Modelling Text

Using their Topics

What is Topic Modelling?

- The driving force behind it is the intuition of how a person writes a text

- Possibility:

A text is just a sequence of words!

➔ Language Models

What is Topic Modelling?

- Assumption behind Latent Dirichlet Allocation (LDA)

- Whenever we write a text, we:
  1. *Decide which content we want to write about, and*
  2. *Express this content in words*

What is Topic Modelling? -Example

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

What is Topic Modelling?

- Assumption behind Latent Dirichlet Allocation (LDA)

- Whenever we write a text, we:
  1. *Decide which content we want to write about, and*
  2. *Express this content in words*

➔The content is modelled as a mixture over topics

➔Words that don’t represent any strong semantic meaning are ignored (stopwords, e.g. „a“, „and“)

What is Topic Modelling?

- LDA – Generative Process (How a document is created)

1. Determine document length N_d
2. Determine a distribution over topics θ_d for document d
3. For each „slot“ i in d $(0 < i \leq N_d)$:
   1. Determine a topic z_i for slot i
   2. Select a word of topic z_i using the word-topic distribution ϕ_{z_i}

LDA – Understanding the Parameters

- There are 2 kinds of parameters:
    1. Distribution over topics for each document θ_d

$$\text{E.g. } \theta_d = \begin{bmatrix} p(topic = Arts|d) \\ p(topic = Budget|d) \\ p(topic = Children|d) \\ p(topic = Education|d) \end{bmatrix} = \begin{bmatrix} 0.31 \\ 0.36 \\ 0.24 \\ 0.09 \end{bmatrix}$$

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

LDA – Understanding the Parameters

- There are 2 kinds of parameters:
  2. Distribution over words for each topic z, ϕ_z

$$\text{E.g. } \phi_{z=\color{red}{Arts}} = \begin{bmatrix} p(w = Opera \mid z = \color{red}{Arts}) \\ p(w = York \mid z = \color{red}{Arts}) \\ p(w = Lincoln \mid z = \color{red}{Arts}) \\ \dots \end{bmatrix}$$

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

LDA – Understanding the Parameters

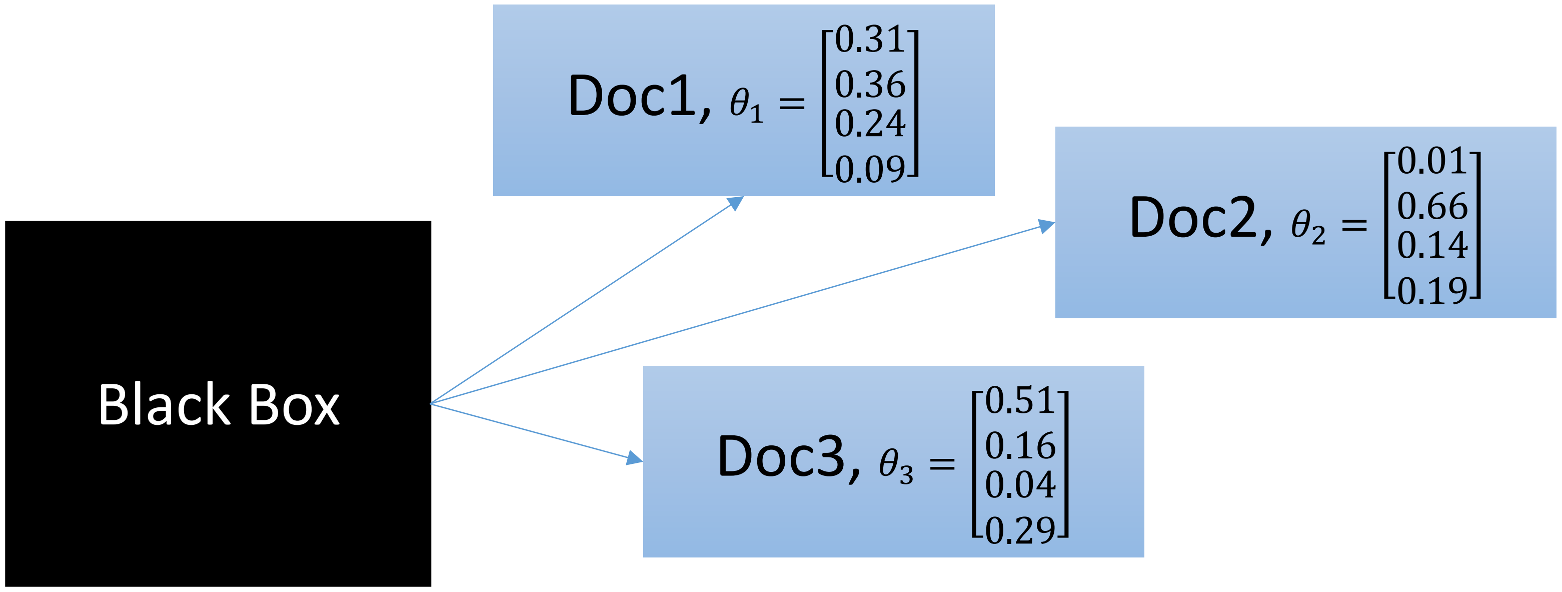
- Assumptions for the parameters:

    1. The distribution of words, given their topic is static for every document (but every topic can have a different distribution over words)

    2. The distribution of topics for a given document changes with every document

LDA – Understanding the Parameters

- When generating a document:
  - We have to sample the distribution θ_d from a distribution that is able to generate different distributions !!



LDA – Understanding the Parameters

- When generating a document:
  - We have to sample the distribution θ_d from a distribution that is able to generate different distributions !!

➔We could just create random vectors for each document, but:
  - We would lose „control“ (e.g. if I know that my documents doesn’t contain any baby toys, I don’t want to generate a distribution with a high baby toys proportion)
  - We would loose mathematical properties (so called „Conjugated Prior“)

LDA – The Dirichlet Distribution

- Rather complex mathematical representation:

$$f(x_1, \dots x_k, \alpha_1, \dots \alpha_k) = \frac{1}{Beta(\vec{\alpha})} \cdot \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

with $Beta(\vec{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)}$

where $\Gamma(\alpha)$ corresponds to the Gamma function and can be seen as the factorial function for continuous inputs

The Dirichlet Distribution: Example

- Given a problem with 3 different outcomes x_1, x_2, x_3 (just think of it as a dice with 3 possible outcomes 1, 2 or 3)

- Let us assume we love throwing the 3-sides dice and have done many throws in our life and found:
  - 1 appeared 7 times α_1
  - 2 appeared 2 times α_2
  - 3 appeared 5 times α_3

