Probabilistic Topic Models

Mark Steyvers – University of California, Irvine
Tom Griffiths – Brown University

Presented by Drew Buck
9/25/2014

Outline

- What are topic models?
  - Generative models
  - Probabilistic Topic Models
- How to extract topics from documents?
  - Gibbs sampling algorithm
  - Examples
- Applications
  - Information retrieval
  - Word association

Topic Models

- Consider a corpus of many documents…
  - Documents contain mixtures of topics
  - Topics are distributions over words
- Topic models are generative models
  - New documents can be generated if the statistical parameters are known
  - The parameters can also be estimated

Topics

Topics are distributions over words.

Topic 247

word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

Topic 5

word	prob.
RED	.202
BLUE	.099
GREEN	.096
YELLOW	.073
WHITE	.048
COLOR	.048
BRIGHT	.030
COLORS	.029
ORANGE	.027
BROWN	.027
PINK	.017
LOOK	.017
BLACK	.016
PURPLE	.015
CROSS	.011
COLORED	.009

Topic 43

word	prob.
MIND	.081
THOUGHT	.066
REMEMBER	.064
MEMORY	.037
THINKING	.030
PROFESSOR	.028
FELT	.025
REMEMBERED	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENT	.020
THINK	.019
THING	.016
WONDER	.014
FORGET	.012
RECALL	.012

Topic 56

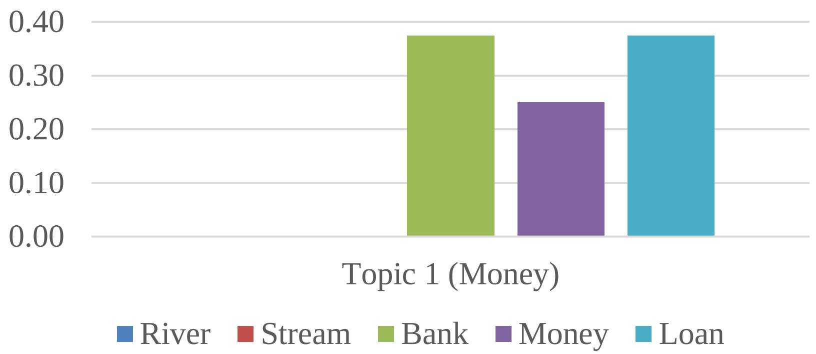
word	prob.
DOCTOR	.074
DR.	.063
PATIENT	.061
HOSPITAL	.049
CARE	.046
MEDICAL	.042
NURSE	.031
PATIENTS	.029
DOCTORS	.028
HEALTH	.025
MEDICINE	.017
NURSING	.017
DENTAL	.015
NURSES	.013
PHYSICIAN	.012
HOSPITALS	.011

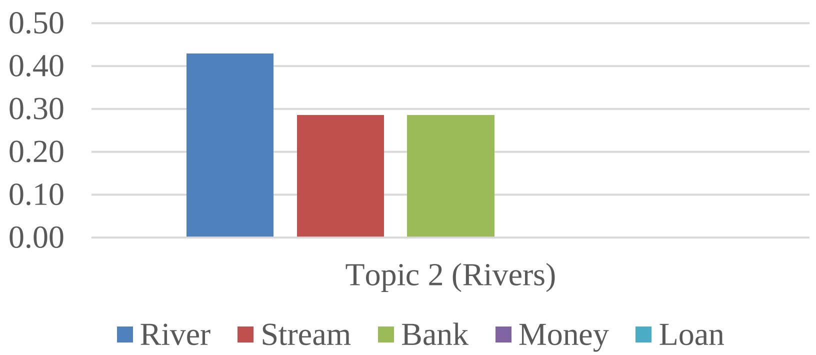
Generative Models

When the distributions in the model are known, documents can be generated by sampling.

Consider two topics:

- Money
  - Money, Bank, Loan
- Rivers
  - River, Bank, Stream



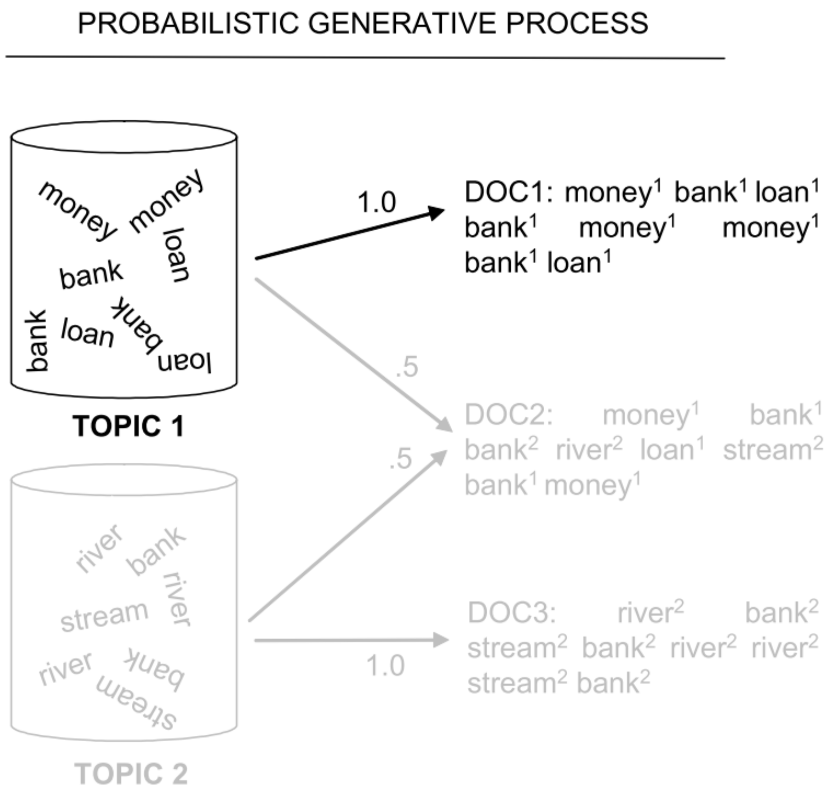


Generative Models

For Document 1, only Topic 1 is used to sample words.

Each word is sampled independently.

