Introduction to Topic Models

Dr Pierre Le Bras

February 2020

Topic Modelling

Topic Modelling

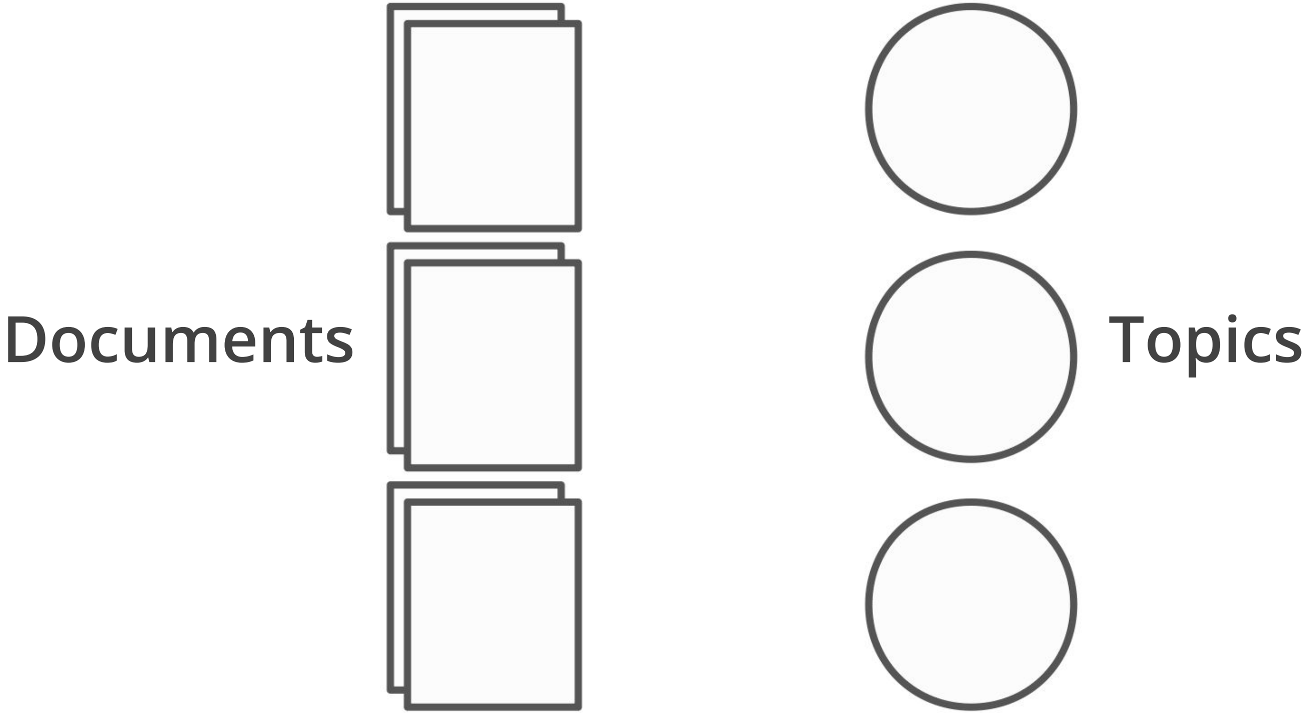
- Any text document
  - Course materials
  - Academic papers
  - Company reports
  - Emails
  - Wikipedia

⇒ Collections of Words or Labels

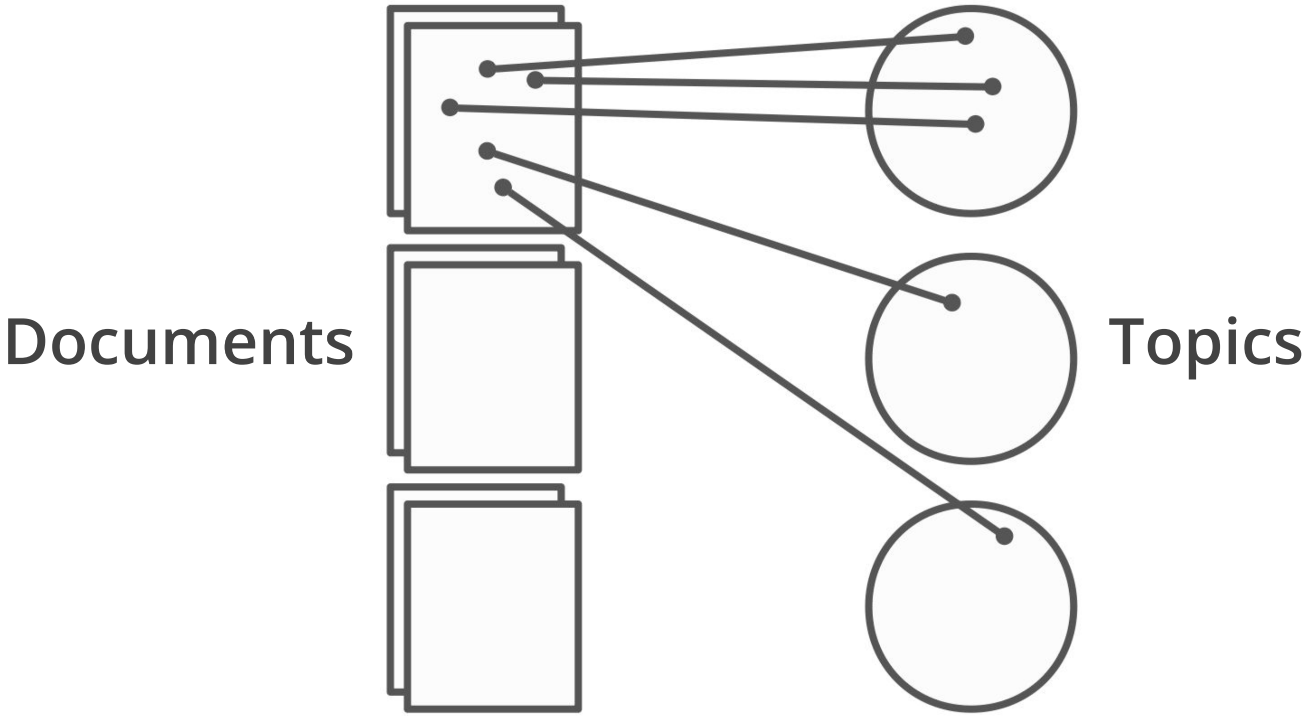
Topic Modelling

?