➔The vector of our alphas $\vec{\alpha} = (7,2,5)$

➔Our belief is that the most probable distribution would be:

$$x_1 = 0.5, x_2 = 0.14 \text{ and } x_3 = 0.36$$

The Dirichlet Distribution: Example

- $f(x_1, \dots x_k, \alpha_1, \dots \alpha_k) = \frac{1}{Beta(\vec{\alpha})} \cdot \prod_{i=1}^{K} x_i^{\alpha_i - 1}$

- Let us calculate the probability for:
  - $\vec{\alpha} = (7,2,5)$
  - $\vec{x} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$

➔We ignore $\frac{1}{Beta(\vec{\alpha})}$ since it is the same for all values

➔$P\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 7,2,5\right) \sim \left(\frac{1}{3}\right)^6 \cdot \left(\frac{1}{3}\right)^1 \cdot \left(\frac{1}{3}\right)^4 \sim 5.6 \times 10^{-6}$

The Dirichlet Distribution: Example

- $f(x_1, \dots x_k, \alpha_1, \dots \alpha_k) = \frac{1}{Beta(\vec{\alpha})} \cdot \prod_{i=1}^{K} x_i^{\alpha_i - 1}$

- Let us calculate the probability for:
  - $\vec{\alpha} = (7,2,5)$
  - $\vec{x} = (\frac{7}{14}, \frac{2}{14}, \frac{5}{14})$

➔ $P\left(\frac{7}{14}, \frac{2}{14}, \frac{5}{14}, 7,2,5\right) \sim \left(\frac{7}{14}\right)^6 \cdot \left(\frac{2}{14}\right)^1 \cdot \left(\frac{5}{14}\right)^4 \sim 3.6 \times 10^{-5}$

The Dirichlet Distribution: Example

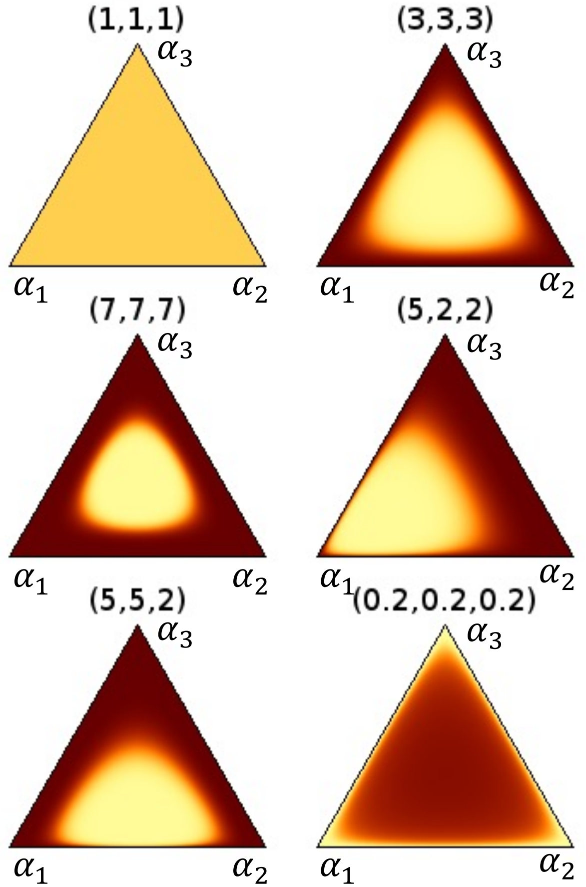
➔ $$\frac{f\left(\frac{7}{14},\frac{2}{14},\frac{5}{14},7,2,5\right)}{f\left(\frac{1}{3},\frac{1}{3},\frac{1}{3},7,2,5\right)} \sim \frac{3.6\times 10^{-5}}{5.6\times 10^{-6}} \sim 6.4$$

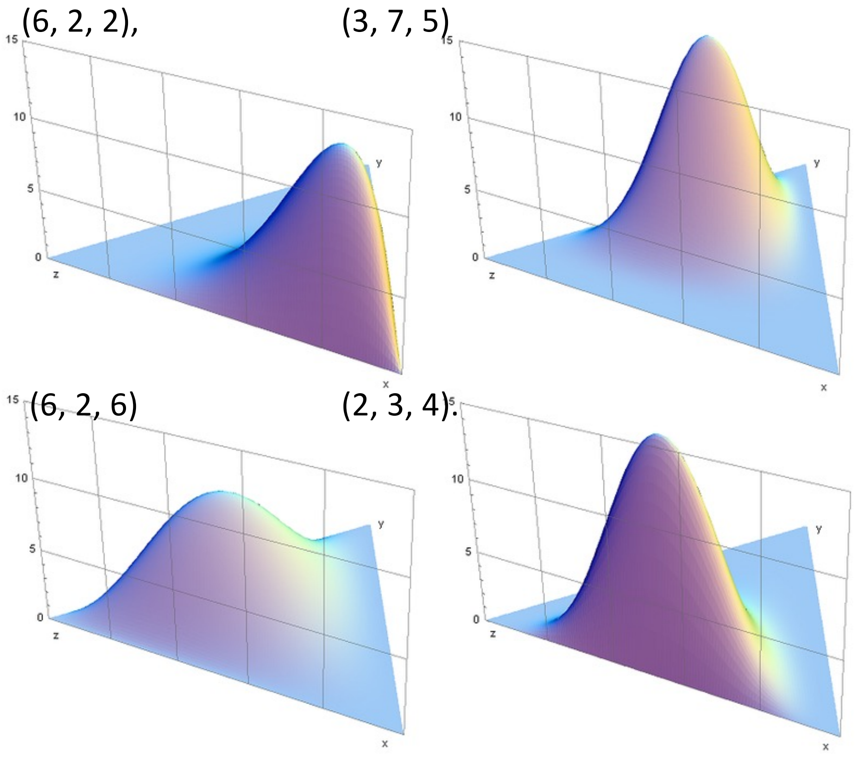
- The chance that our Dirichlet Distribution produces a uniformly distributed distribution is about 6.4 times lower than to produce a biased distribution

The Dirichlet Distribution: Example

- Figures: Dirichlet distributions with three parameters $(\alpha_1, \alpha_2, \alpha_3)$

$$f(x_1, \dots x_k, \alpha_1, \dots \alpha_k) = \frac{1}{Beta(\vec{\alpha})} \cdot \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$





Taken from http://ml4dummies.blogspot.com/2018/01/the-dirichlet-distribution-is-conjugate.html
And https://de.wikipedia.org/wiki/Dirichlet-Verteilung