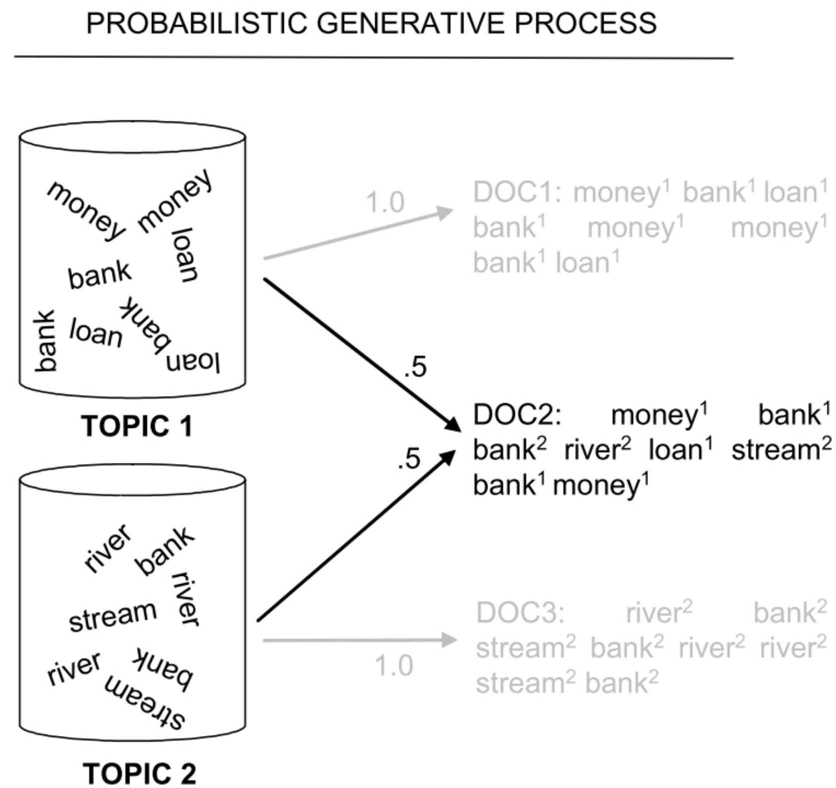
<u>Bag-of-words assumption</u>



Generative Models

For Document 2, both topics are chosen with equal probability.

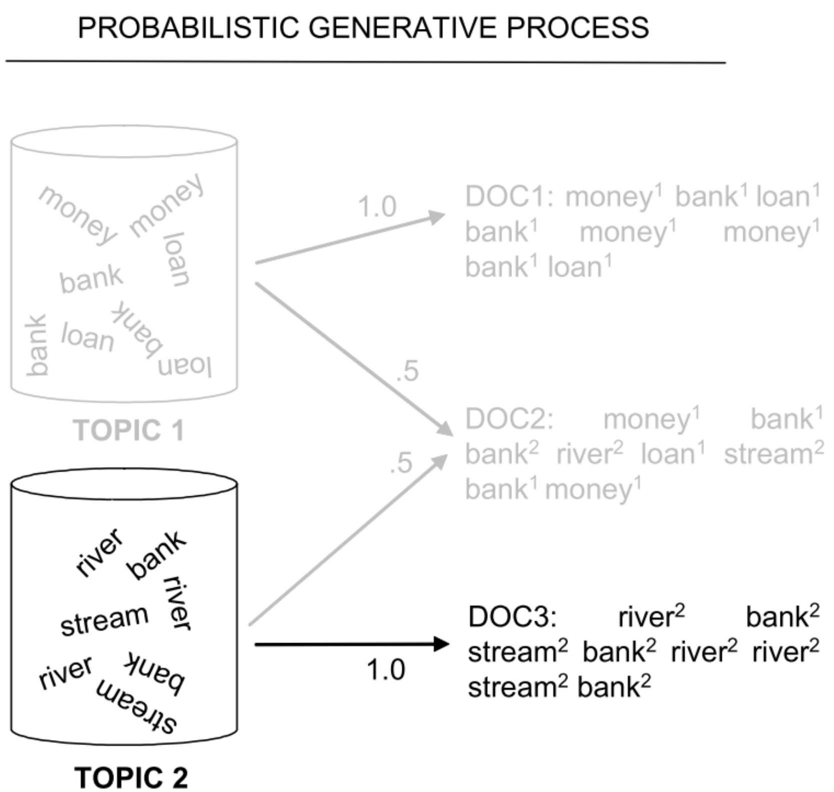
First, a topic is chosen. Then, a word is sampled from the topic's distribution over words.



Generative Models

For Document 3, only Topic 2 is used.

Words having multiple meanings (polysemy) can appear in multiple topics.

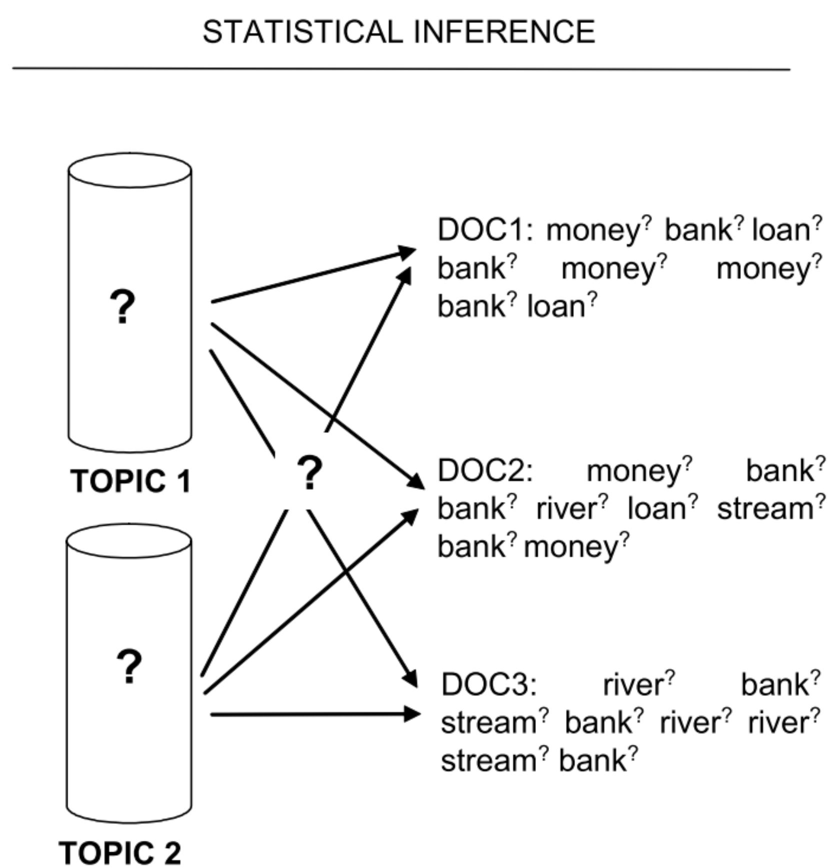


Generative Models

Now, suppose that we don't know the topics.

We want to determine:

- What is the distribution over words for each topic?
- Which topics appear in each document?



Probabilistic Topic Models

Notation:

$P(z)$	Distribution over topics z in a particular document
$P(w\|z)$	Distribution over words w given topic z
$P(z_i = j)$	Probability that the j^{th} topic was sampled for the i^{th} word token
$P(w_i\|z_i = j)$	Probability of word w_i under topic j

Probabilistic Topic Models

Distribution over words within a document:

$$P(w_i) = \sum_{j=1}^{T} P(w_i | z_i = j) P(z_i = j)$$

where T is the number of topics.