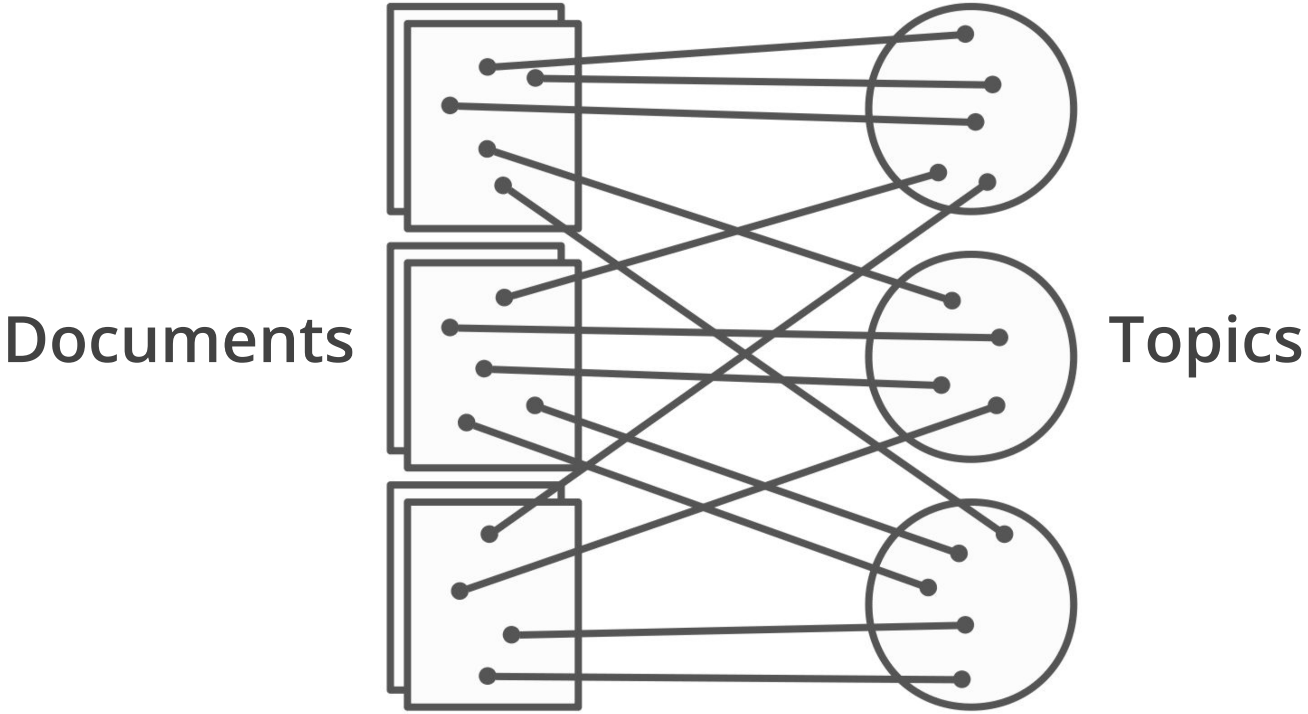
Topic Modelling



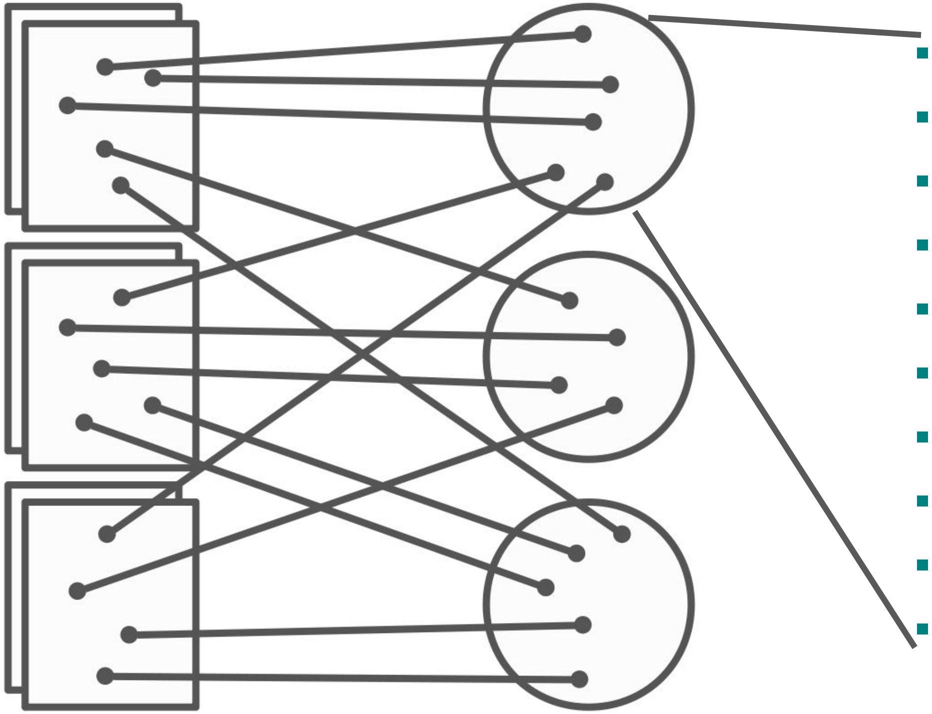
Topic Modelling



Topic Modelling

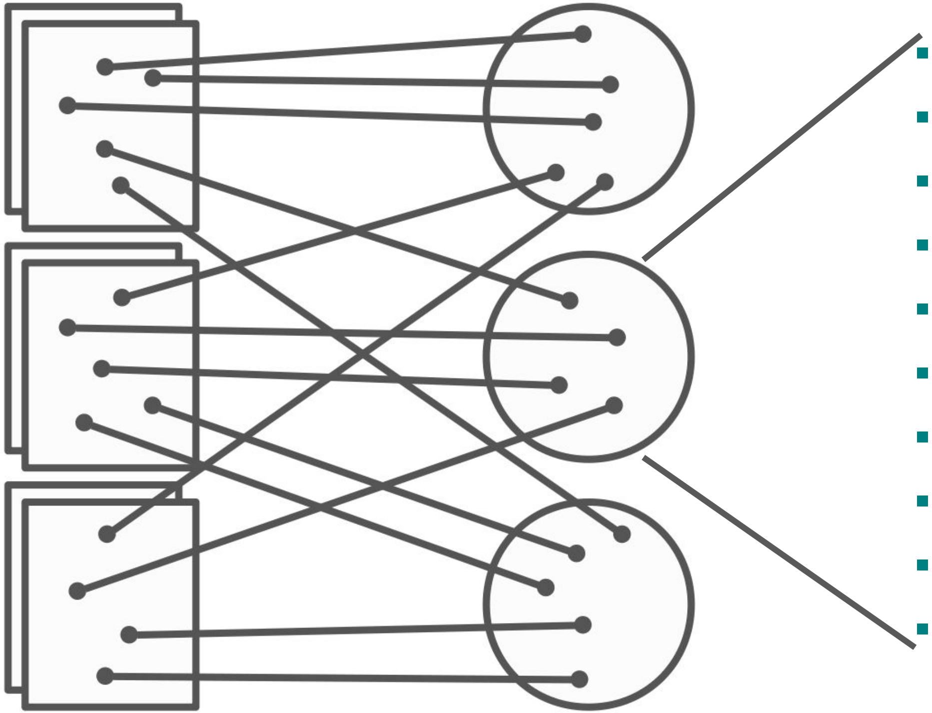


Topic Modelling



- User
- Information
- Technology
- Digital
- Data
- Media
- Design
- Online
- Social
- Interaction

Topic Modelling



- Care
- Health
- Patient
- Service
- Practice
- People
- Study
- Support
- Interview
- Experience

LDA

Latent Dirichlet Allocation

How are documents created ?

How are documents created ?

$$\theta(t|d) \sim Dir(\alpha)$$

$$\varphi(w|t) \sim Dir(\beta)$$

$$\nu(i,d) \sim \theta(t|d)$$

$$\omega(i,d) \sim \varphi(w|\nu(i,d))$$

LDA

Documents
Document 1
Document 2
Document 3
Document 4
…
Document M

Topics	
	Topic 1
	Topic 2
	Topic 3
	Topic 4
	…
	Topic N

Words	
	Word 1
	Word 2
	Word 3
	Word 4
	…
	Word O

LDA

$$\theta(t \mid d) \sim Dir(\alpha)$$

$$\varphi(w \mid t) \sim Dir(\beta)$$

$$\nu(i,d) \sim \theta(t \mid d)$$

$$\omega(i,d) \sim \varphi(w \mid \nu(i,d))$$

LDA

Documents
Document 1
Document 2
Document 3
Document 4
…
Document M

$\theta(t \mid D3)$

Topics	
0.06	Topic 1
0.15	Topic 2
0.1	Topic 3
0.08	Topic 4
	…
0.03	Topic N

Words	
	Word 1
	Word 2
	Word 3
	Word 4
	…
	Word O

LDA

Documents
Document 1
Document 2
Document 3
Document 4
…
Document M

Topics	
0.07	Topic 1
0.03	Topic 2
0.08	Topic 3
0.2	Topic 4
	…
0.1	Topic N

$\theta(t|D2)$

Words	
	Word 1
	Word 2
	Word 3
	Word 4
	…
	Word O

LDA

- Documents are distributions of Topics
    - *A document is made of multiple topics*

LDA

$$\theta(t|d) \sim Dir(\alpha)$$

$$\varphi(w|t) \sim Dir(\beta)$$

$$\nu(i,d) \sim \theta(t|d)$$

$$\omega(i,d) \sim \varphi(w|\nu(i,d))$$

LDA

Documents
Document 1
Document 2
Document 3
Document 4
…
Document M

Topics	
	Topic 1
	Topic 2
	Topic 3
	Topic 4
	…
	Topic N

Words	
0.08	Word 1
0.1	Word 2
0.12	Word 3
0.04	Word 4
	…
0.15	Word O

$\varphi(w|T2)$

LDA

Documents
Document 1
Document 2
Document 3
Document 4
…
Document M

Topics	
	Topic 1
	Topic 2
	Topic 3
	Topic 4
	…
	Topic N

Words	
0.14	Word 1
0.01	Word 2
0.13	Word 3
0.08	Word 4
	…
0.05	Word O

$\varphi(w|T4)$

LDA

- Documents are distributions of Topics
- Topics are distributions of Words
    - *A topic is a collection of words*

Token 36 in Document 2 = ?