LDA – Generative Process

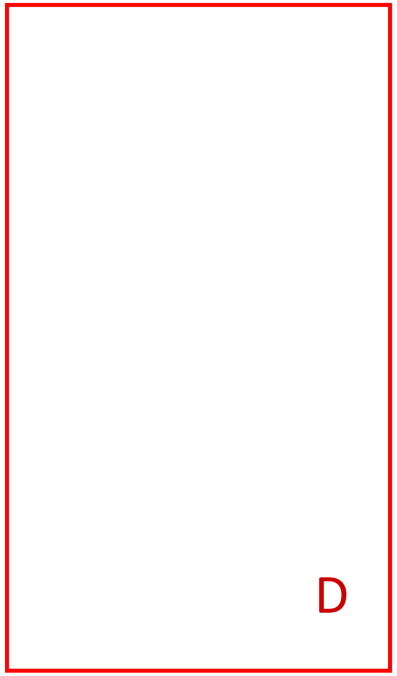
- For each Topic z
  - Sample a distribution ϕ_z from a Dirichlet α

- For each document
  1. Select a document length N
  2. Sample a distribution of topics from θ_d from a Dirichlet β
  3. For each word in $w_i \in d$:
     i. Topic z_i sampled from θ_d
     ii. Actual word w_i sampled from ϕ_{z_i}

LDA – Graphical Representation

- For each Topic z
  - Sample a distribution ϕ_z from a Dirichlet α

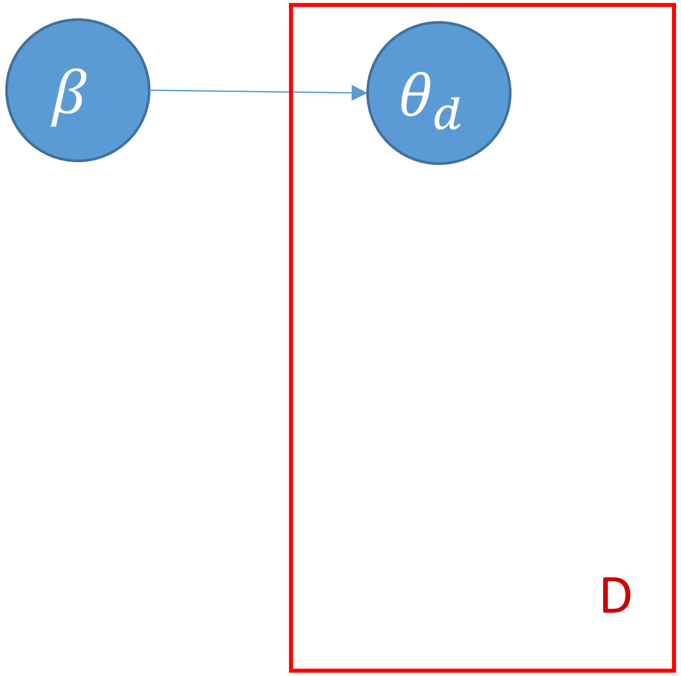
- For each document
  1. Select a document length N
  2. Sample a distribution of topics from θ_d from a Dirichlet β
  3. For each word in $w_i \in d$:
     i. Topic z_i sampled from θ_d
     ii. Actual word w_i sampled from ϕ_{z_i}



LDA – Graphical Representation

- For each Topic z
  - Sample a distribution ϕ_z from a Dirichlet α

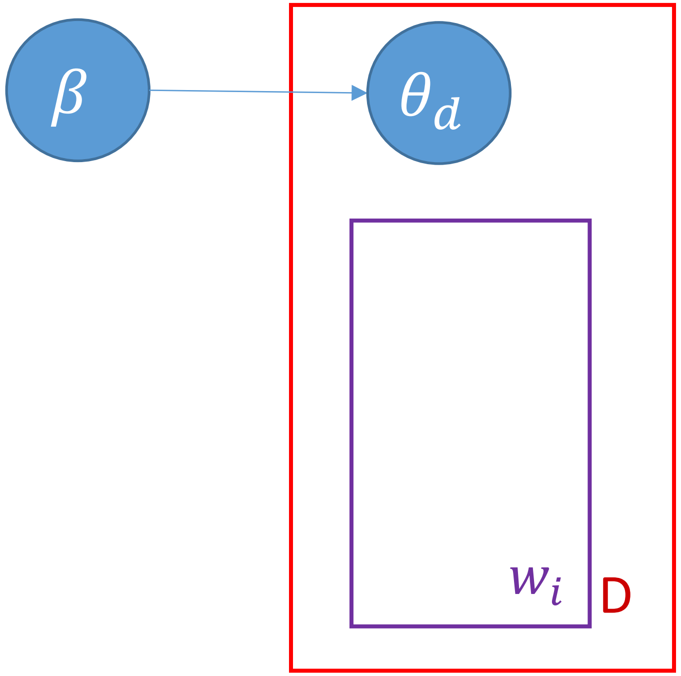
- For each document
  1. Select a document length N
  2. Sample a distribution of topics from θ_d from a Dirichlet β
  3. For each word in $w_i \in d$:
     1. Topic z_i sampled from θ_d
     2. Actual word w_i sampled from ϕ_{z_i}



LDA – Graphical Representation

- For each Topic z
  - Sample a distribution ϕ_z from a Dirichlet α

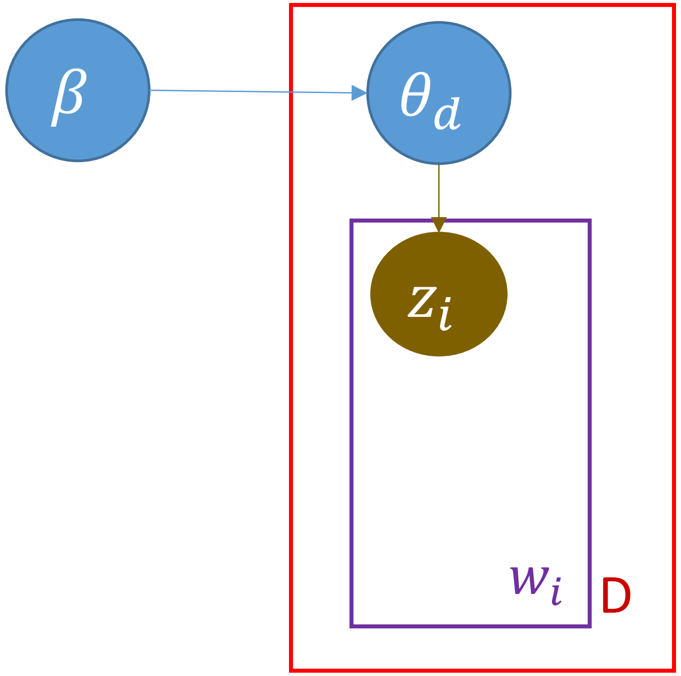
- For each document
  1. Select a document length N
  2. Sample a distribution of topics from θ_d from a Dirichlet β
  3. For each word in $w_i \in d$:
     i. Topic z_i sampled from θ_d
     ii. Actual word w_i sampled from ϕ_{z_i}