Let:

- $\phi^{(j)} = P(w|z = j) =$ multinomial distribution over words for topic j
- $\theta^{(d)} = P(z) =$ multinomial distribution over topics for document d
- $D =$ number of documents, each containing N_d word tokens
- $N =$ total number of word tokens (i.e., $N = \Sigma\, N_d$)

Probabilistic Topic Models

- ϕ and θ are both multinomial distributions.
  - ϕ indicates which words are important for a particular topic.
  - θ indicates which topics are important for a particular document.

What is the domain of possible distributions for ϕ and θ?

Consider the multinomial distribution $p = (p_1, \ldots, p_T)$.

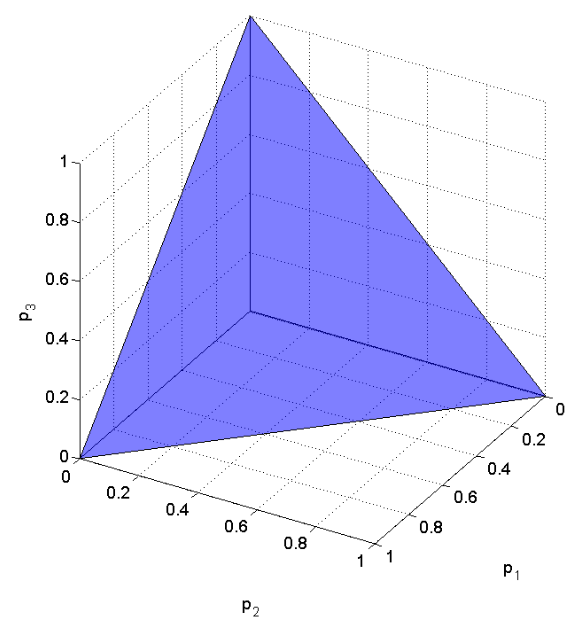
To be a probability distribution, we must have $\sum_j p_j = 1$.

Probabilistic Topic Models

Given $p = (p_1, \dots, p_T)$, there are T parameters to define.

In T-dimensional space, the points that satisfy $\sum_j p_j = 1$ form a (T-1)-dimensional probability simplex.

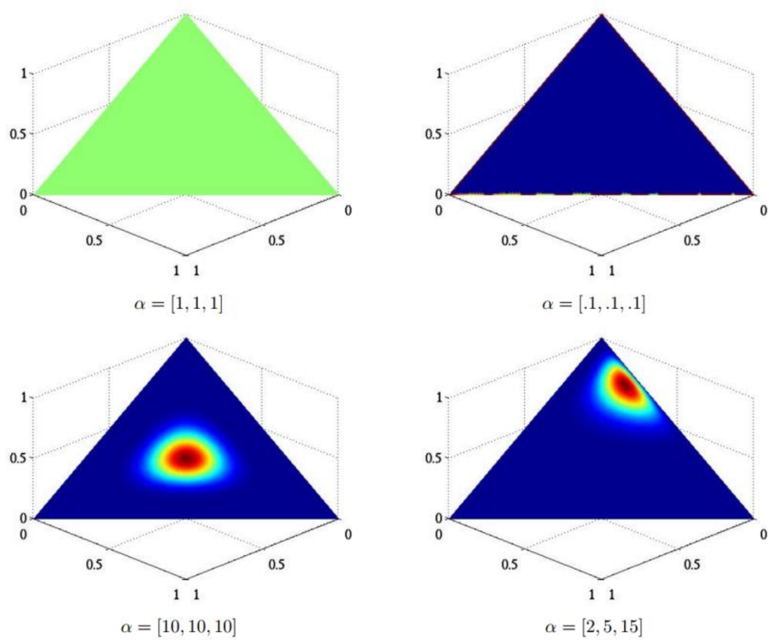
Points on this simplex are valid probability distributions.

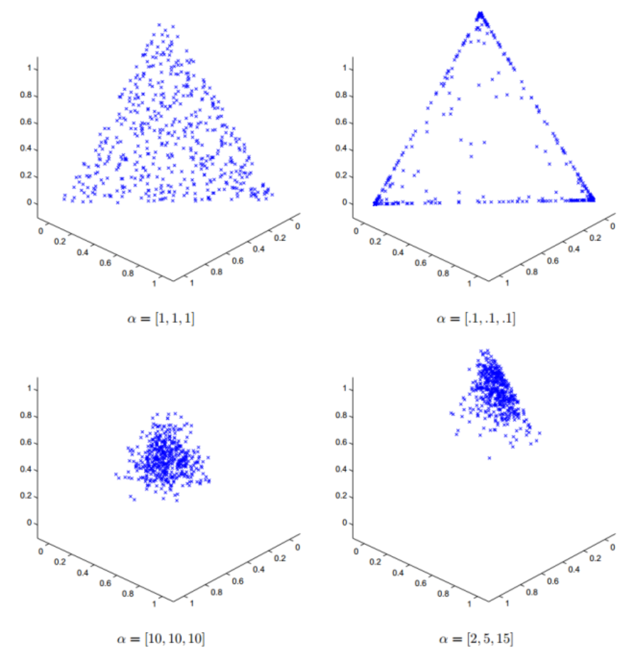


Dirichlet Distribution

The probability density of a T dimensional Dirichlet distribution over the multinomial distribution $p = (p_1, \dots, p_T)$ is defined by:

$$\mathrm{Dir}(\alpha_1, \dots, \alpha_T) = \frac{\Gamma\left(\sum_j \alpha_j\right)}{\prod_j \Gamma\left(\alpha_j\right)} \prod_{j=1}^{T} p_j^{\alpha_j - 1}$$

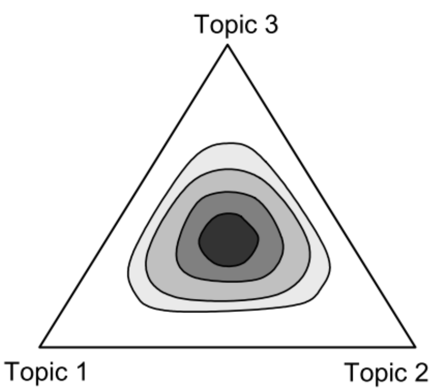




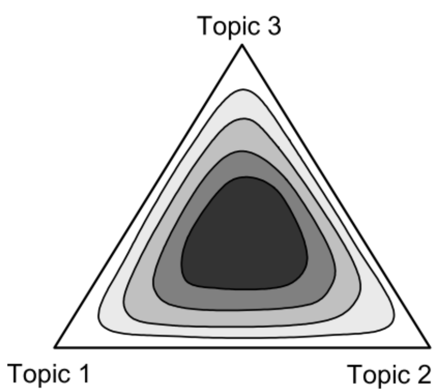
Bela A. Frigyik, Amol Kapila, and Maya R. Gupta, *Introduction to the Dirichlet Distribution and Related Processes*

Dirichlet Distribution

- The parameters $\alpha_1 \dots \alpha_T$ define the distribution.
- For convenience, we set $\alpha_1 = \alpha_2 = \cdots = \alpha_T = \alpha$.
  - Larger values for α give more smoothing (away from corners).
  - For $\alpha < 1$, the modes are located at the corners of the simplex, favoring topic distributions with only a few topics.



