# Probabilistic Topic Models

Reza Soltani, Helen Zhang

# Structure

- Latent Semantic Analysis (LSA)
- Topic Models
- Latent Dirichlet Allocation (LDA)
- Algorithm for Extracting Topics (Gibbs sampling)
- Polysemy with Topics
- Computing Similarities between Documents or between Words
- Canvas Questions Discussion

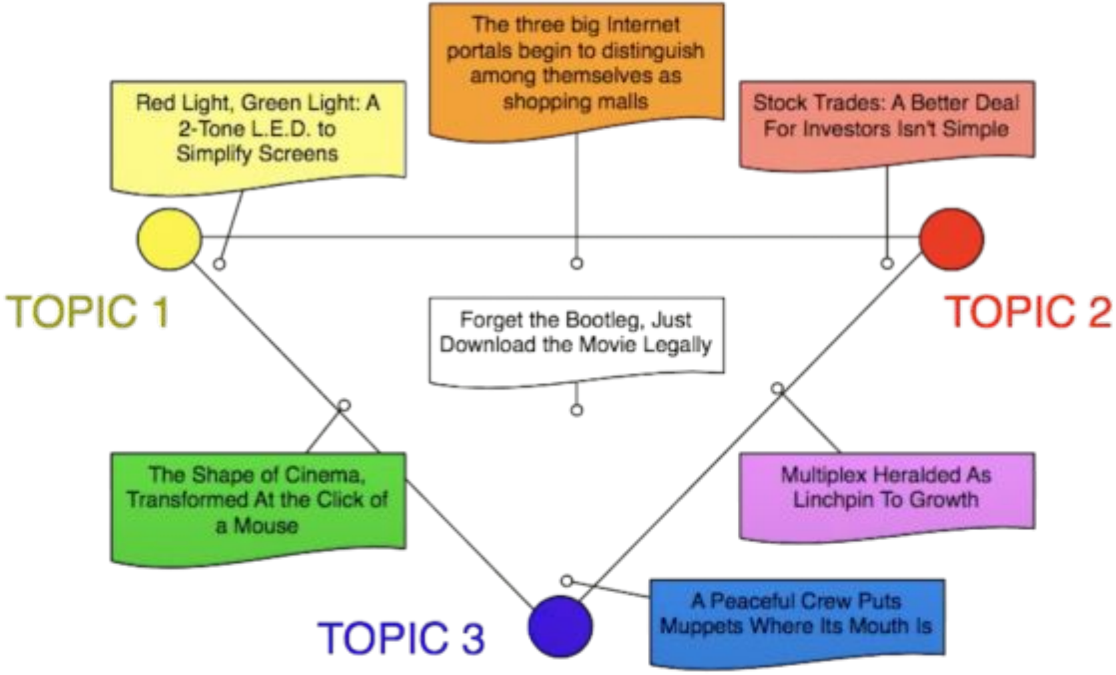
# Latent Semantic Analysis (LSA)

Statistical method that can be applied to large databases and yield insight into words and documents.

1. Semantic information can be derived from a **word-document co-occurrence** matrix.
2. Dimensionality reduction is an essential part of this derivation.
3. Words and documents can be represented as points in Euclidean space.

# Topic Models

- Documents are the mixtures of different topics.



# Topic Models

- A topic is a probability distribution over words
- Each topic is individually interpretable