LDA – Graphical Representation

- For each Topic z
  - Sample a distribution ϕ_z from a Dirichlet α

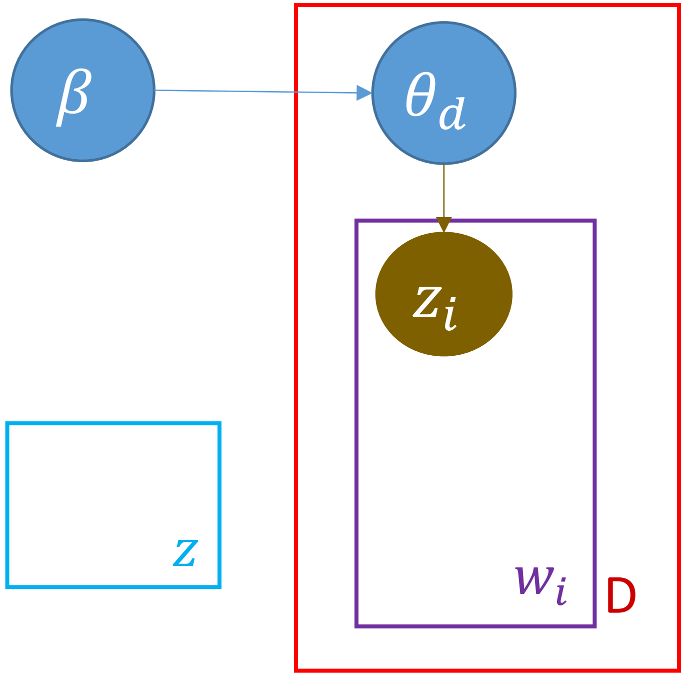
- For each document
  1. Select a document length N
  2. Sample a distribution of topics from θ_d from a Dirichlet β
  3. For each word in $w_i \in d$:
     i. Topic z_i sampled from θ_d
     ii. Actual word w_i sampled from ϕ_{z_i}



LDA – Graphical Representation

- For each Topic z
  - Sample a distribution ϕ_z from a Dirichlet α

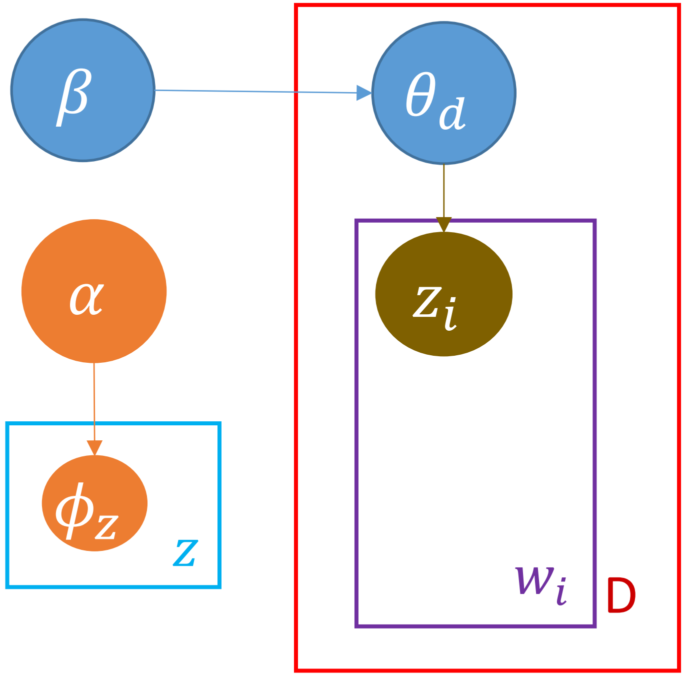
- For each document
  1. Select a document length N
  2. Sample a distribution of topics from θ_d from a Dirichlet β
  3. For each word in $w_i \in d$:
     i. Topic z_i sampled from θ_d
     ii. Actual word w_i sampled from ϕ_{z_i}



LDA – Graphical Representation

- For each Topic z
  - Sample a distribution ϕ_z from a Dirichlet α

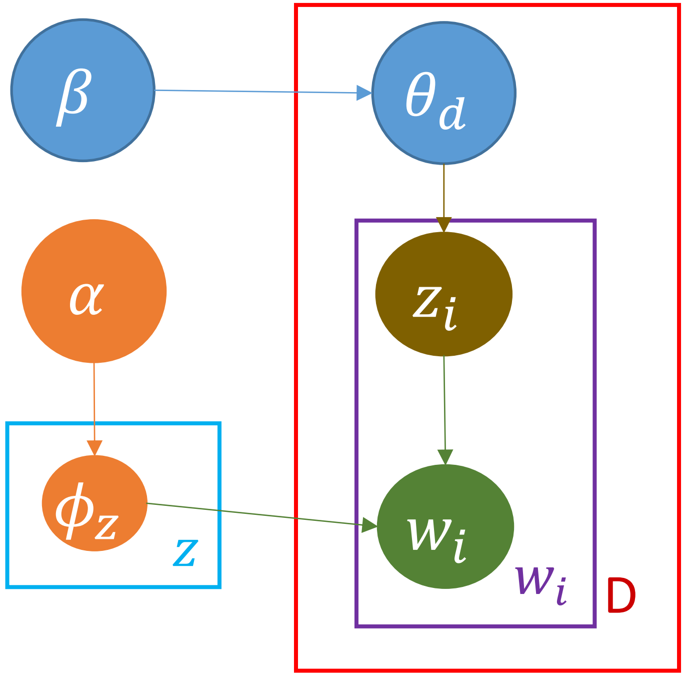
- For each document
  1. Select a document length N
  2. Sample a distribution of topics from θ_d from a Dirichlet β
  3. For each word in $w_i \in d$:
     i. Topic z_i sampled from θ_d
     ii. Actual word w_i sampled from ϕ_{z_i}



LDA – Graphical Representation

- For each Topic z
  - Sample a distribution ϕ_z from a Dirichlet α

- For each document
  1. Select a document length N
  2. Sample a distribution of topics from θ_d from a Dirichlet β
  3. For each word in $w_i \in d$:
     i. Topic z_i sampled from θ_d
     ii. Actual word w_i sampled from ϕ_{z_i}