LDA

$$\theta(t|d) \sim Dir(\alpha)$$

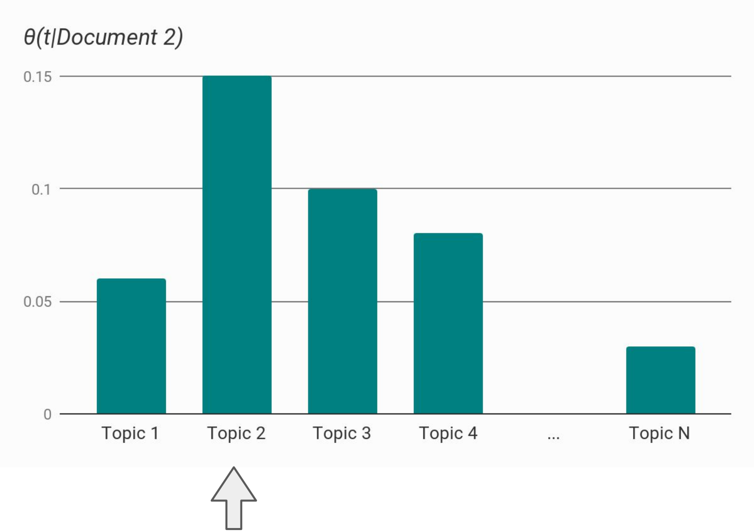
$$\varphi(w|t) \sim Dir(\beta)$$

$$\nu(i,d) \sim \theta(t|d)$$

$$\omega(i,d) \sim \varphi(w|\nu(i,d))$$

LDA

Token 36 in Document 2 = ?



$$v(36, D2) \sim \theta(t \mid D2)$$

LDA

- Documents are distributions of Topics
- Topics are distributions of Words
- To *write* a word in a document:
    - We select a topic from the document's topic distribution

LDA

$$\theta(t|d) \sim Dir(\alpha)$$

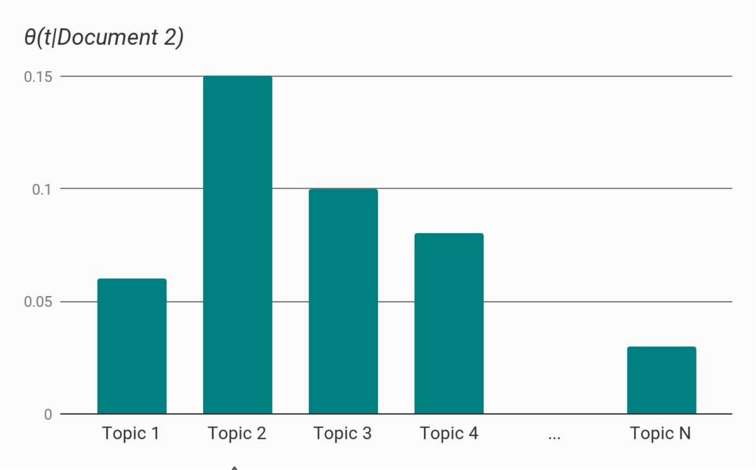
$$\varphi(w|t) \sim Dir(\beta)$$

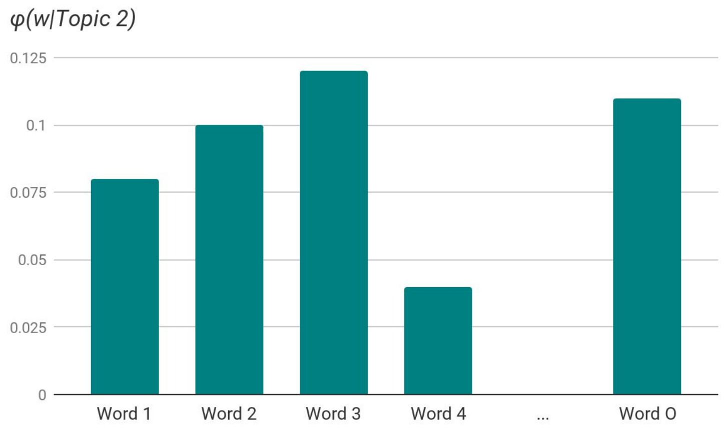
$$\nu(i,d) \sim \theta(t|d)$$

$$\omega(i,d) \sim \varphi(w|\nu(i,d))$$

LDA

Token 36 in Document 2 = ?





$v(36, D2) \sim \theta(t|D2)$

$\omega(36,D2) \sim \varphi(w|v(36,D2))$

LDA

- Documents are distributions of Topics
- Topics are distributions of Words
- To *write* a word in a document:
  - We select a topic from the document’s topic distribution
  - We select a word value from the topic’s word distribution

In Practice

$\theta(t \mid d)$?

$\varphi(w \mid t)$?

$\nu(i,d)$?

$\omega(i,d)$ ✓

Collapsed Gibbs Sampling

Collapsed Gibbs Sampling

- Collapsed ?
  - θ and φ are integrated out
- Sampling ?
  - Iteratively samples $v(i,d)$ (topic assignment) for every token i in every document d

Collapsed Gibbs Sampling

- Input
  - Lemmatised Documents (i.e. clean list of labels)
  - Number of topics desired
- Output
  - Document to topic distributions
  - Topic to word distributions

Collapsed Gibbs Sampling

- Markov chain Monte Carlo (MCMC) algorithm
  a. Randomly assign topics to all words in all documents
  b. Select a word in a document
  c. Remove topic assignment
  d. Build the new probabilities of topic assignment
  e. Assign new topic using probability
  f. Repeat for all words, for all documents
  g. Repeat multiple times until the model gets *stable*

Collapsed Gibbs Sampling

$$P(v(i,d) = t) \propto \frac{n^{\neg i,d}(t,d) + \alpha_t}{\sum_k n(k,d) + \alpha_k} \; x \; \frac{v^{\neg i,d}(\omega(i,d),t) + \beta_{\omega(i,d)}}{\sum_w v(w,t) + \beta_w}$$

Inspired by Jordan Boyd-Graber's lecture: https://www.youtube.com/watch?v=u7l5hhmdc0M

Collapsed Gibbs Sampling

$$P(v(i,d) = t) \propto \frac{n^{\neg i,d}(t,d) + \alpha_t}{\sum_k n(k,d) + \alpha_k} \times \frac{v^{\neg i,d}(\omega(i,d),t) + \beta_{\omega(i,d)}}{\sum_w v(w,t) + \beta_w}$$

- The probability that the observed token (*i* in document *d*) belongs to topic *t*

Inspired by Jordan Boyd-Graber's lecture: https://www.youtube.com/watch?v=u7l5hhmdc0M

Collapsed Gibbs Sampling

$$P(v(i,d) = t) \propto \frac{n^{\neg i,d}(t,d) + \alpha_t}{\sum_k n(k,d) + \alpha_k} \times \frac{v^{\neg i,d}(\omega(i,d),t) + \beta_{\omega(i,d)}}{\sum_w v(w,t) + \beta_w}$$

- Number of times document *d* uses topic *t* (minus current token)

Inspired by Jordan Boyd-Graber's lecture: https://www.youtube.com/watch?v=u7l5hhmdc0M

Collapsed Gibbs Sampling

$$P(v(i,d) = t) \propto \frac{n^{\neg i,d}(t,d) + \alpha_t}{\sum_k n(k,d) + \alpha_k} \times \frac{v^{\neg i,d}(\omega(i,d),t) + \beta_{\omega(i,d)}}{\sum_w v(w,t) + \beta_w}$$

- Dirichlet parameter for document to topic distribution

Inspired by Jordan Boyd-Graber's lecture: https://www.youtube.com/watch?v=u7l5hhmdc0M

Collapsed Gibbs Sampling

$$P(v(i,d) = t) \propto \frac{n^{\neg i,d}(t,d) + \alpha_t}{\sum_k n(k,d) + \alpha_k} \times \frac{v^{\neg i,d}(\omega(i,d),t) + \beta_{\omega(i,d)}}{\sum_w v(w,t) + \beta_w}$$

- In proportion, how much document *d* likes topic *t*

Inspired by Jordan Boyd-Graber's lecture: https://www.youtube.com/watch?v=u7l5hhmdc0M

Collapsed Gibbs Sampling

$$P(v(i,d) = t) \propto \frac{n^{\neg i,d}(t,d) + \alpha_t}{\sum_k n(k,d) + \alpha_k} \; x \; \frac{v^{\neg i,d}(\omega(i,d),t) + \beta_{\omega(i,d)}}{\sum_w v(w,t) + \beta_w}$$

- Number of times topic *t* uses word *ω(i,d)* (minus current token)

Inspired by Jordan Boyd-Graber's lecture: https://www.youtube.com/watch?v=u7l5hhmdc0M