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 |

**Figure 1.** An illustration of four (out of 300) topics extracted from the TASA corpus.

# Generative Model

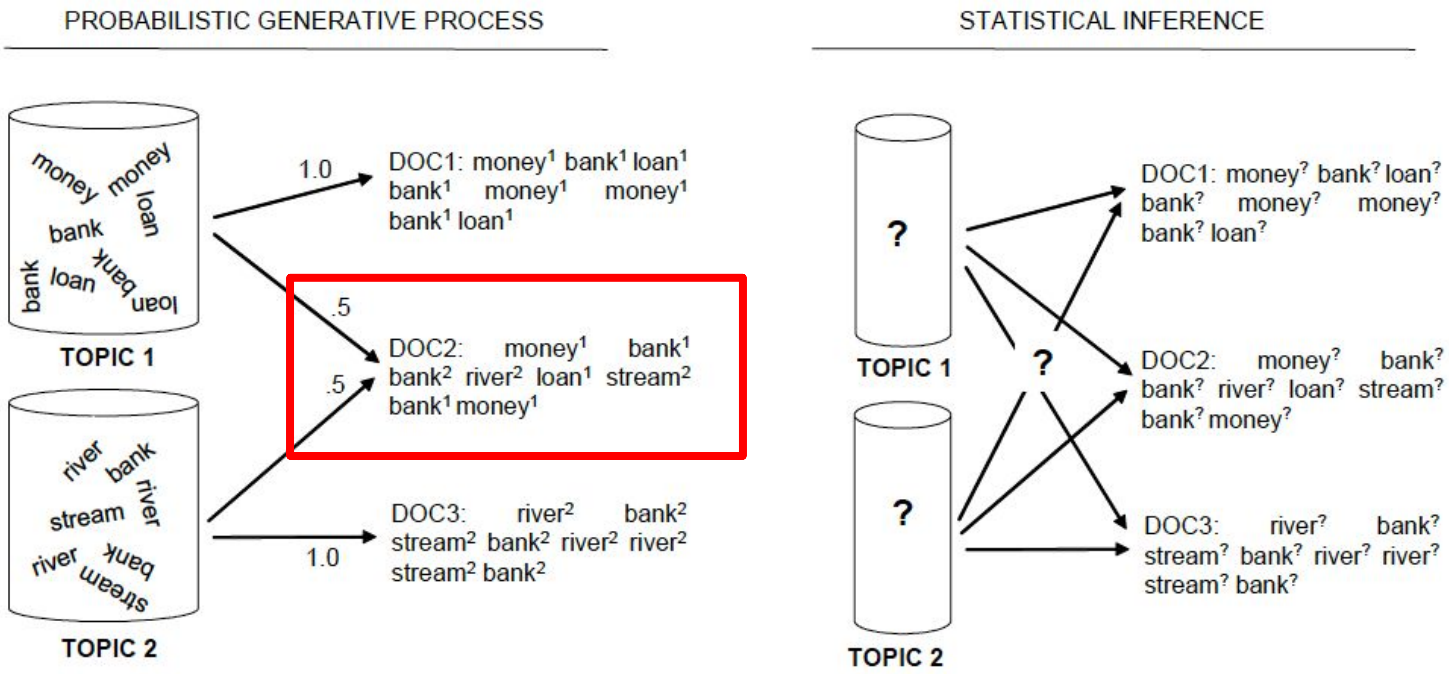



**Figure 2.** Illustration of the generative process and the problem of statistical inference underlying topic models

# Words Order

- Integrating Topics and Syntax (Griffiths, Steyvers, Blei, and Tenenbaum 2005)
  - Combining syntactic and semantic generative models

- Topic Segmentation with An Ordering-Based Topic Model (Lan Du, John K Pate and Mark Johnson 2019)

# Probabilistic Topic Models

- $\theta^{(d)} = P(z)$ refer to the **multinomial** distribution over topics for document d
- $\phi^{(j)} = P(w \mid z{=}j)$ refer to the **multinomial** distribution over words given topic j
- Distribution over words within a document: $P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j) P(z_i = j)$
- T is the number of topics

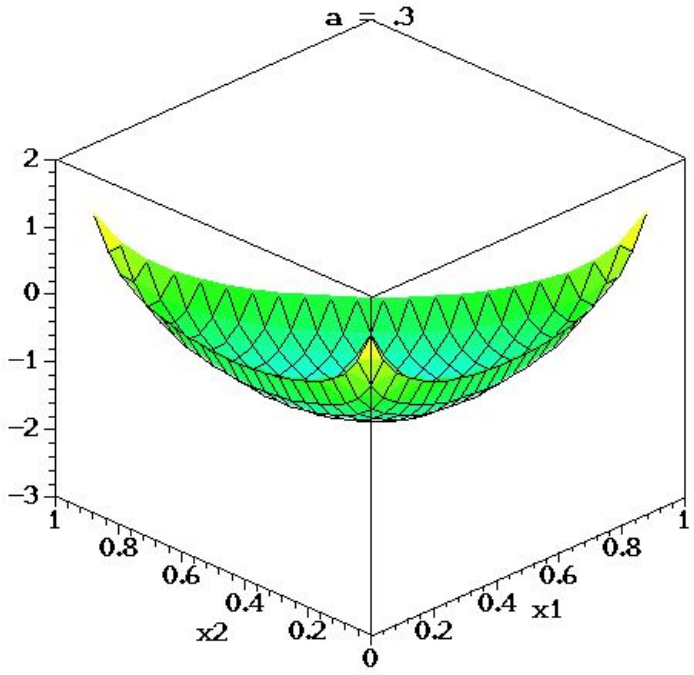
# Dirichlet Distribution

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

- $\Gamma$ is the extension of the factorial function to complex numbers.
- Dirichlet distribution is the **conjugate prior** of the multinomial distribution.
- Smoothing by symmetric Dirichlet distribution with a single hyperparameter $\alpha$

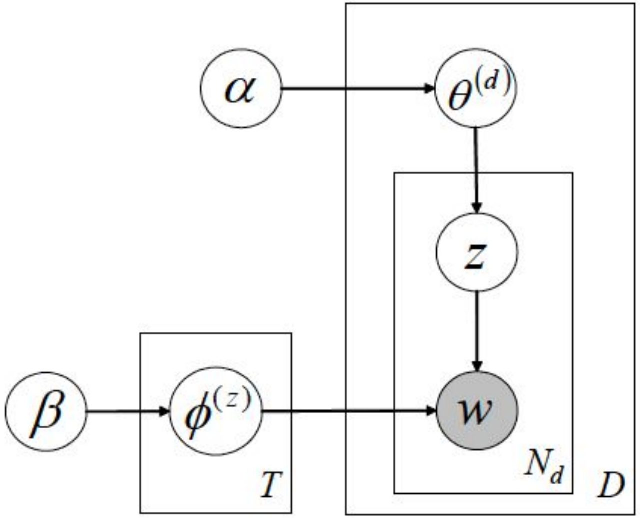
# How Dirichlet Distribution Helps?