LDA- Inference – Gibbs Sampling

- Gibbs Sampling is a form of „Markov-Chain Monte Carlo“ short MCMC

- Idea:
  1. Initially assign random values to the variables (our topics for each word)
  2. Define an ordering of those variables z_i
  3. Repeat k times:
  4. Assign a new value for z_i given all other assignments $z_{\neg i}$

$$p(z_i|z_{\neg i}) = \frac{p(z_i, z_{\neg i})}{p(z_{\neg i})} = \frac{p(z)}{p(z_{\neg i})}$$

Gibbs Sampling for LDA

- Initialisation

1. For every word w_i assign a random topic z_i

➔ We can now count:

- how often a topic k appeared in a document d: n_d^k
- and how often a topic k was assign to a term n_k^t

2 (rather huge) matrices

Gibbs Sampling for LDA

- Gibbs Sampling „Burn-in“

1. For a fix amount of epochs, do:

   For each word w_i:

   - remove the current topic-assignment from the counts n_d^k and n_k^t

This leaves us with $z_{\neg i}$

Gibbs Sampling for LDA

- Gibbs Sampling „Burn-in“

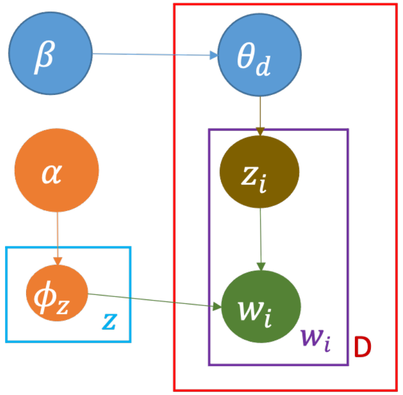
1. For a fix amount of epochs, do:

   For each word w_i:

   - remove the current topic-assignment from the counts n_d^k and n_k^t
   - Sample a new topic z_i from $p(z_i|z_{\neg i})$
   - Update counts n_d^k and n_k^t according to z_i

- Convergence:
  - Calculate parameters θ_d and ϕ_z using our counts n_d^k and n_k^t

Gibbs Sampling for LDA



- Gibbs Sampling „Burn-in“

1. For a fix amount of epochs, do:

   For each word w_i:

   - remove the current topic-assignment from the counts n_d^k and n_k^t
   - Sample a new topic z_i from $p(z_i|z_{\neg i})$
   - Update counts n_d^k and n_k^t according to z_i

But how?

- Convergence:
  - Calculate parameters θ_d and ϕ_z using our counts n_d^k and n_k^t

Gibbs Sampling for LDA

- Derivation of $p(z_i = k | z_{\neg i})$:

! Goal: make use of Dirichlet properties to sample z_i directly using counts n_d^k and n_k^t

Gibbs Sampling for LDA

- Derivation of $p(z_i = k | z_{\neg i})$:

$$p(z_i = k | z_{\neg i}) = p(z_i = k | z_{\neg i}, w)$$

$$= \frac{p(w, z_i, z_{\neg i})}{p(z_{\neg i}, w)} = \frac{p(w, z)}{p(z_{\neg i}, w)}$$

Conditional Probability

Gibbs Sampling for LDA

- Derivation of $p(z_i = k | z_{\neg i})$:

$$= \frac{p(w, z_i, z_{\neg i})}{p(z_{\neg i}, w)} = \frac{p(w, z)}{p(z_{\neg i}, w)}$$

$$= \frac{p(w|z)}{p(w|z_{\neg i})} \cdot \frac{p(z)}{p(z_{\neg i})}$$

Conditional Probability

Gibbs Sampling for LDA

- Derivation of $p(z_i = k|z_{\neg i})$:

$$\frac{p(w|z)}{p(w\,|z_{\neg i})} \cdot \frac{p(z)}{p(z_{\neg i})}$$

$$= \frac{p(w|z)}{p(w_i\,,w_{\neg i}|z_{\neg i})} \cdot \frac{p(z)}{p(z_{\neg i})}$$

$$= \frac{p(w|z)}{p\left(w_i\;\;|z_{\neg i}\right) \cdot p(w_{\neg i}\;|z_{\neg i})} \cdot \frac{p(z)}{p(z_{\neg i})}$$

$$\frac{p(w|z)}{p(w_i) \cdot p(w_{\neg i}\;|z_{\neg i})} \cdot \frac{p(z)}{p(z_{\neg i})}$$

Gibbs Sampling for LDA

- Derivation of $p(z_i = k|z_{\neg i})$:

$$p(z_i = k|z_{\neg i}) = \frac{p(w|z)}{p(w_i) \cdot p(w_{\neg i}\,|z_{\neg i})} \cdot \frac{p(z)}{p(z_{\neg i})}$$

$$\sim \frac{p(w|z)}{p(w_{\neg i}\,|z_{\neg i})} \cdot \frac{p(z)}{p(z_{\neg i})}$$

Same for every k
➔ignore

Gibbs Sampling for LDA

- Derivation of $p(z_i = k | z_{\neg i})$

$$\sim \frac{p(w|z)}{p(w_{\neg i} \, | z_{\neg i})} \cdot \frac{p(z)}{p(z_{\neg i})}$$

- We basically need to describe 2 different distributions:

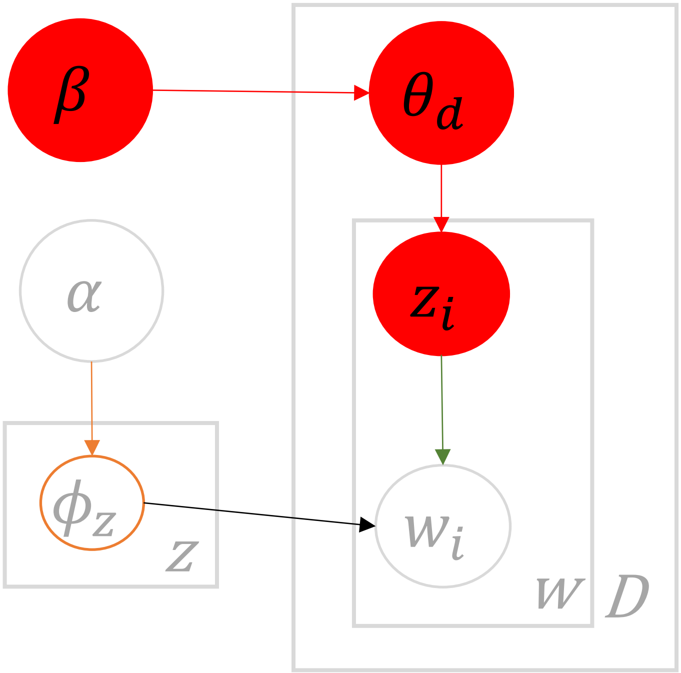
1. $p(w|z)$
2. $p(z)$

Gibbs Sampling for LDA

- Derivation of the term $p(z)$:

$$\Rightarrow p(z) = \int p(z|\theta) \cdot p(\theta|\beta)\, d\theta$$

Marginalising θ



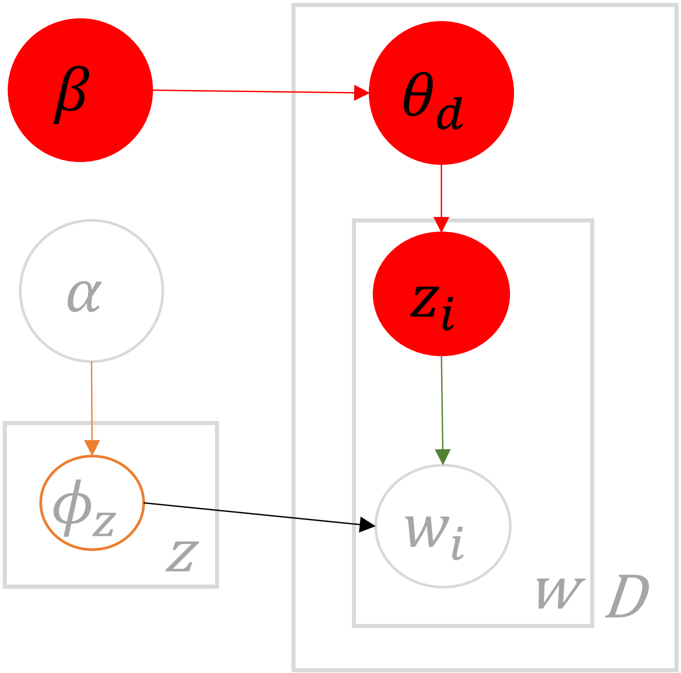
Gibbs Sampling for LDA

- Derivation of the term $p(z)$:

$$\Rightarrow p(z) = \int p(z|\theta) \cdot p(\theta|\beta)\, d\theta$$

Multinomial

Dirichlet



➔ The product results in a Dirichlet with different pseudo counts $\overrightarrow{\beta'}$

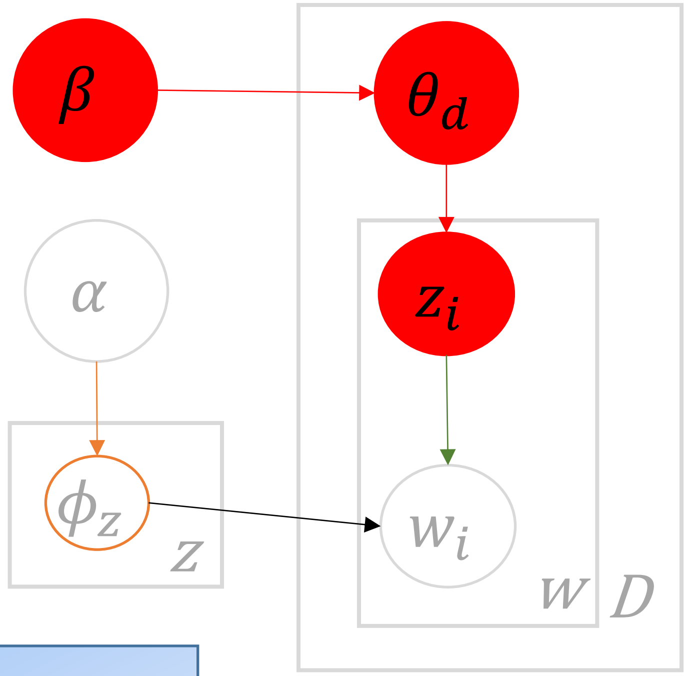
Gibbs Sampling for LDA

- Derivation of the term $p(z)$:

$$\Rightarrow p(z) = \int p(\theta|\beta + n_d - 1)\, d\theta$$

➔The integral results in a Beta-Function

$$\frac{Beta(n_d+\beta)}{Beta(\beta)}$$



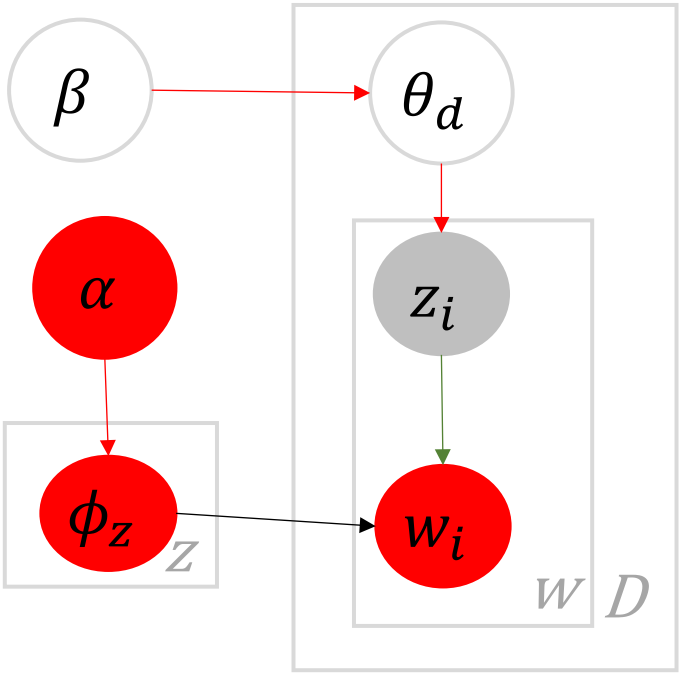
n_d: How often topic k was assigned to a document