Collapsed Gibbs Sampling

$$P(v(i,d) = t) \propto \frac{n^{\neg i,d}(t,d) + \alpha_t}{\sum_k n(k,d) + \alpha_k} \times \frac{v^{\neg i,d}(\omega(i,d),t) + \beta_{\omega(i,d)}}{\sum_w v(w,t) + \beta_w}$$

- Dirichlet parameter for topic to word distribution

Inspired by Jordan Boyd-Graber's lecture: https://www.youtube.com/watch?v=u7l5hhmdc0M

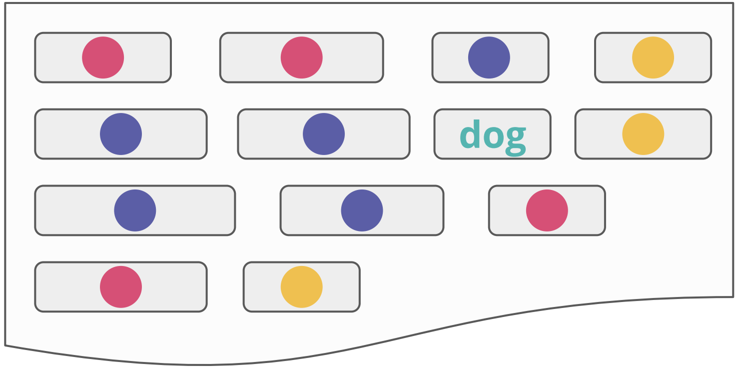
Collapsed Gibbs Sampling

$$P(v(i,d) = t) \propto \frac{n^{\neg i,d}(t,d) + \alpha_t}{\sum_k n(k,d) + \alpha_k} \times \frac{v^{\neg i,d}(\omega(i,d),t) + \beta_{\omega(i,d)}}{\sum_w v(w,t) + \beta_w}$$

- In proportion, how much topic *t* likes word *ω(i,d)*

Inspired by Jordan Boyd-Graber's lecture: https://www.youtube.com/watch?v=u7l5hhmdc0M

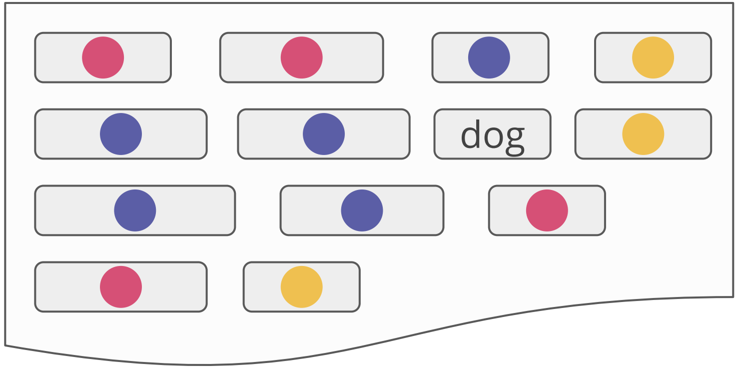
Collapsed Gibbs Sampling



	Topic 0	Topic 1	Topic 2	Topic 3
	…			
dog	12	3	8	7
	…			

Inspired by Jordan Boyd-Graber's lecture: https://www.youtube.com/watch?v=u7l5hhmdc0M

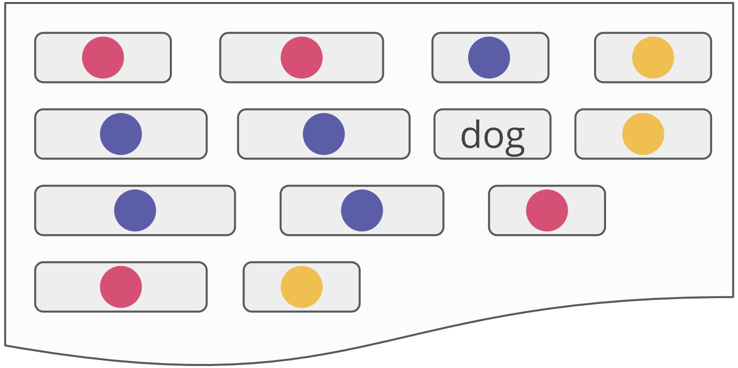
Collapsed Gibbs Sampling



	Topic 0	Topic 1	Topic 2	Topic 3
		…		
dog	12	3	8	~~7~~ 6
		…		

Inspired by Jordan Boyd-Graber's lecture: https://www.youtube.com/watch?v=u7l5hhmdc0M

Collapsed Gibbs Sampling



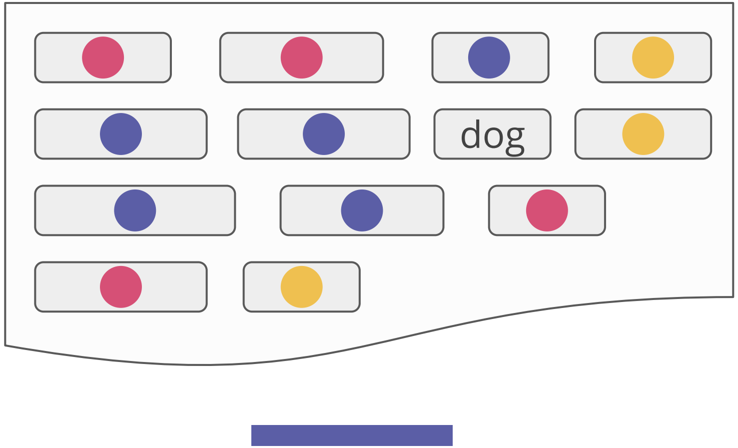
	Topic 0	Topic 1	Topic 2	Topic 3
…				
dog	12	3	8	6
…				

How much document likes topic

How much topic likes word

Inspired by Jordan Boyd-Graber's lecture: https://www.youtube.com/watch?v=u7l5hhmdc0M

Collapsed Gibbs Sampling



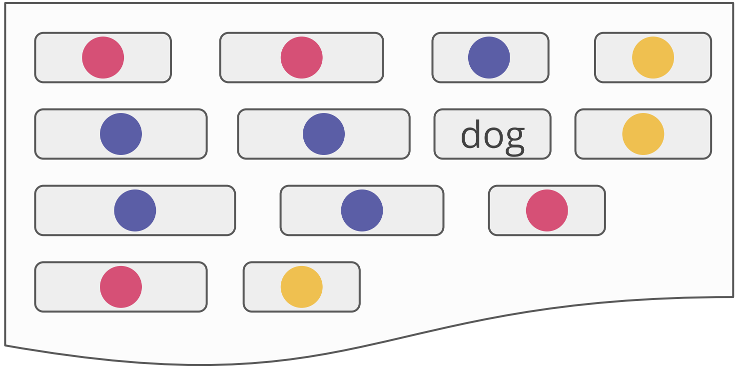
	Topic 0	Topic 1	Topic 2	Topic 3
		...		
dog	12	3	8	6
		...		

How much document likes topic

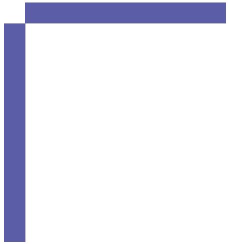
How much topic likes word

Inspired by Jordan Boyd-Graber's lecture: https://www.youtube.com/watch?v=u7l5hhmdc0M

Collapsed Gibbs Sampling



	Topic 0	Topic 1	Topic 2	Topic 3
…				
dog	12	3	8	6
…				

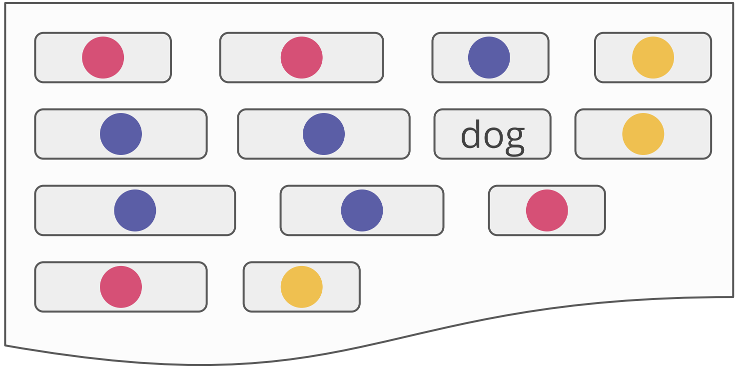


How much document likes topic

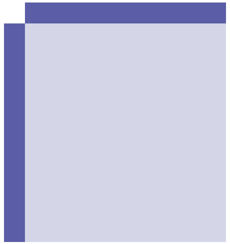
How much topic likes word

Inspired by Jordan Boyd-Graber's lecture: https://www.youtube.com/watch?v=u7l5hhmdc0M

Collapsed Gibbs Sampling



	Topic 0	Topic 1	Topic 2	Topic 3
...				
dog	12	3	8	6
...				



