in the laboratory lab experimenters...
quick fix for freeways thanks to a team of fast working...
feathers fly in grouse population dispute researchers...

...

Word index and counts

1897:1 1467:1 1351:1 731:2 800:5 682:1 315:6 3668:1 14:1
4261:2 518:1 271:6 2734:1 2662:1 2432:1 683:2 1631:7
2724:1 107:3 518:1 141:3 3208:1 32:1 2444:1 182:1 250:1
2552:1 1993:1 116:1 539:1 1630:1 855:1 1422:1 182:3 2432:1
1372:1 1351:1 261:1 501:1 1938:1 32:1 14:1 4067:1 98:2
4384:1 1339:1 32:1 4107:1 2300:1 229:1 529:1 521:1 2231:1
569:1 3617:1 3781:2 14:1 98:1 3596:1 3037:1 1482:12 665:2

...



- LDA discovers themes through posterior inference.
  - We have defined an “appropriate” model for the data.
  - Now we want to learn that model.
- We then use these learned values for tasks we care about.

The New York Times

music	book	art	game	show
band	life	museum	knicks	film
songs	novel	show	nets	television
rock	story	exhibition	points	movie
album	books	artist	team	series
jazz	man	artists	season	says
pop	stories	paintings	play	life
song	love	painting	games	man
singer	children	century	night	character
night	family	works	coach	know

theater	clinton	stock	restaurant	budget
play	bush	market	sauce	tax
production	campaign	percent	menu	governor
show	gore	fund	food	county
stage	political	investors	dishes	mayor
street	republican	funds	street	billion
broadway	dole	companies	dining	taxes
director	presidential	stocks	dinner	plan
musical	senator	investment	chicken	legislature
directed	house	trading	served	fiscal

Why does LDA "work"?

- LDA trades off two goals.

  1. In each **document**, allocate its words to a **few topics**.
  2. In each **topic**, assign high probability to a **few words**.

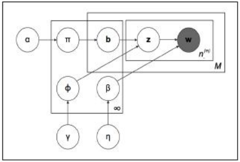
- These goals are competing with one another.

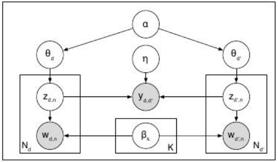
  - Putting a document in a single topic makes #2 hard:
    All of its words must have probability under that topic.
  - Putting very few words in each topic makes #1 hard:
    To cover a documents words, it must assign many topics to it.

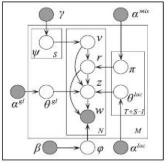
- Trading off these goals finds groups of co-occurring words.

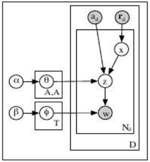
Part II

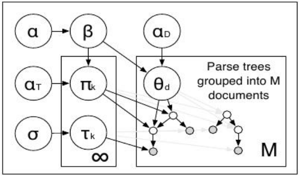
Developing Latent Dirichlet Allocation

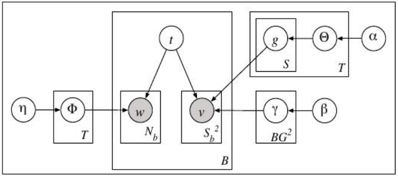


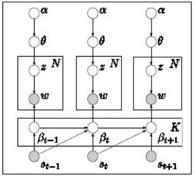




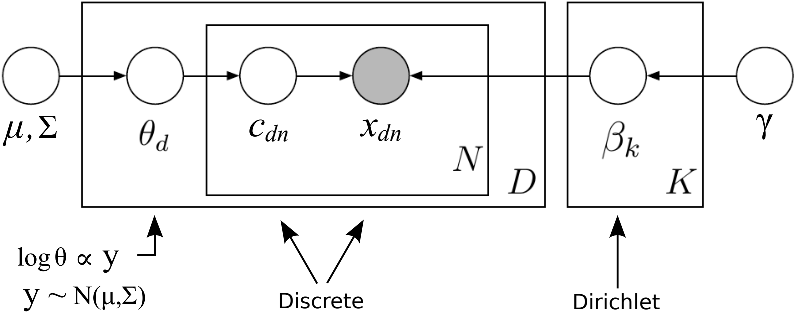








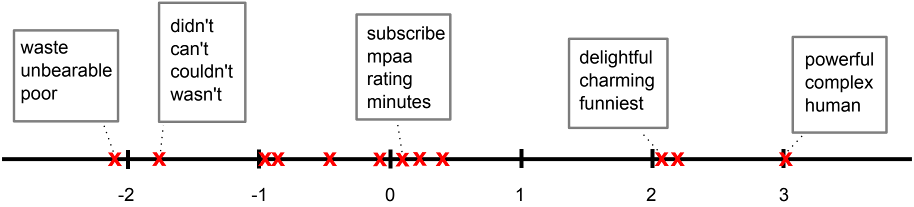
- LDA is a simple building block that enables many applications.
- Can capture assumptions with new **distributions**.
- Can be **embedded** into more complex model structures.



Example 1

- LDA is an "uncorrelated" model.
  - One topic doesn't directly influence the probability of another.
  - Again, all combinations of topics are equally probable.

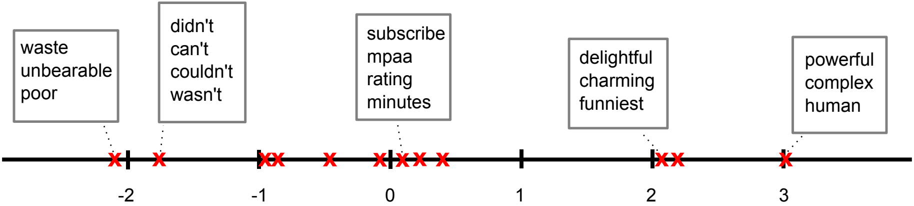
- **Correlated topic models** learn topic covariance structure.