Gibbs Sampling for LDA

- Derivation of the term $p(w|z)$:

$$\Rightarrow p(w|z) = \int p(w|z,\phi) \cdot p(\phi|\alpha)\, d\phi$$

Marginalising ϕ



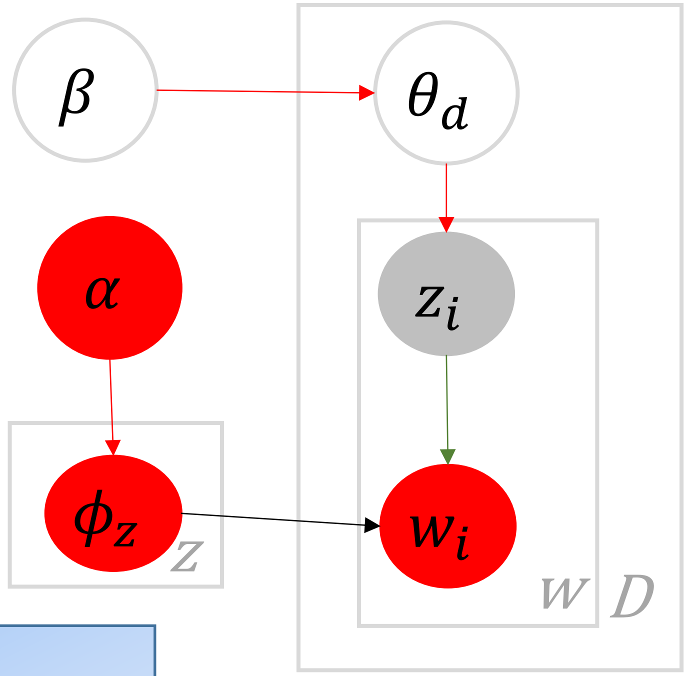
Gibbs Sampling for LDA

- Derivation of the term $p(w|z)$:

$$\Rightarrow p(w|z) = \int p(w|z,\phi) \cdot p(\phi|\alpha)\, d\phi$$

Multinomial

Dirichlet

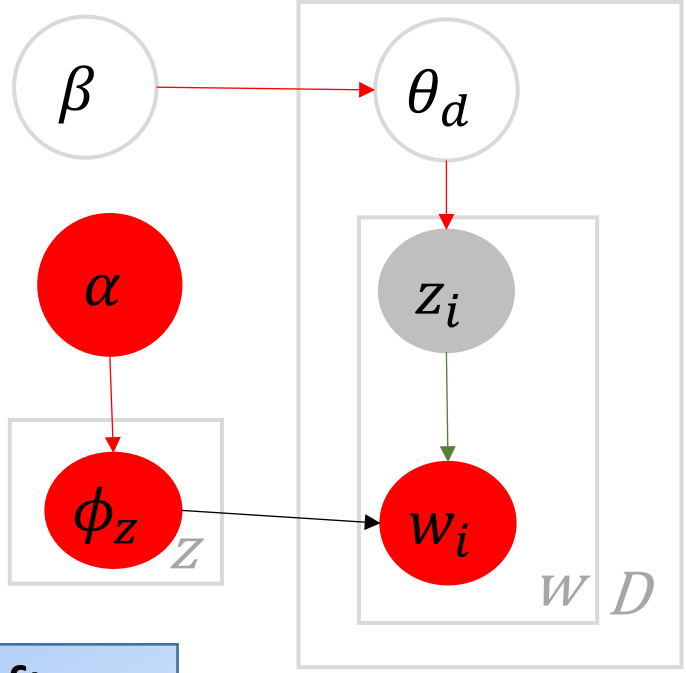


Gibbs Sampling for LDA

- Derivation of the term $p(w|z)$:

$$\Rightarrow p(w|z) = \int p(w|z,\phi) \cdot p(\phi|\alpha)\, d\phi$$

$$= \frac{Beta(n_z + \alpha)}{Beta(\alpha)}$$



n_z: How often topic z was found

Gibbs Sampling for LDA

- Derivation of $p(z_i = k|z_{\neg i})$

$$\sim \frac{p(w|z)}{p(w_{\neg i}\,|z_{\neg i})} \cdot \frac{p(z)}{p(z_{\neg i})}$$

$$p(z_i|z_{\neg i}) \sim \frac{Beta(n_z + \alpha)}{Beta(n_{z_{\neg i}} + \alpha)} \cdot \frac{Beta(n_d + \beta)}{Beta(n_{d_{\neg i}} + \beta)}$$

Gibbs Sampling for LDA

Requires law of the Gamma-Distribution
$\Gamma(x+1) = x\Gamma(x)$

- Derivation of $p(z_i = k|z_{\neg i})$

$$p(z_i|z_{\neg i}) \sim \frac{Beta(n_z + \alpha)}{Beta(n_{z\neg i} + \alpha)} \cdot \frac{Beta(n_d + \beta)}{Beta(n_{d\neg i} + \beta)}$$

$$\boxed{p(z_i = k|z_{\neg i}) = \frac{\mathrm{n}_{\mathrm{k}\neg i}^{t} + \alpha_t}{\sum_{t'=1}^{V}(n_{k\neg i}^{t'} + \alpha_{t'})} \cdot \frac{\mathrm{n}_{\mathrm{d}\neg i}^{k} + \beta_k}{\left[\sum_{k'=1}^{K}(n_d^{k'} + \beta_{k'})\right] - 1}}$$

Understanding the result

$$p(z_i = k|z_{\neg i}) = \frac{\mathrm{n}^{t}_{\mathrm{k}\neg i} + \alpha_t}{\sum_{t'=1}^{V}(n^{t'}_{k\neg i} + \alpha_{t'})} \cdot \frac{\mathrm{n}^{k}_{\mathrm{d}\neg i} + \beta_k}{\left[\sum_{k'=1}^{K}(n^{k'}_{d} + \beta_{k'})\right] - 1}$$

n^{t}_{k}: How often topic *k* was assigned to term *t*

α_t: How often was a term *t* found with topic *k* apriori

$\mathrm{n}^{k}_{\mathrm{d}}$: How often topic *k* was assigned to document *d*

β_k: How often topic *k* was assigned apriori

Example

- Given a document d

A beaver plays soccer. The ball is made of wood and is as round as the head of an owl.

- Two different topics (let us call them „sports“ and „animals“)
- Apriori counts for $\vec{\alpha}$ and $\vec{\beta}$, set to $\vec{1}$

Initialisation

- Remove stopwords

~~A~~ beaver plays soccer~~.~~ ~~The~~ ball ~~is made of~~ wood ~~and is as~~ round ~~as the~~ head ~~of an~~ owl~~.~~

- Assign random topics $\{0 = \text{sports}; 1 = \text{animals}\}$ to each remaining word

beaver	plays	soccer	ball	wood	round	head	owl
0	1	1	1	0	0	1	0

Initialisation

- Count the following statistics:
  1. Topic to term assignments

↓Token | →Topic	sports	animals
beaver	1	22
plays	20	17
soccer	15	1
ball	12	4
wood	1	33
round	4	14
head	6	12
owl	1	40

Initialisation

- Count the following statistics:
  1. Topic to term assignments
  2. Topic-distribution:

$$\vec{z} = (60{,}143)$$

↓Token | →Topic	sports	animals
beaver	1	22
plays	20	17
soccer	15	1
ball	12	4
wood	1	33
round	4	14
head	6	12
owl	1	40

