Latent Dirichlet Allocation

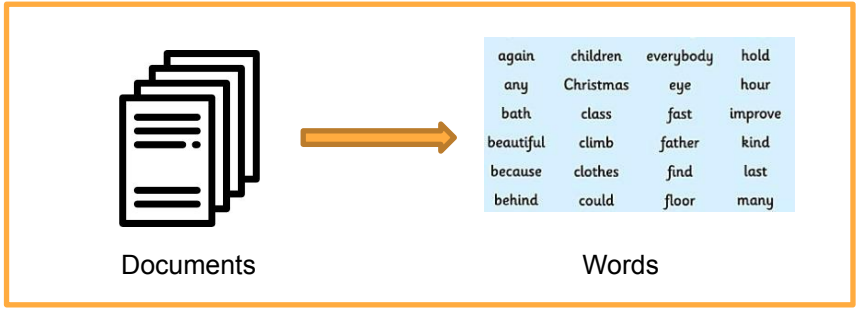
Latent Dirichlet Allocation

- Generative probabilistic model

Latent Dirichlet Allocation

- Generative probabilistic model

  Finds topics from a corpus
  Annotates documents with topics



Analytics Vidhya

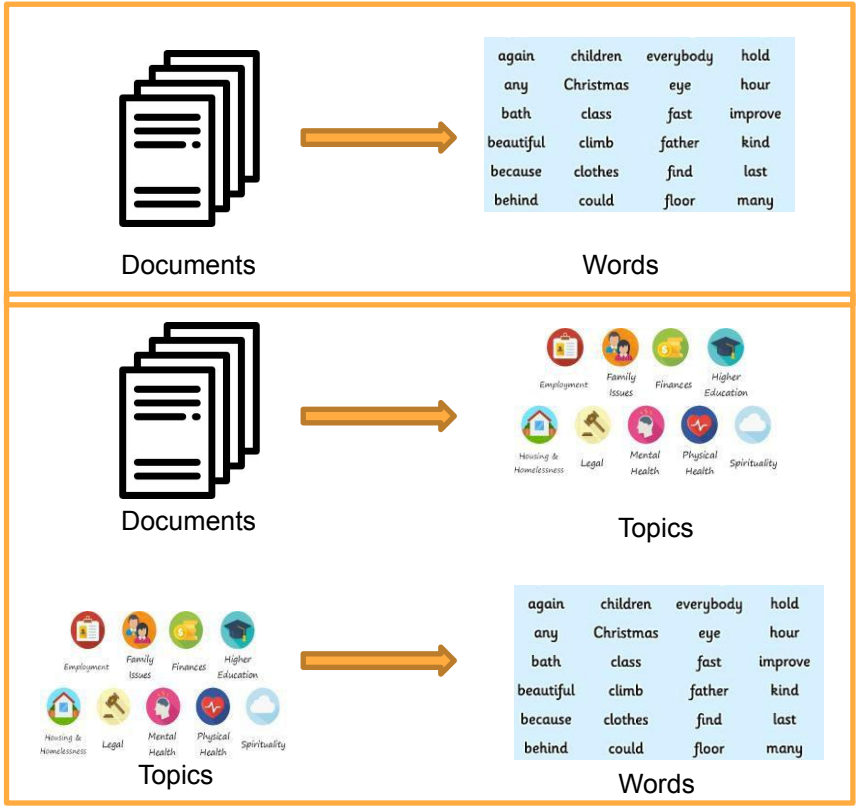
Latent Dirichlet Allocation

- Generative probabilistic model

  Finds topics from a corpus
  Annotates documents with topics

- LDA Assumptions

  Documents = mixture of topics
  Topics = mixture of words



Analytics Vidhya

Latent Dirichlet Allocation

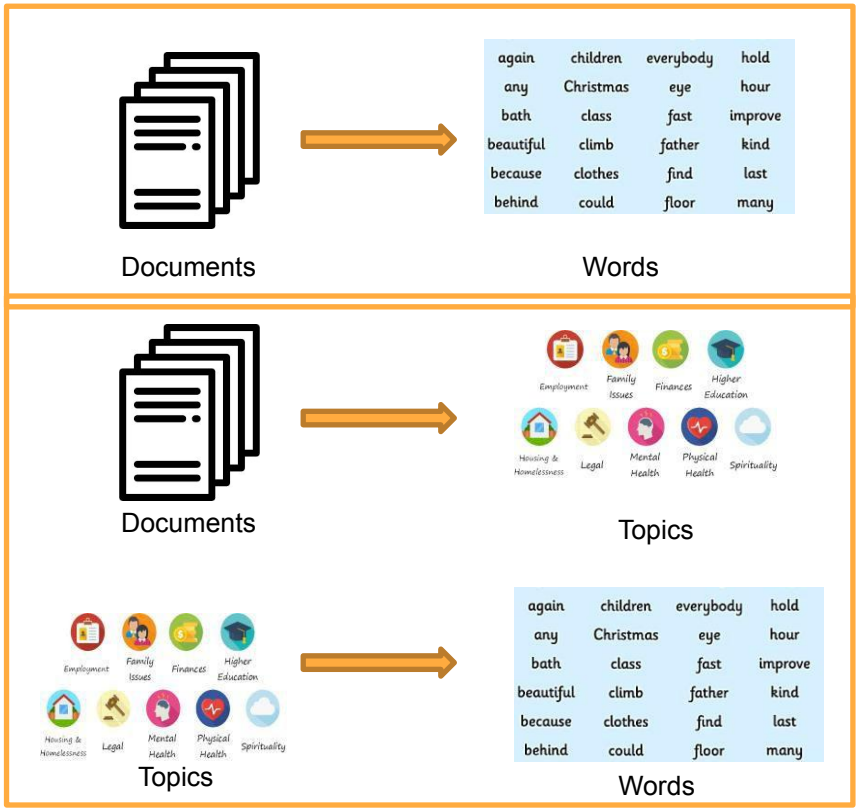
- Generative probabilistic model

  Finds topics from a corpus
  Annotates documents with topics

- LDA Assumptions

  Documents = mixture of topics
  Topics = mixture of words

- Documents : Probability Distributions of Topics
  Topics : Probability Distributions of Words