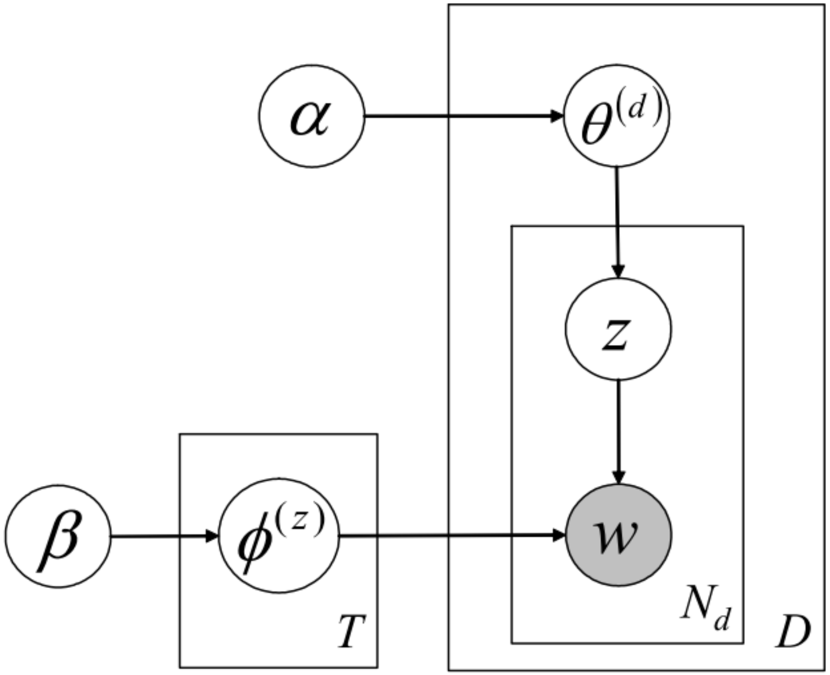
$\alpha = 4$



$\alpha = 2$

Graphical Model

- We use a symmetric $\mathrm{Dirichlet}(\alpha)$ prior on θ.
  - Represents the prior observation count for topics within documents.
- We also use a symmetric $\mathrm{Dirichlet}(\beta)$ prior on ϕ.
  - Represents the prior observation count for words within topics.
- Suggested Values
  - $\alpha = 50/T$
  - $\beta = 0.01$



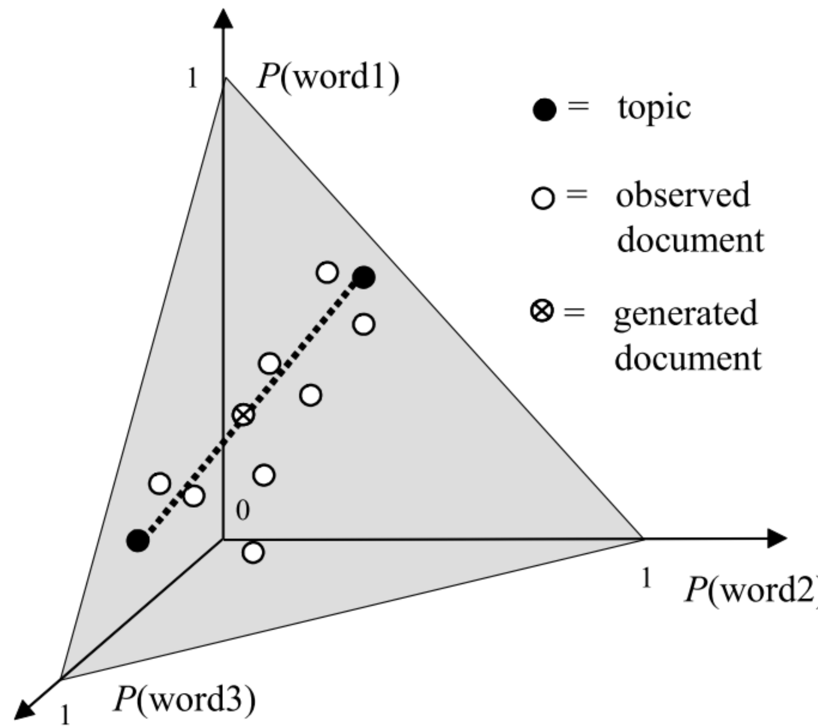
Geometric Interpretation

Imagine a W-dimensional space where each axis represents the probability of observing word w.

Points on the (W-1)-dimensional simplex represent probability distributions over words.

Each generated document lies on the (T-1)-dimensional subsimplex.

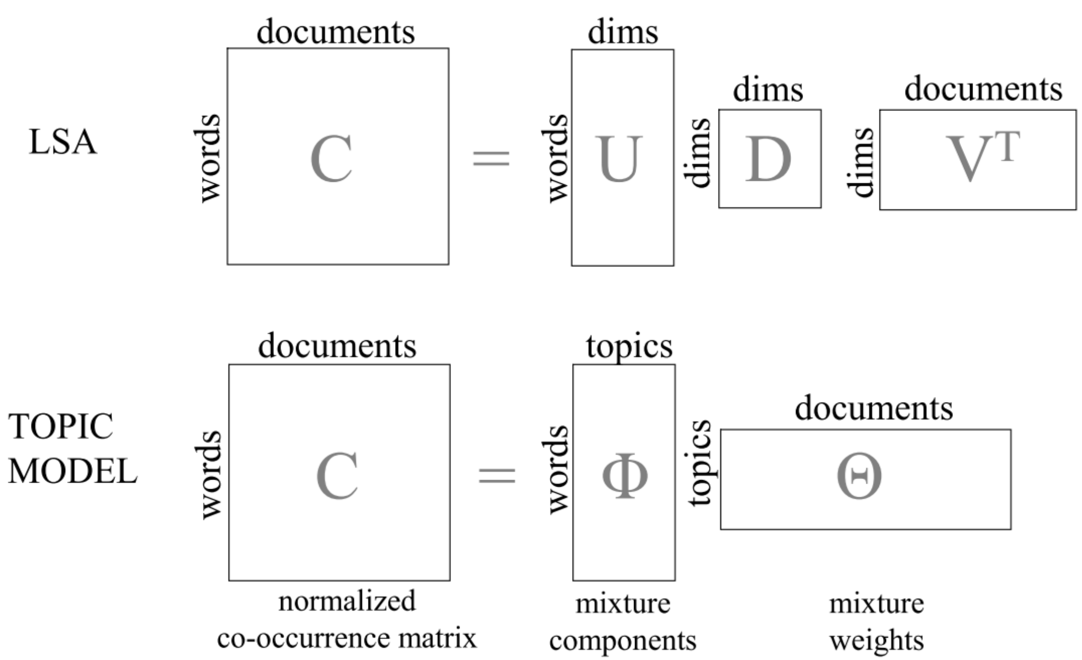
When $T \ll W$, this can be thought of as dimensionality reduction.