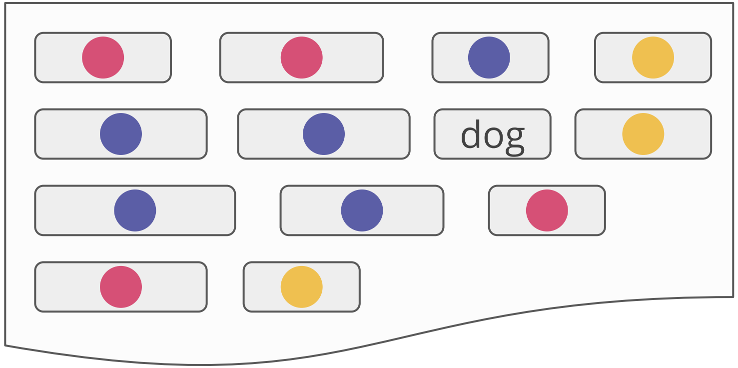
P(v(7,d) = T0)

How much document likes topic

How much topic likes word

Inspired by Jordan Boyd-Graber's lecture: https://www.youtube.com/watch?v=u7l5hhmdc0M

Collapsed Gibbs Sampling

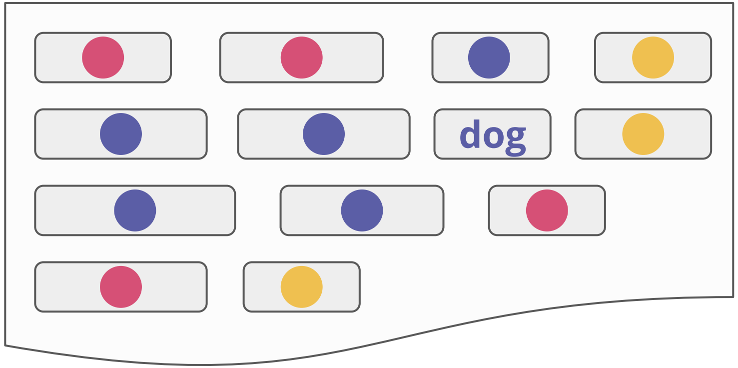


	Topic 0	Topic 1	Topic 2	Topic 3
		…		
dog	12	3	8	6
		…		

P(v(7,d) = T0) *P(v(7,d) = T1)* *P(v(7,d) = T2)* *P(v(7,d) = T3)*

Inspired by Jordan Boyd-Graber's lecture: https://www.youtube.com/watch?v=u7l5hhmdc0M

Collapsed Gibbs Sampling



	Topic 0	Topic 1	Topic 2	Topic 3
...				
dog	~~12~~ 13	3	8	6
...				

Inspired by Jordan Boyd-Graber's lecture: https://www.youtube.com/watch?v=u7l5hhmdc0M

Results

Document to Topic Matrix

	T_0	T_1	...	T_N
D_0	0.2	0.09		0.03
D_1	0.03	0.13		0.11
D_2	0.04	0.12		0.08
...				
D_M	0.09	0.01		0.2

Topic to Word Matrix

	W_0	W_1	...	W_O
T_0	0.02	0.01		0.13
T_1	0.01	0.02		0.14
T_2	0.09	0.01		0.03
...				
T_N	0.11	0.23		0.01

Topic Model Data

Topic Model Data

- Orders of Magnitude:

Topics << Documents <<<< Words

	~100	~10,000	~100,000
For the User:	Reasonable	Too much	

Topics to Words Matrix

Topic 1		
	Label	Weight
W_0	book	0.02
W_1	tulip	0.01
...		
W_O	fox	0.16

Topic 2		
	Label	Weight
W_0	book	0.12
W_1	tulip	0.02
...		
W_O	fox	0.001

Topics Labels

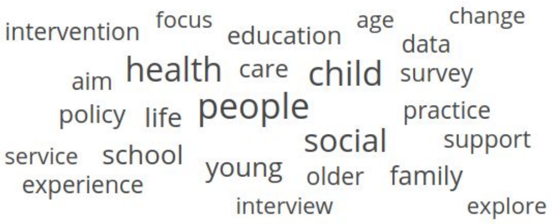
Topic 1		
	Label	Weight
W_A	dog	0.25
W_B	canine	0.2
W_c	fox	0.16
...		

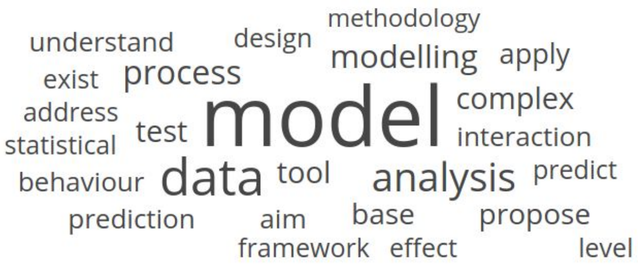
Topic 2		
	Label	Weight
W_D	page	0.18
W_E	ink	0.16
W_F	book	0.12
...		

Topics Labels


```
> dataModel
<· ▼{topics: Array(30), topicsAsocIndex: {…}, refEntries: Array(1819)}
  ▼topics: Array(30)
    ▼0:
      topicNumber: 0
      ▼words:
        first3words: "food-environmental-science"
        ▼wordCloudAsArrayOfObjects: Array(20)
          ▶0: {weight: 1639, label: "food"}
          ▶1: {weight: 1373, label: "environmental"}
          ▶2: {weight: 1211, label: "science"}
          ▶3: {weight: 825, label: "research"}
          ▶4: {weight: 704, label: "plant"}
          ▶5: {weight: 689, label: "facility"}
          ▶6: {weight: 653, label: "environment"}
          ▶7: {weight: 644, label: "marine"}
          ▶8: {weight: 627, label: "change"}
          ▶9: {weight: 622, label: "conservation"}
          ▶10: {weight: 602, label: "animal"}
          ▶11: {weight: 587, label: "development"}
          ▶12: {weight: 569, label: "ecology"}
          ▶13: {weight: 500, label: "nerc"}
          ▶14: {weight: 487, label: "management"}
          ▶15: {weight: 475, label: "work"}
          ▶16: {weight: 469, label: "ecosystem"}
          ▶17: {weight: 442, label: "soil"}
          ▶18: {weight: 441, label: "include"}
          ▶19: {weight: 420, label: "agriculture"}
           length: 20
          ▶__proto__: Array(0)
        ▶__proto__: Object
      ▶topDocuments: {fullInfo: Array(100)}
      ▶similarities: (30) [1, 0, 0, 0, 0, 0, 0, 0, 0.07207604062325, 0.00797398165
      ▶__proto__: Object
    ▶1: {topicNumber: 1, words: {…}, topDocuments: {…}, similarities: Array(30)}
    ▶2: {topicNumber: 2, words: {…}, topDocuments: {…}, similarities: Array(30)}
```

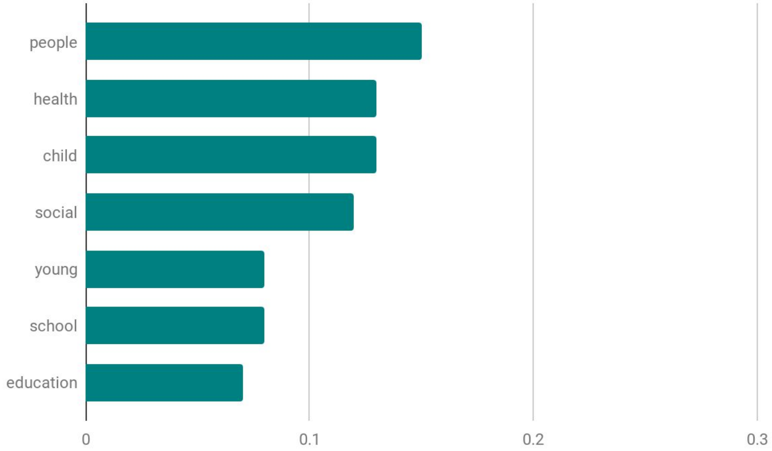
Topics Labels

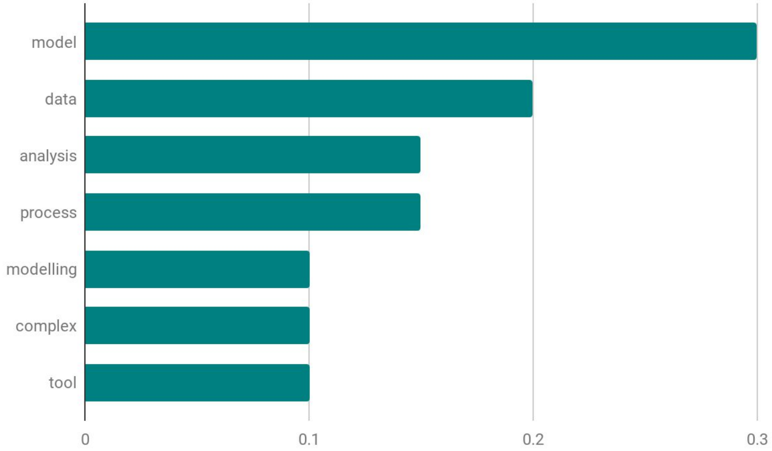




Source: [Strategic Futures Laboratory]()