Initialisation

- Count the following statistics:
  1. Topic to term assignments
  2. Topic-distribution:

$$\vec{z} \quad = (60{,}143)$$

  3. Topic-document-distribution:

$$\overrightarrow{z_d} \quad = (4{,}4)$$

↓Token | →Topic	sports	animals
beaver	1	22
plays	20	17
soccer	15	1
ball	12	4
wood	1	33
round	4	14
head	6	12
owl	1	40

Loop

- Determine topic for beaver:
  1. Remove the assignment
  2. Update counts

$$\vec{z}_{\neg i} = (59{,}143)$$

$$\overrightarrow{z_{\neg i,d}} = (3{,}4)$$

↓Token | →Topic	sports	animals
beaver	0	22
plays	20	17

beaver plays soccer ball wood round head owl

~~0~~ 1 1 1 0 0 1 0

Loop

- Calculate topic probabilities for beaver:

↓Token | →Topic	sports	animals
beaver	0	22
plays	20	17

$$p(z = sports) \sim \frac{0 + 1}{59 + 8 \cdot 1} \cdot \frac{3 + 1}{8 + 2 - 1}$$

$$\vec{z}_{\neg i} = (59{,}143)$$

$$p(z = sports) \sim 0.0066$$

$$\overrightarrow{z_{\neg i,d}} = (3{,}4)$$

$$p(z_i = k|z_{\neg i}) = \frac{\mathrm{n}^{t}_{\mathrm{k}\neg i} + \alpha_t}{\sum_{t'=1}^{V}(n^{t'}_{k\neg i} + \alpha_{t'})} \cdot \frac{\mathrm{n}^{k}_{\mathrm{d}\neg i} + \beta_k}{\left[\sum_{k'=1}^{K}(n^{k'}_{d} + \beta_{k'})\right] - 1}$$

Loop

- Calculate topic probabilities for beaver:

↓Token | →Topic	sports	animals
beaver	0	22
plays	20	17

$$p(z = animals) \sim \frac{22 + 1}{143 + 8 \cdot 1} \cdot \frac{4 + 1}{8 + 2 - 1}$$

$$\vec{z}_{\neg i} = (59, 143)$$

$$p(z = animals) \sim 0.0846$$

$$\overrightarrow{z_{\neg i,d}} = (3, 4)$$

$$p(z_i = k | z_{\neg i}) = \frac{\mathrm{n}_{\mathrm{k}\neg i}^{t} + \alpha_t}{\sum_{t'=1}^{V}(n_{k\neg i}^{t'} + \alpha_{t'})} \cdot \frac{\mathrm{n}_{\mathrm{d}\neg i}^{k} + \beta_k}{\left[\sum_{k'=1}^{K}(n_d^{k'} + \beta_{k'})\right] - 1}$$

Loop

- Renormalize

$p(z = sports) \sim 0.0066$
$p(z = animals) \sim 0.0846$

$\rightarrow p(z = sports) = \frac{0.0066}{0.0066+0.0846} = 7.23\%$
$\rightarrow p(z = animals) = \frac{0.0846}{0.0066+0.0846} = 92.77\%$

$\rightarrow$ We would most likely assign animal to beaver

Loop

- Assign new value

beaver	plays	soccer	ball	wood	round	head	owl
1	1	1	1	0	0	1	0

➔ Increment counts

➔ Continue with the word plays

➔ Loop until convergence or end of iterations

After convergence – Read params

- We still need our parameters

1. Distribution over words for each topic k, ϕ_k^t

$$\phi_k^t = \frac{n_k^t + \alpha_t}{\sum_{t'} n_k^{t'} + \alpha_{t'}}$$

2. Distribution over topics for each document θ_d^k

$$\theta_d^k = \frac{n_d^k + \beta_k}{\sum_{k'} n_d^{k'} + \beta_{k'}}$$

Gibbs Sampling for LDA in full glory

⊡ initialisation
zero all count variables, $n_m^{(k)}, n_m, n_k^{(t)}, n_k$
for all documents $m \in [1, M]$ **do**
    for all words $n \in [1, N_m]$ in document m **do**
        sample topic index $z_{m,n}=k \sim \text{Mult}(1/K)$
        increment document–topic count: $n_m^{(k)} + 1$
        increment document–topic sum: $n_m + 1$
        increment topic–term count: $n_k^{(t)} + 1$
        increment topic–term sum: $n_k + 1$
    end for
end for
⊡ Gibbs sampling over burn-in period and sampling period
while not finished **do**
    for all documents $m \in [1, M]$ **do**
        for all words $n \in [1, N_m]$ in document m **do**
            ⊡ for the current assignment of k to a term t for word $w_{m,n}$:
            decrement counts and sums: $n_m^{(k)} - 1; n_m - 1; n_k^{(t)} - 1; n_k - 1$
            ⊡ multinomial sampling acc. to Eq. 79 (decrements from previous step):
            sample topic index $\tilde{k} \sim p(z_i|\vec{z}_{\neg i}, \vec{w})$
            ⊡ use the new assignment of $z_{m,n}$ to the term t for word $w_{m,n}$ to:
            increment counts and sums: $n_m^{(\tilde{k})} + 1; n_m + 1; n_{\tilde{k}}^{(t)} + 1; n_{\tilde{k}} + 1$
        end for
    end for
    ⊡ check convergence and read out parameters
    if converged and L sampling iterations since last read out **then**
        ⊡ the different parameters read outs are averaged.
        read out parameter set $\underline{\Phi}$ according to Eq. 82
        read out parameter set $\underline{\Theta}$ according to Eq. 83
    end if
end while

Heinrich, Gregor. *Parameter estimation for text analysis*. Technical report, 2005.

Topic Modelling

Applications for LDA

Applications

- Given the algorithm seen before, we can infer the topics of the words in a new document $\hat{d}$ by simply:

$$p(z_i = k | z_{\neg i}) = \frac{\mathrm{n}_{\mathrm{k}\neg i}^{t} + c_{k,\neg i}^{t} + \alpha_t}{\sum_{t=1}^{V} n_{k\neg i}^{t} + c_{k,\neg i}^{t} + \alpha_t} \cdot \frac{\mathrm{n}_{\hat{d}\neg i}^{k} + \beta_k}{\left[\sum_{k=1}^{K} n_{\hat{d}}^{k} + \beta_k\right] - 1}$$

- $c_{k,\neg i}^{t}$ is just the count of the document $\hat{d}$

Applications

- Used for determining a document similarity in Information Retrieval
- Used as features in classification
- Used for clustering
- To analyse the shift of meaning over time
- …