Matrix Factorization Interpretation

In Latent Semantic Analysis (LSA), the word document co-occurrence matrix C is decomposed using singular value decomposition.

In our model, C is split into a topic matrix Φ and a document matrix Θ.



Algorithm for Extracting Topics

- Approach:
  - Estimate the posterior distribution over z (the assignment of word tokens to topics) given the observed words $\boldsymbol{w}$, while marginalizing out ϕ and θ.
  - Use a Gibbs sampling algorithm to sequentially sample from the posterior distribution of z.
  - Continue generating samples until the sampled values approximate the target distribution.
  - Compute estimates of ϕ and θ using the posterior estimates of z.

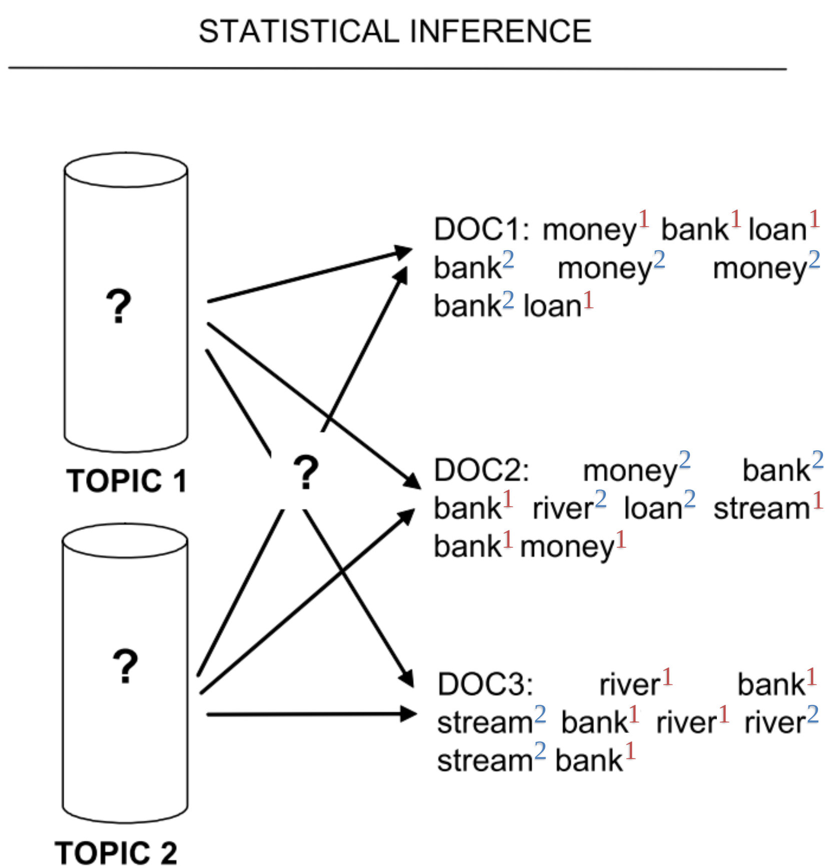
Algorithm for Extracting Topics

Initialization:

- Assign each word token to a random topic in $[1 \ldots T]$.
- Compute the count matrices C^{WT} and C^{DT}.

C^{WT}	Topic 1	Topic 2
River	2	2
Stream	1	2
Bank	6	3
Money	2	3
Loan	2	1

C^{DT}	Topic 1	Topic 2
DOC1	4	4
DOC2	4	4
DOC3	5	3



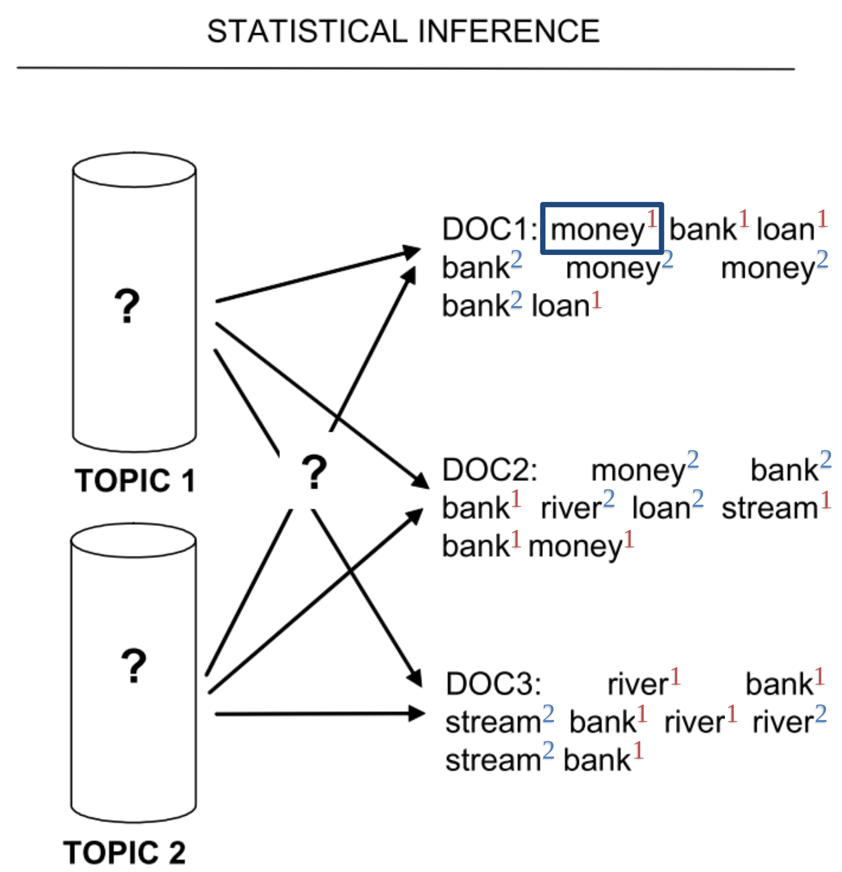
Algorithm for Extracting Topics

Gibbs Sampling Algorithm:

- Look at each word token in turn
- Decrement the corresponding entries in C^{WT} and C^{DT}

C^{WT}	Topic 1	Topic 2
River	2	2
Stream	1	2
Bank	6	3
Money	$\boxed{1}$	3
Loan	2	1

C^{DT}	Topic 1	Topic 2
DOC1	$\boxed{3}$	4
DOC2	4	4
DOC3	5	3



Algorithm for Extracting Topics

- Gibbs Sampling Algorithm:
  - Estimate the posterior distribution over z_i

$$P(z_i = j | z_{-i}, w_i, d_i, \cdot) \propto \frac{C^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha}$$

z_{-i} refers to the topic assignments of all other word tokens

w_i is the current word token

d_i is the current document

$\cdot$ refers to all other known information, such as all other word and document indices w_{-i} and d_{-i} and hyperparameters α and β.