Topics Labels





Documents to Topics Matrix

	T_0	T_1	...	T_N
D_0	0.2	0.09		0.03
D_1	0.03	0.13		0.11
D_2	0.04	0.12		0.08
...				
D_M	0.09	0.01		0.2

Topics to Documents Matrix

	D_0	D_1	...	D_M
T_0	0.2	0.03		0.09
T_1	0.09	0.13		0.01
T_2	0.05	0.08		0.14
...				
T_N	0.03	0.11		0.2

- Top documents per topic

Topics to Documents Matrix

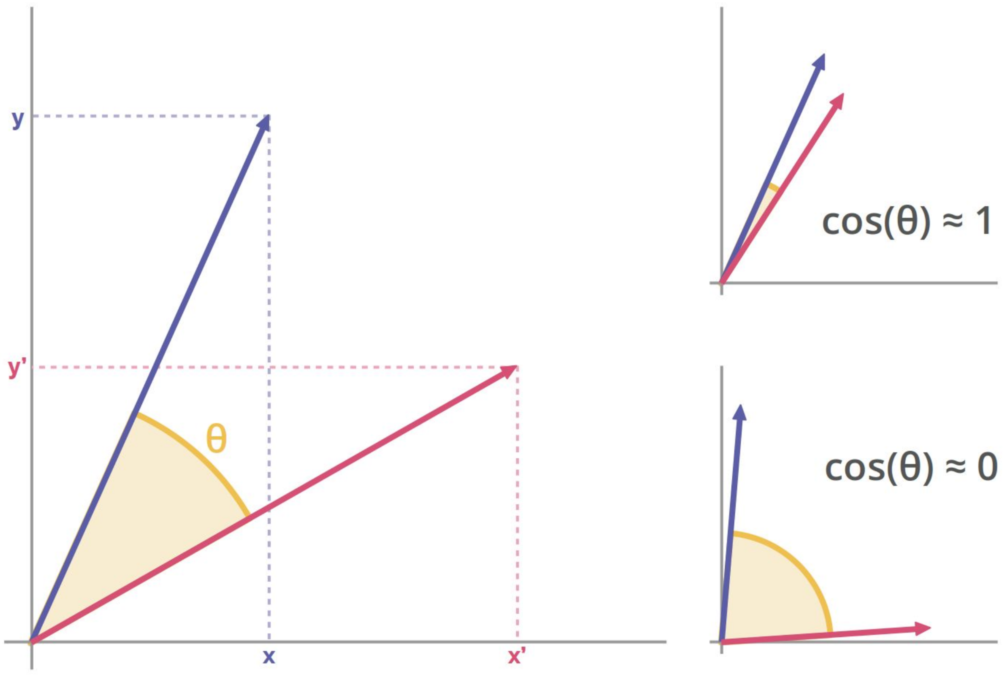

```
> dataModel
<· ▼{topics: Array(30), topicsAsocIndex: {…}, refEntries: Array(1819)} i
  ▼topics: Array(30)
    ▼0:
      topicNumber: 0
     ▶words: {first3words: "food-environmental-science", wordCloudAsArrayOfObjects: Array(20)}
     ▼topDocuments:
       ▼fullInfo: Array(100)
         ▼0:
            topicWeight: 0.5197057277982133
            wordCount: 3972
            docID: "10007799-6-"
           ▶docInfo: {UoAString: " Agriculture, Veterinary and Food Science", Institution name: "Newcas
           ▶__proto__: Object
         ▶1: {topicWeight: 0.5148090413094311, wordCount: 5242, docID: "10007857-6-", docInfo: {…}}
         ▶2: {topicWeight: 0.5074045206547155, wordCount: 5242, docID: "10007856-6-", docInfo: {…}}
         ▶3: {topicWeight: 0.4523875241512559, wordCount: 3748, docID: "10007804-6-", docInfo: {…}}
         ▶4: {topicWeight: 0.43951985226223456, wordCount: 3286, docID: "10007822-6-", docInfo: {…}}
         ▶5: {topicWeight: 0.4359951845906902, wordCount: 5087, docID: "10007802-6-", docInfo: {…}}
         ▶6: {topicWeight: 0.4173706441393875, wordCount: 3932, docID: "10007857-7-", docInfo: {…}}
         ▶7: {topicWeight: 0.41129883843717, wordCount: 3932, docID: "10007856-7-", docInfo: {…}}
         ▶8: {topicWeight: 0.39144845873384154, wordCount: 3091, docID: "10040812-6-", docInfo: {…}}
```

Topics to Documents Matrix

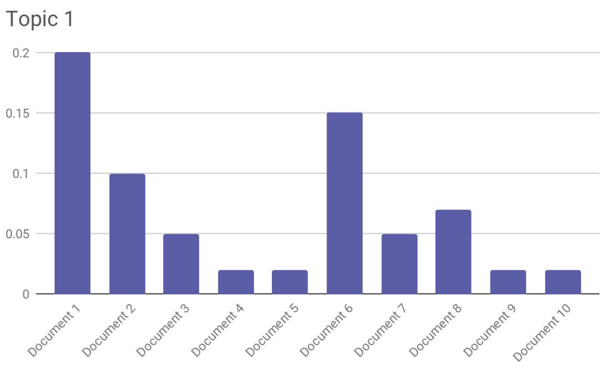
	D_0	D_1	...	D_M
T_0	0.2	0.03		0.09
T_1	0.09	0.13		0.01
T_2	0.05	0.08		0.14
...				
T_N	0.03	0.11		0.2

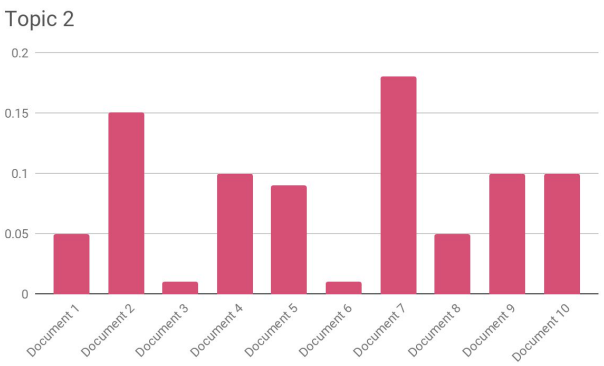
- Top documents per topic
- Document vectors per topic