Example 2: Supervised LDA

1. Draw topic proportions $\theta \sim \text{Dirichlet}(\alpha)$
2. For each word
   - Draw topic assignment $c_n \sim \text{Discrete}(\theta)$
   - Draw word observation $x_n \sim \text{Discrete}(\beta_{c_n})$
3. Draw a response variable $y \sim p(w^\top \theta)$

Some versions use the histogram of $(c_1, \ldots, c_N)$ instead of θ.

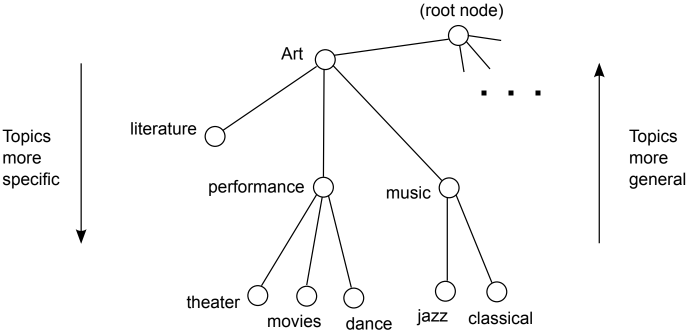


Example 2: Supervised LDA

- A closer look at $y \sim p(w^\top \theta)$ shows how topics are predictive.
- Imagine a set of movie reviews with star ratings and prediction

$$\text{Rating:} \quad y \;\approx\; w^\top \theta \;=\; \sum_{k=1}^{K} w_k \theta_k$$

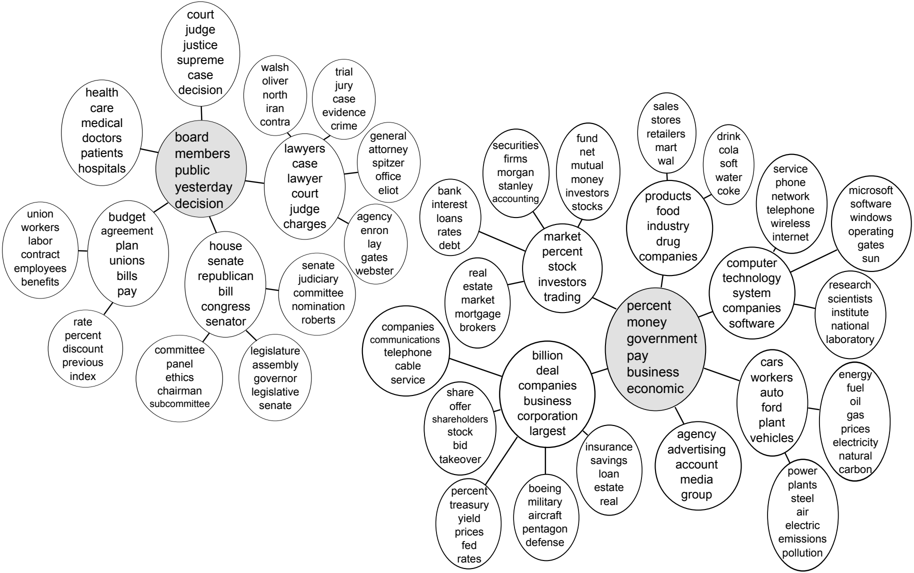
- If $w_k \gg 0$, then $\theta_k > 0$ increases rating. Therefore, words with high probability in topic β_k should be positive. We see two examples above.



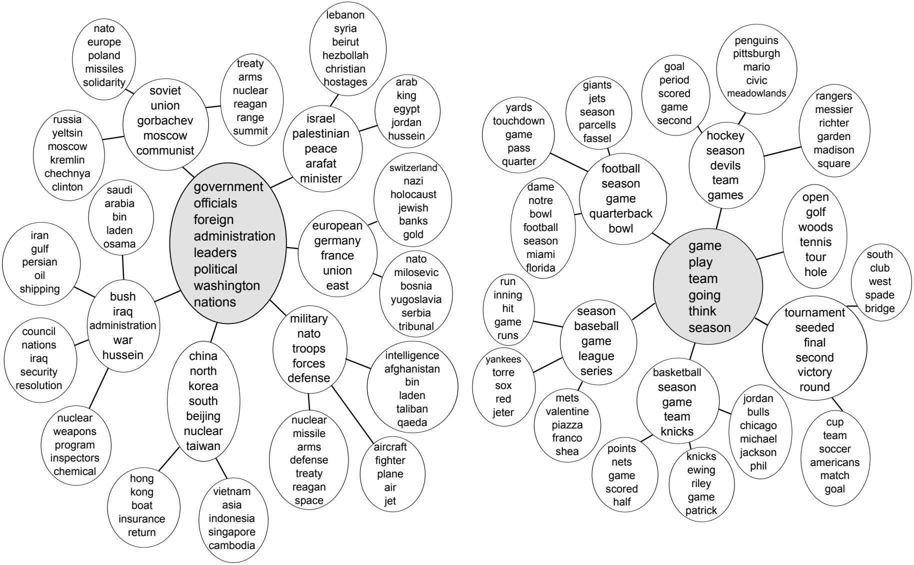
Example 3

- LDA is a "flat" model.
    - There is no structural dependency among the topics.
    - All combinations of topics are *a priori* equally probable.

- **Hierarchical topic models** capture detailed relationships.



- **Good news:** By increasing the number of parameters, we increase capacity of the model to learn interesting things.



- **Challenges:** However, we also increase the amount of data necessary to learn, and so efficient algorithms are necessary.