Algorithm for Extracting Topics

For this example assume $\alpha = 25$ and $\beta = 0.01$.

$$P(z_i = j | z_{-i}, w_i, d_i, \cdot) \propto \frac{C_{w_i j}^{WT} + \beta}{\sum_{w=1}^{W} C_{wj}^{WT} + W\beta} \frac{C_{d_i j}^{DT} + \alpha}{\sum_{t=1}^{T} C_{d_i t}^{DT} + T\alpha}$$

$$P(z_i = 1 | z_{-i}, w_i, d_i, \cdot) \propto \frac{1 + 0.01}{12 + 0.05} \frac{3 + 25}{7 + 50} = 0.0412$$

$$P(z_i = 2 | z_{-i}, w_i, d_i, \cdot) \propto \frac{3 + 0.01}{11 + 0.05} \frac{4 + 25}{7 + 50} = 0.139$$

Normalize and sample a new topic for this word token.

$$P(z_i = 1 | z_{-i}, w_i, d_i, \cdot) = 0.229$$
$$P(z_i = 2 | z_{-i}, w_i, d_i, \cdot) = 0.771$$

C^{WT}	Topic 1	Topic 2
River	2	2
Stream	1	2
Bank	6	3
Money	1	3
Loan	2	1

C^{DT}	Topic 1	Topic 2
DOC1	3	4
DOC2	4	4
DOC3	5	3

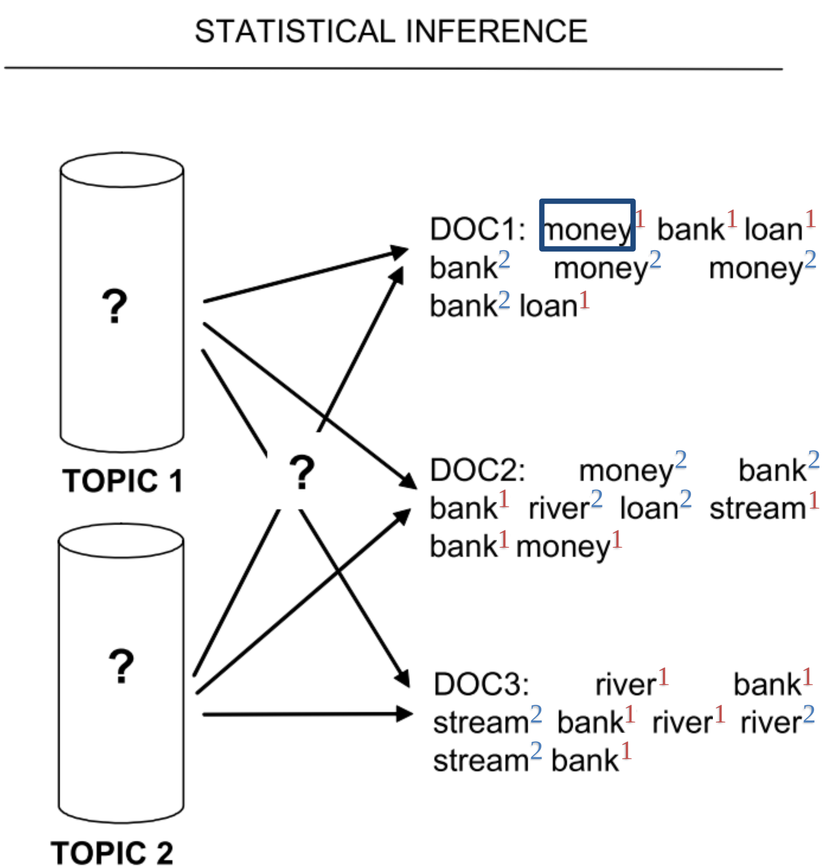
Algorithm for Extracting Topics

Gibbs Sampling Algorithm:

- Update C^{WT} and C^{DT}.
- Repeat with the next word token.
- Continue until the samples approximate the target distribution.

C^{WT}	Topic 1	Topic 2
River	2	2
Stream	1	2
Bank	6	3
Money	1	$\boxed{4}$
Loan	2	1

C^{DT}	Topic 1	Topic 2
DOC1	3	$\boxed{5}$
DOC2	4	4
DOC3	5	3



Algorithm for Extracting Topics

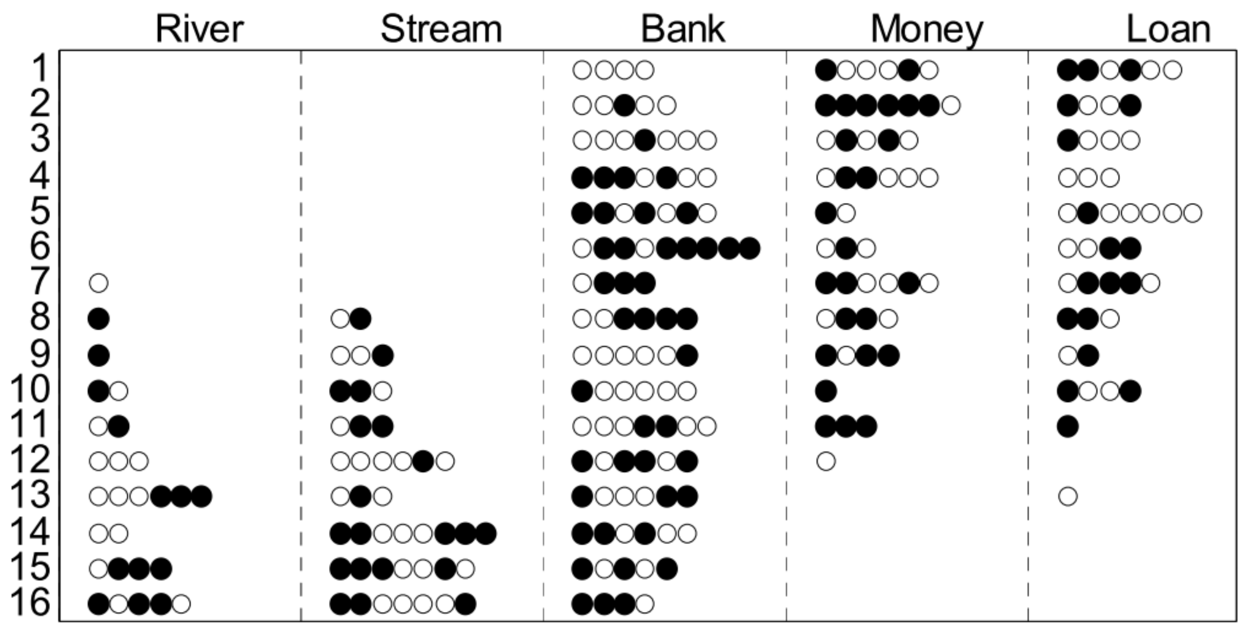
- Gibbs Sampling Algorithm:
  - The first several passes through the corpus will produce poor samples and should be ignored (burn-in period).
  - After the burn-in period, use samples at regularly spaced intervals to prevent correlations between samples.