Cosine Similarity

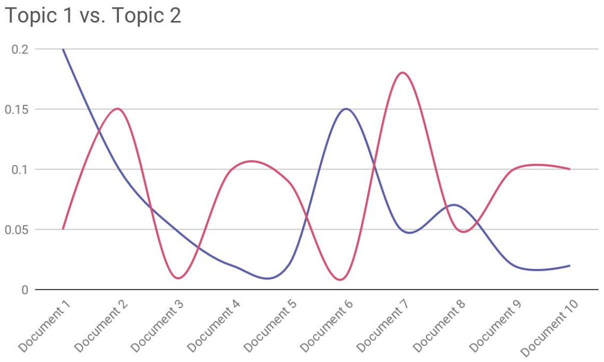


Cosine Similarity

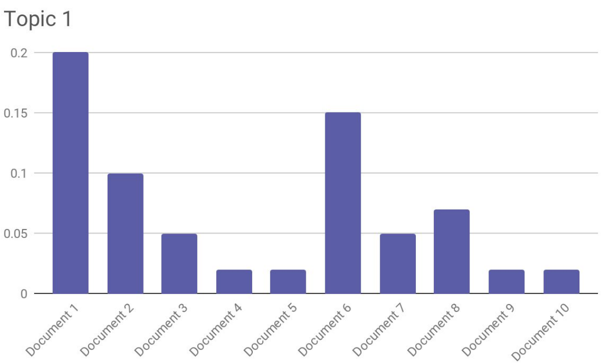


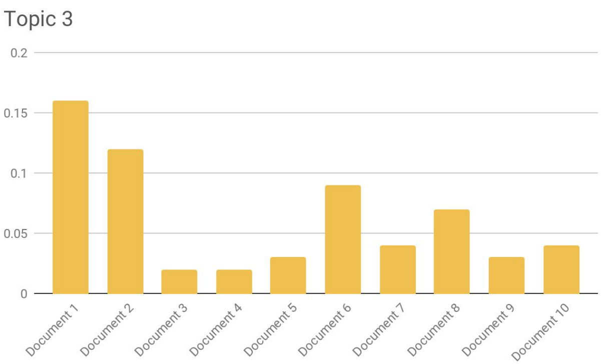


Similarity ~ 52%

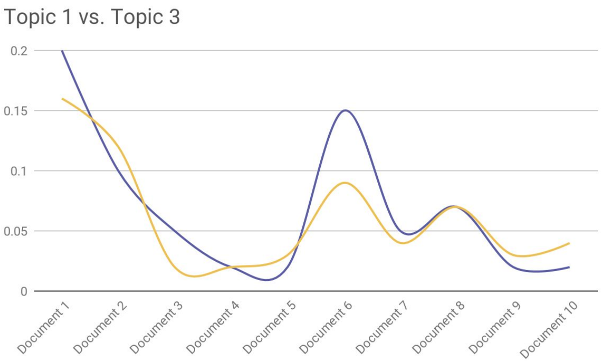


Cosine Similarity





Similarity ~ 96%



Topics to Documents Matrix

	D_0	D_1	...	D_M
T_0	0.2	0.03		0.09
T_1	0.09	0.13		0.01
T_2	0.05	0.08		0.14
...				
T_N	0.03	0.11		0.2

- Top documents per topic
- Document vectors per topic
  - ➜ Topic to Topic Similarity

Topics to Topic Similarity


```
> dataModel
<· ▼{topics: Array(30), topicsAsocIndex: {…}, refEntries: Array(1819)}
  ▼topics: Array(30)
    ▼0:
      topicNumber: 0
     ▶words: {first3words: "food-environmental-science", wordCloudAsArrayOfObjects: Array(20)}
     ▶topDocuments: {fullInfo: Array(100)}
     ▼similarities: Array(30)
        0: 1
        1: 0
        2: 0
        3: 0
        4: 0
        5: 0
        6: 0
        7: 0
        8: 0.07207604062325
        9: 0.00797398165098
        10: 0.02409086293046
        11: 0
        12: 0.00553670267199
        13: 0.01827718882005
        14: 0.08280989828378
        15: 0
        16: 0.20354517225645
        17: 0.02644391490685
        18: 0.10612095996277
        19: 0.00466132686394
        20: 0
        21: 0
        22: 0
        23: 0
        24: 0.00734715054184
        25: 0.00928436669004
        26: 0
        27: 0.01919855096201
        28: 0
        29: 0
        length: 30
       ▶__proto__: Array(0)
     ▶__proto__: Object
   ▶1: {topicNumber: 1, words: {…}, topDocuments: {…}, similarities: Array(30)}
```

Topics to Documents Matrix

	D_0	D_1	...	D_M
T_0	0.2	0.03		0.09
T_1	0.09	0.13		0.01
T_2	0.05	0.08		0.14
...				
T_N	0.03	0.11		0.2

- Top documents per topic
- Document vectors per topic
  - ➜ Topic to Topic Similarity
    - ➜ Clusters
    - ➜ Visual Layouts

Agglomerative Clustering

	A	B	C	D	E	
A	1	0.2	0.6	0.1	0.3	
B	0.2	1	0.3	0.7	0.5	
C	0.6	0.3	1	0.2	0.4	
D	0.1	0.7	0.2	1	0.4	
E	0.3	0.5	0.4	0.4	1	

Agglomerative Clustering

	A	B	C	D	E	
A	1	0.2	0.6	0.1	0.3	
B	0.2	1	0.3	**0.7**	0.5	
C	0.6	0.3	1	0.2	0.4	
D	0.1	**0.7**	0.2	1	0.4	
E	0.3	0.5	0.4	0.4	1	

Agglomerative Clustering

	A	B	C	D	E	(B,D)
A	1	0.2	0.6	0.1	0.3	0.1
B	0.2	1	0.3	**0.7**	0.5	
C	0.6	0.3	1	0.2	0.4	0.2
D	0.1	**0.7**	0.2	1	0.4	
E	0.3	0.5	0.4	0.4	1	0.4
(B,D)	0.1		0.2		0.4	1