Analytics Vidhya

Latent Dirichlet Allocation

Latent Dirichlet Allocation

- Corpus : Document Word Matrix
- Document Word Matrix = Document Topic Matrix + Topic Word Matrix

Latent Dirichlet Allocation

- Corpus : Document Word Matrix

Latent Dirichlet Allocation

- Corpus : Document Word Matrix
- Document Word Matrix = Document Topic Matrix + Topic Word Matrix

	W1	W2	W3	Wn
D1	0	2	1	3
D2	1	4	0	0
D3	0	2	3	1
Dn	1	1	3	0

	K1	K2	K3	K
D1	1	0	0	1
D2	1	1	0	0
D3	1	0	0	1
Dn	1	0	1	0

	W1	W2	W3	Wm
K1	0	1	1	1
K2	1	1	1	0
K3	1	0	0	1
K	1	1	0	0

Latent Dirichlet Allocation

- Corpus : Document Word Matrix
- Document Word Matrix = Document Topic Matrix + Topic Word Matrix

	W1	W2	W3	Wn
D1	0	2	1	3
D2	1	4	0	0
D3	0	2	3	1
Dn	1	1	3	0

	K1	K2	K3	K
D1	1	0	0	1
D2	1	1	0	0
D3	1	0	0	1
Dn	1	0	1	0

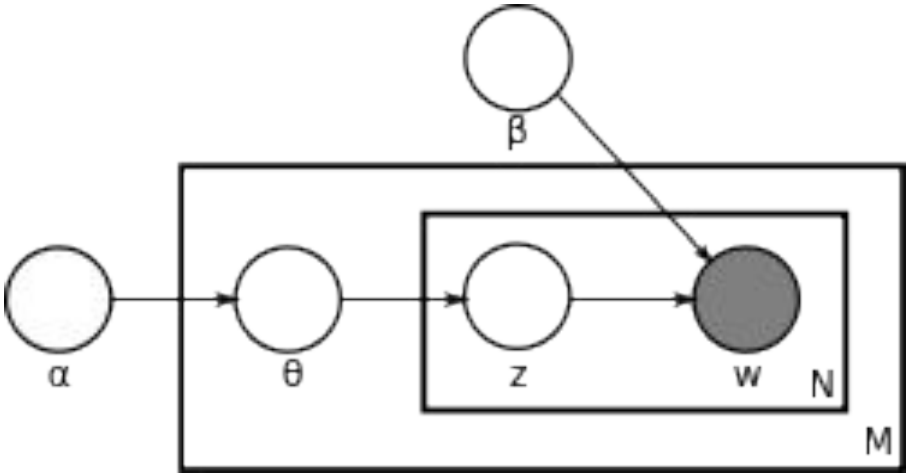
	W1	W2	W3	Wm
K1	0	1	1	1
K2	1	1	1	0
K3	1	0	0	1
K	1	1	0	0