- Estimating ϕ and θ:

$$\phi_i'^{(j)} = \frac{C_{ij}^{WT} + \beta}{\sum_{k=1}^{W} C_{kj}^{WT} + W\beta} \qquad \theta_j'^{(d)} = \frac{C_{dj}^{DT} + \alpha}{\sum_{k=1}^{T} C_{dk}^{DT} + T\alpha}$$

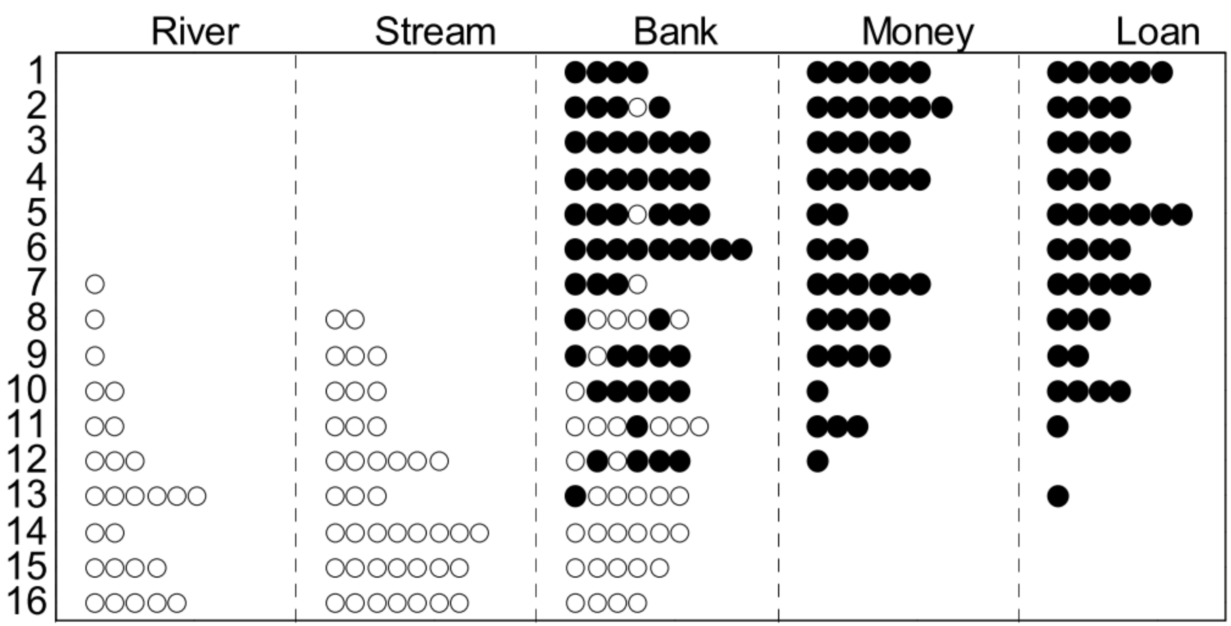
An Example

- Generate artificial data from a known topic model:
  - Topic 1 (black): $\phi^{(1)}_{MONEY} = \phi^{(1)}_{LOAN} = \phi^{(1)}_{BANK} = 1/3$
  - Topic 2 (white): $\phi^{(2)}_{RIVER} = \phi^{(2)}_{STREAM} = \phi^{(2)}_{BANK} = 1/3$



An Example

- After 64 iterations of Gibbs sampling,
  - $\phi'^{(1)}_{MONEY} = 0.32 \quad \phi'^{(1)}_{LOAN} = 0.29 \quad \phi'^{(1)}_{BANK} = 0.39$
  - $\phi'^{(2)}_{RIVER} = 0.25 \quad \phi'^{(2)}_{STREAM} = 0.4 \quad \phi'^{(2)}_{BANK} = 0.35$



Stability of Topics

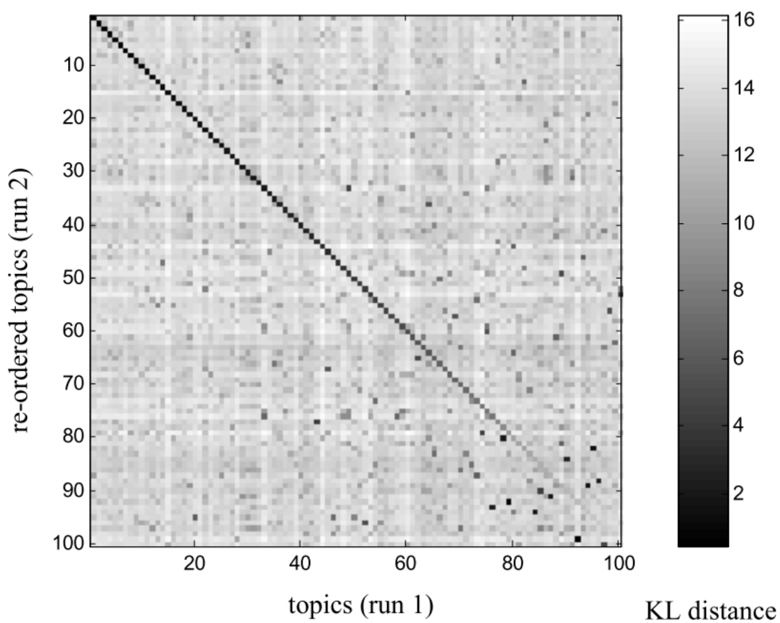
- There is no a priori ordering on the topics that will make the topics identifiable between runs of the algorithm.
- In some applications, we want to know which topics are stable (appearing across many runs of the algorithm) versus idiosyncratic for a particular run.
- We measure the distance between topics j_1 and j_2 with the symmetrized Kullback Liebler (KL) distance:

$$KL(j_1, j_2) = \frac{1}{2}\sum_{k=1}^{W} {\phi'}_k^{(j_1)} \log_2 \frac{{\phi'}_k^{(j_1)}}{{\phi''}_k^{(j_2)}} + \frac{1}{2}\sum_{k=1}^{W} {\phi''}_k^{(j_2)} \log_2 \frac{{\phi''}_k^{(j_2)}}{{\phi'}_k^{(j_1)}}$$

Stability of Topics

Alignment of topics between runs

- TASA corpus
  - W=26,414; D=37,651; N=5,628,867; T=100; α=50/T=0.5; β=0.01
  - 2000 iterations



Worst Pair of Aligned Topics
KL distance = 9.4

Run 1		Run 2	
MONEY	.094	MONEY	.086
GOLD	.044	PAY	.033
POOR	.034	BANK	.027
FOUND	.023	INCOME	.027
RICH	.021	INTEREST	.022
SILVER	.020	TAX	.021
HARD	.019	PAID	.016
DOLLARS	.018	TAXES	.016
GIVE	.016	BANKS	.015
WORTH	.016	INSURANCE	.015
BUY	.015	AMOUNT	.011
WORKED	.014	CREDIT	.010
LOST	.013	DOLLARS	.010
SOON	.013	COST	.008
PAY	.013	FUNDS	.008

Polysemy with Topics

Many words in natural language are polysemous, having multiple senses, which must be resolved through context.