Agglomerative Clustering

	A	B	C	D	E	(B,D)
A	1		0.6		0.3	0.1
B						
C	0.6		1		0.4	0.2
D						
E	0.3		0.4		1	0.4
(B,D)	0.1		0.2		0.4	1

Agglomerative Clustering

	A	C	E	(B,D)	
A	1	0.6	0.3	0.1	
C	0.6	1	0.4	0.2	
E	0.3	0.4	1	0.4	
(B,D)	0.1	0.2	0.4	1	

B, D, 0.7

Agglomerative Clustering

	E	(B,D)	(A,C)	
E	1	0.4	0.3	
(B,D)	0.4	1	0.1	
(A,C)	0.3	0.1	1	

B, D, 0.7

A, C, 0.6

Agglomerative Clustering

	(A,C)	((B,D),E)	
(A,C)	1	0.1	
((B,D),E)	0.1	1	

B, D, 0.7

A, C, 0.6

(B,D), E, 0.4

Agglomerative Clustering

	((B,D),E),(A,C))	
((B,D),E),(A,C))	1	

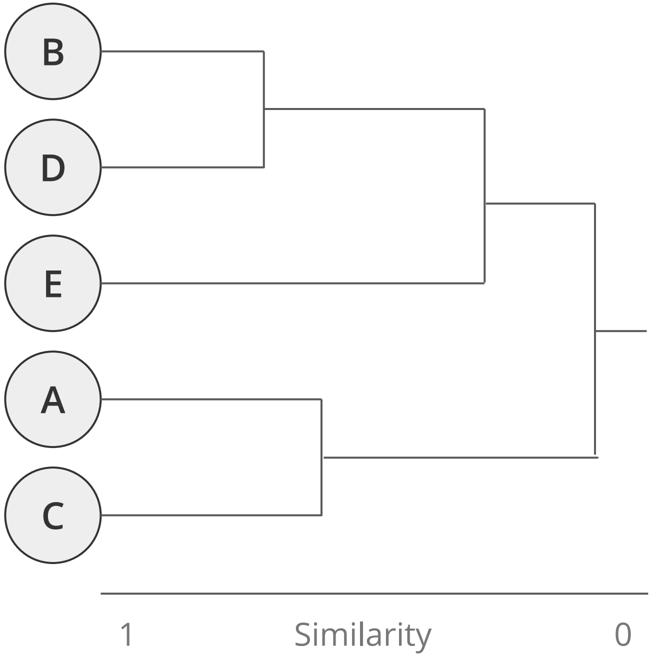
B, D, 0.7

A, C, 0.6

(B,D), E, 0.4

((B,D),E), (A,C), 0.1

Agglomerative Clustering



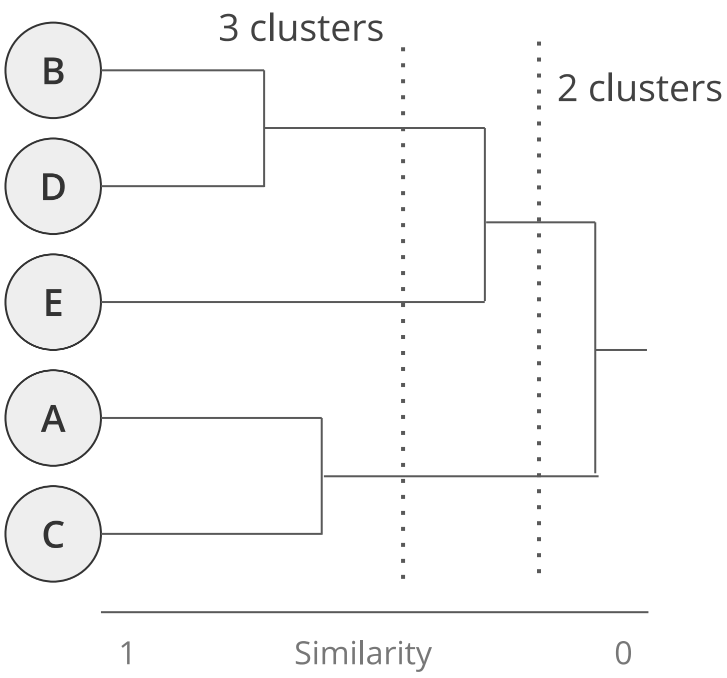
B, D, 0.7

A, C, 0.6

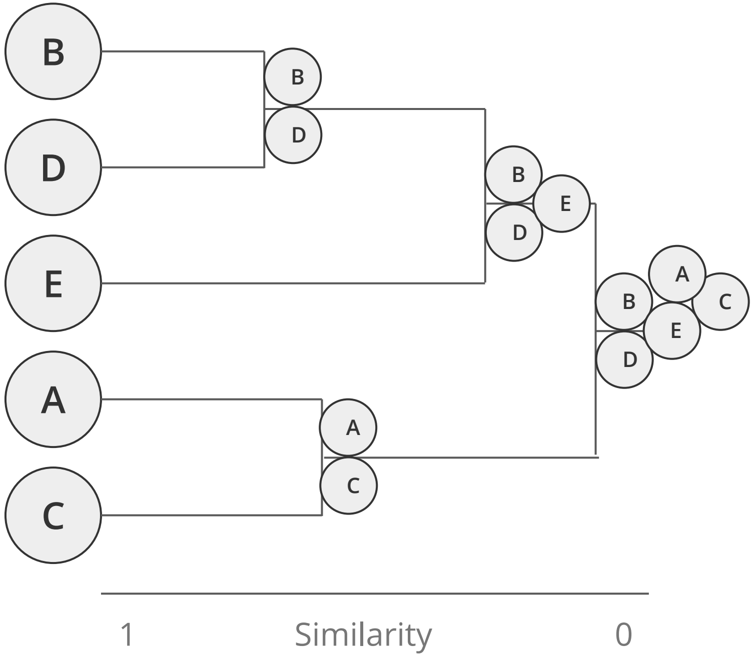
(B,D), E, 0.4

((B,D),E), (A,C), 0.1

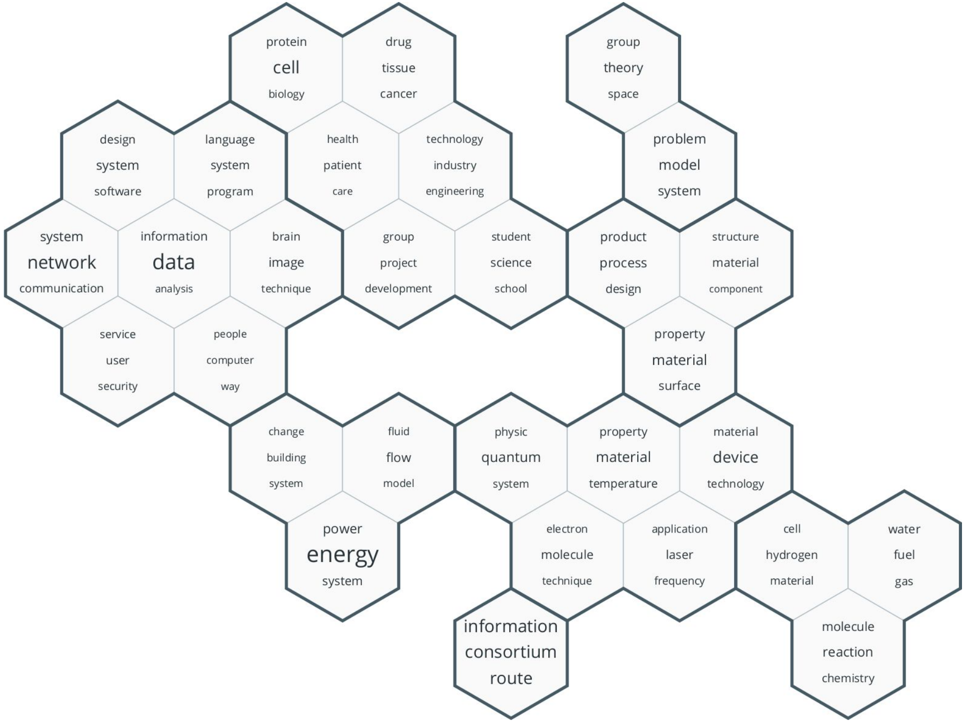
Agglomerative Clustering



Agglomerative Layout

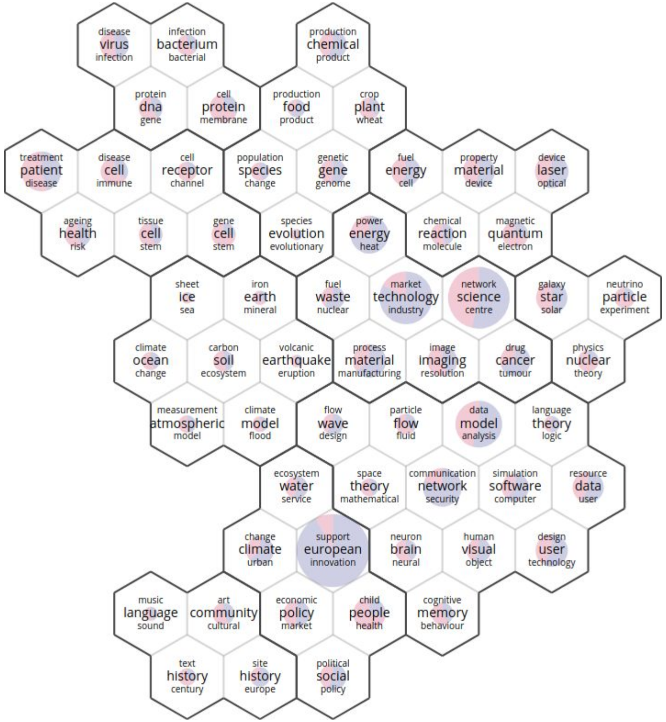


Topic Maps



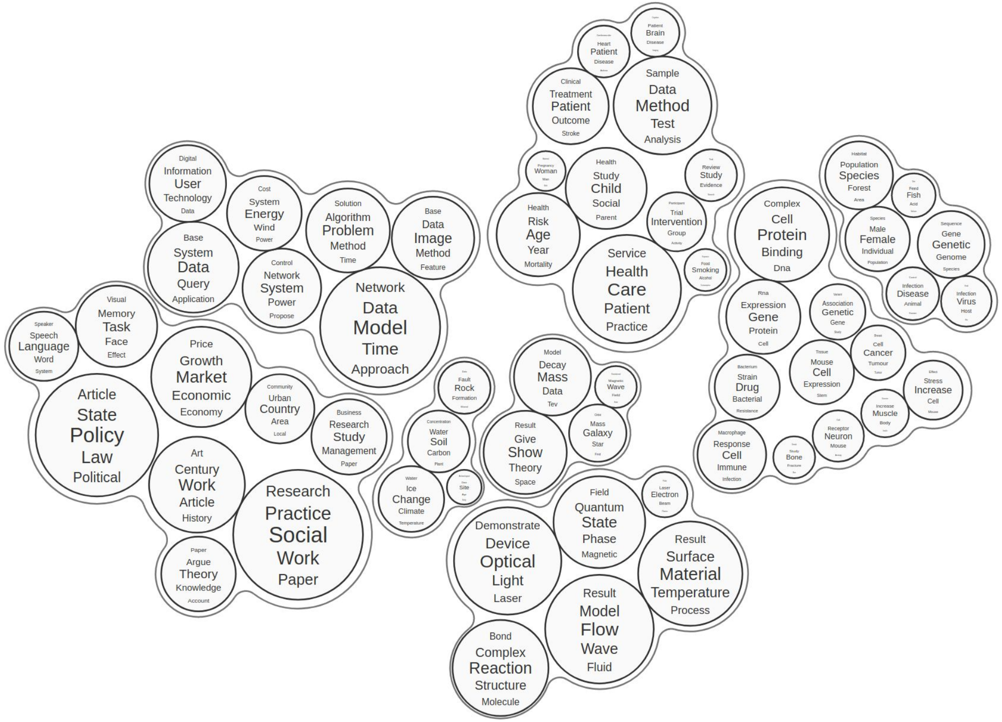
Source: [Strategic Futures Laboratory]()

Topic Maps



Source: [Strategic Futures Laboratory]

Topic Maps



Source: [Strategic Futures Laboratory]

Conclusion

- Topic Modelling Unsupervised classification of documents into themes
- LDA Describes a document generative model
- Collapsed Gibbs Sampling MCMC algorithm sampling document to topic distributions and topic to word distributions
- Agglomerative Clustering Builds a hierarchy from similarity data