- Goal – Optimize representations

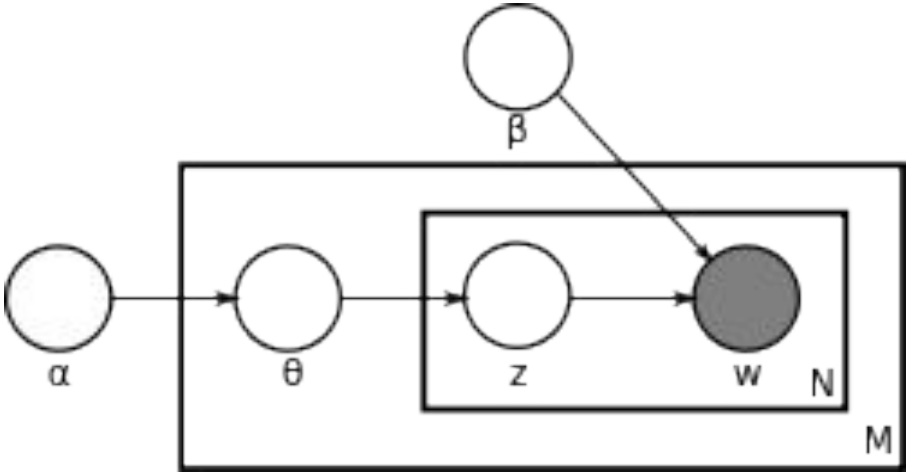
  Document Topic distributions
  Topic Terms distributions

Latent Dirichlet Allocation

Latent Dirichlet Allocation

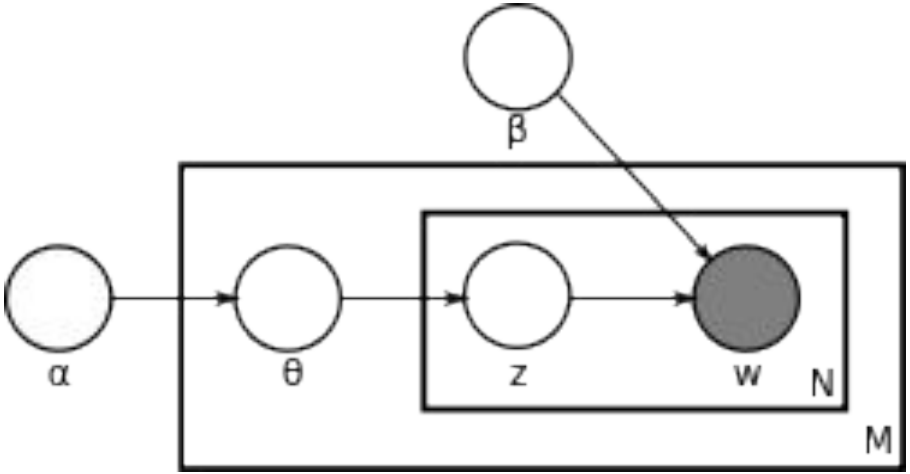


Latent Dirichlet Allocation



- M : Total Documents in Corpus
  N : No of words in a Document
  w : Word in a document
  z : Latent topic assigned to the word
  theta : Topic Distribution

Latent Dirichlet Allocation



- M : Total Documents in Corpus
  N : No of words in a Document
  w : Word in a document
  z : Latent topic assigned to the word
  theta : Topic Distribution

- Alpha, Beta – LDA model parameters

Analytics Vidhya

Latent Dirichlet Allocation

Latent Dirichlet Allocation

- Corpus:

D1 = (w1, w2, w3, w4, …..... wn)
D2 = (w'1, w'2, w'3, w'4, …....
w'n)
D3 = (w''1, w''2, w''3, w''4, …..... w''n)
...
...
Dm = (w1, w2, w3, w4, …..... wn)

Latent Dirichlet Allocation

- Corpus:

D1 = (w1, w2, w3, w4, …..... wn)
D2 = (w'1, w'2, w'3, w'4, …....
w'n)
D3 = (w''1, w''2, w''3, w''4, …..... w''n)
...
...
Dm = (w1, w2, w3, w4, …..... wn)

- First step : Assign random topics to each word

Latent Dirichlet Allocation

- Corpus:

D1 = (w1, w2, w3, w4, ……. wn)
D2 = (w'1, w'2, w'3, w'4, …….
w'n)
D3 = (w''1, w''2, w''3, w''4, ……. w''n)
...
...
Dm = (w1, w2, w3, w4, ……. wn)

- First step : Assign random topics to each word

D1 = (w1 (k4), w2 (k2), w3 (k2), w4 (k2), ……. wn (k3))
D2 = (w'1 (k1), w'2 (k7), w'3 (k3), w'4 (k6), ……. w'n (k2))
D3 = (w''1(k5), w''2 (k4), w''3 (k1), w''4 (k5), ……. w''n
(k1))
...
...
Dm = (w1 (k4), w2 (k2), w3 (k6), w4 (k1), ……. wn (k2))

Analytics Vidhya

Latent Dirichlet Allocation

D1 = (w1 (k4), w2 (k2), w3 (k2), w4 (k2), ……. wn (k3))
D2 = (w'1 (k1), w'2 (k7), w'3 (k3), w'4 (k6), ……. w'n (k2))
D3 = (w''1(k5), w''2 (k4), w''3 (k1), w''4 (k5), ……. w''n (k1))

Documents : Mixture of Topics:

D1 = k4 + k2 + k2 + k2 + … k3 D2 = k1 + k7 + k3 + k6 + … k2 D3 = k5 + k4 + k1 + k5 + … k1 Dn = ...
.

Analytics Vidhya

Latent Dirichlet Allocation

D1 = (w1 (k4), w2 (k2), w3 (k2), w4 (k2), ……. wn (k3))
D2 = (w'1 (k1), w'2 (k7), w'3 (k3), w'4 (k6), ……. w'n (k2))
D3 = (w''1(k5), w''2 (k4), w''3 (k1), w''4 (k5), ……. w''n (k1))

Documents : Mixture of Topics:

D1 = k4 + k2 + k2 + k2 + … k3 D2 = k1 + k7 + k3 + k6 + … k2 D3 = k5 + k4 + k1 + k5 + … k1 Dn = ...

Topics : Mixture of Terms:

k1 = w'1 + w''3

k2 = w2 + w3 + w4 + …
…
kn = wi + ...

Analytics Vidhya

Latent Dirichlet Allocation

Optimization Steps:

Iterate : each document d
Iterate : each word *w*

Latent Dirichlet Allocation

Optimization Steps:

Iterate : each document d
Iterate : each word *w*

-Assume that all topic assignments except the current word are correct

Latent Dirichlet Allocation

Optimization Steps:

Iterate : each document d
Iterate : each word *w*

-Assume that all topic assignments except the current word are correct

-compute p1, p2

Latent Dirichlet Allocation

Optimization Steps:

Iterate : each document d
Iterate : each word *w*

-Assume that all topic assignments except the current word are correct

-compute p1, p2

p1 = proportion (topic t / document d) *p2 = proportion (word w / topic t)*

p1 -> proportion of words in document d that are currently assigned to topic t
p2 -> proportion of assignments to topic t that come from w, over all documents

Analytics Vidhya

Latent Dirichlet Allocation

- Reassign word w of document d a new topic k’

  - Where we choose topic k’ with a new probability = p1 * p2

Latent Dirichlet Allocation

- Reassign word w of document d a new topic k’

  - Where we choose topic k’ with a new probability = p1 * p2

- Repeated large number of times until steady state

Thank You

Analytics Vidhya