Topic 77

word	prob.
MUSIC	.090
DANCE	.034
SONG	.033
PLAY	.030
SING	.026
SINGING	.026
BAND	.026
PLAYED	.023
SANG	.022
SONGS	.021
DANCING	.020
PIANO	.017
PLAYING	.016
RHYTHM	.015
ALBERT	.013
MUSICAL	.013

Topic 82

word	prob.
LITERATURE	.031
POEM	.028
POETRY	.027
POET	.020
PLAYS	.019
POEMS	.019
PLAY	.015
LITERARY	.013
WRITERS	.013
DRAMA	.012
WROTE	.012
POETS	.011
WRITER	.011
SHAKESPEARE	.010
WRITTEN	.009
STAGE	.009

Topic 166

word	prob.
PLAY	.136
BALL	.129
GAME	.065
PLAYING	.042
HIT	.032
PLAYED	.031
BASEBALL	.027
GAMES	.025
BAT	.019
RUN	.019
THROW	.016
BALLS	.015
TENNIS	.011
HOME	.010
CATCH	.010
FIELD	.010

Polysemy with Topics

Document #29795

Bix beiderbecke, at age^{060} fifteen^{207}, sat^{174} on the slope^{071} of a bluff^{055} overlooking^{027} the mississippi^{137} river^{137}. He was listening^{077} to music^{077} coming^{009} from a passing^{043} riverboat. The music^{077} had already captured^{006} his heart^{157} as well as his ear^{119}. It was jazz^{077}. Bix beiderbecke had already had music^{077} lessons^{077}. He showed^{002} promise^{134} on the piano^{077}, and his parents^{035} hoped^{268} he might consider^{118} becoming a concert^{077} pianist^{077}. But bix was interested^{268} in another kind^{050} of music^{077}. He wanted^{268} to $\boxed{\text{play}^{077}}$ the cornet. And he wanted^{268} to $\boxed{\text{play}^{077}}$ jazz^{077}...

Document #1883

There is a simple^{050} reason^{106} why there are so few periods^{078} of really great theater^{082} in our whole western^{046} world. Too many things^{300} have to come right at the very same time. The dramatists must have the right actors^{082}, the actors^{082} must have the right playhouses, the playhouses must have the right audiences^{082}. We must remember^{288} that plays^{082} exist^{143} to be performed^{077}, not merely^{050} to be read^{254}. (even when you read^{254} a $\boxed{\text{play}^{082}}$ to yourself, try^{288} to perform^{062} it, to put^{174} it on a stage^{078}, as you go along.) as soon^{028} as a $\boxed{\text{play}^{082}}$ has to be performed^{082}, then some kind^{126} of theatrical^{082}...

Document #21359

Jim^{296} has a game^{166} book^{254}. Jim^{296} reads^{254} the book^{254}. Jim^{296} sees^{081} a game^{166} for one. Jim^{296} plays^{166} the game^{166}. Jim^{296} likes^{081} the game^{166} for one. The game^{166} book^{254} helps^{081} jim^{296}. Don^{180} comes^{040} into the house^{038}. Don^{180} and jim^{296} read^{254} the game^{166} book^{254}. The boys^{020} see a game^{166} for two. The two boys^{020} $\boxed{\text{play}^{166}}$ the game^{166}. The boys^{020} $\boxed{\text{play}^{166}}$ the game^{166} for two. The boys^{020} like the game^{166}. Meg^{282} comes^{040} into the house^{282}. Meg^{282} and don^{180} and jim^{296} read^{254} the book^{254}. They see a game^{166} for three. Meg^{282} and don^{180} and jim^{296} $\boxed{\text{play}^{166}}$ the game^{166}. They $\boxed{\text{play}^{166}}$...

Similarity Between Documents

Two documents are similar to the extent that the same topics appear in both of those documents.

To compare documents d_1 and d_2, we compare their corresponding topic distributions $\theta^{(d_1)}$ and $\theta^{(d_2)}$.

The KL divergence gives the difference between distributions p and q:

$$D(p,q) = \sum_{j=1}^{T} p_j \log_2 \frac{p_j}{q_j}$$

Symmetric KL divergence:

$$KL(p,q) = \frac{1}{2}[D(p,q) + D(q,p)]$$

Symmetric JS divergence:

$$JS(p,q) = \frac{1}{2}\left[D\left(p, \frac{(p+q)}{2}\right) + D\left(q, \frac{(p+q)}{2}\right)\right]$$

Similarity Between Documents

Information Retrieval:

- Find the most similar documents to a query q.
- Retrieve the documents that maximize the conditional probability of the query given the candidate document.
- Using topic models:

$$P(q|d_i) = \prod_{w_k \in q} P(w_k|d_i)$$

$$= \prod_{w_k \in q} \sum_{j=1}^{T} P(w_k|z = j)P(z = j|d_i)$$

Similarity Between Words

Two words w_1 and w_2 are similar to the extent that they share the same topics.

We can use the symmetrized KL or JS divergence to measure the difference between $\theta^{(1)}$ and $\theta^{(2)}$, where $\theta^{(1)} = P(z|w_i = w_1)$ and $\theta^{(2)} = P(z|w_i = w_2)$

Similarity Between Words

An alternative approach is to use human word association.

Based on the topic interpretation of the observed word, predict the likelihood of new words in the same context.

$$P(w_2|w_1) = \sum_{j=1}^{T} P(w_2|z=j)P(z=j|w_1)$$

HUMANS

FUN	.141
BALL	.134
GAME	.074
WORK	.067
GROUND	.060
MATE	.027
CHILD	.020
ENJOY	.020
WIN	.020
ACTOR	.013
FIGHT	.013
HORSE	.013
KID	.013
MUSIC	.013

TOPICS

BALL	.036
GAME	.024
CHILDREN	.016
TEAM	.011
WANT	.010
MUSIC	.010
SHOW	.009
HIT	.009
CHILD	.008
BASEBALL	.008
GAMES	.007
FUN	.007
STAGE	.007
FIELD	.006

Observed and predicted responses for the cue word PLAY.

Conclusion

Generative models for text, such as the topic model, provide a deeper understanding of human language.

Statistical analysis of large document collections can identify the latent structure of text and capture more of the language content.

Topic models can be extended to identify some interesting properties of language, such as the hierarchical semantic relations between words and the interaction between syntax and semantics.