- In practice, $\alpha < 1$ is used.
- Pressure to pick topic distributions favoring just a few topics.
- And each topic favoring a few words.



# LDA: Graphical model

- w in the only observed variable
- variable in the lower right corner referring to the number of samples



# LDA: Geometric Interpretation

W = number of distinct words in vocabulary

T = number of topics

Any distribution over words can be represented as a point in the W-1 dimensional simplex (generalization of triangle).

Topics and documents can be represented over this simplex.

Topics spans a low-dimensional subsimplex and the projection of documents onto this subsimplex is a reduction in dimensionality.

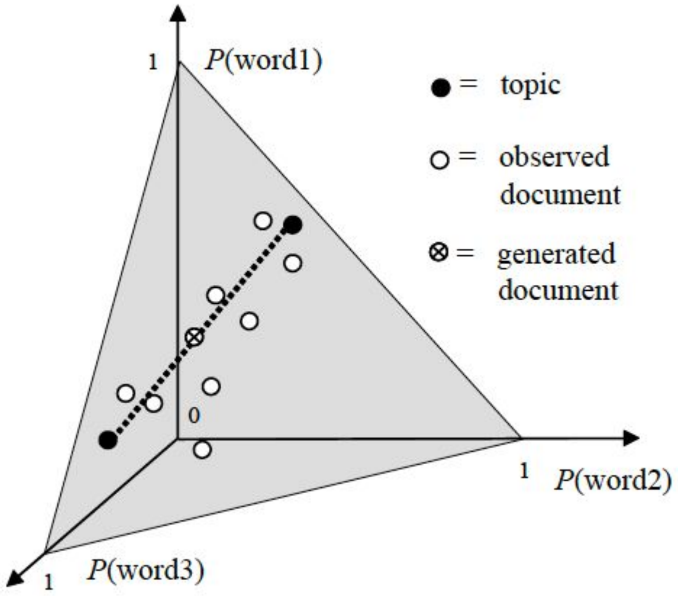
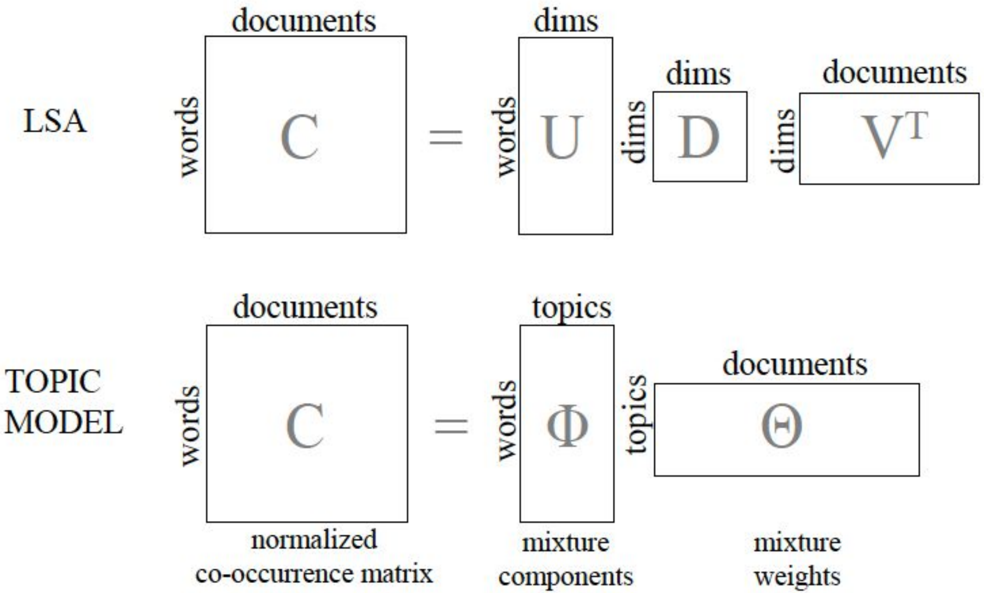



**Figure 5**. A geometric interpretation of the topic model.

# Matrix Factorization

- LSA and topic models both of find a low-dimensional representation for the content of a set of documents.
- Matrix D can be absorbed in V or U to make the similarity more clear.
- In topic model, feature values are non-negative and sum up to one.

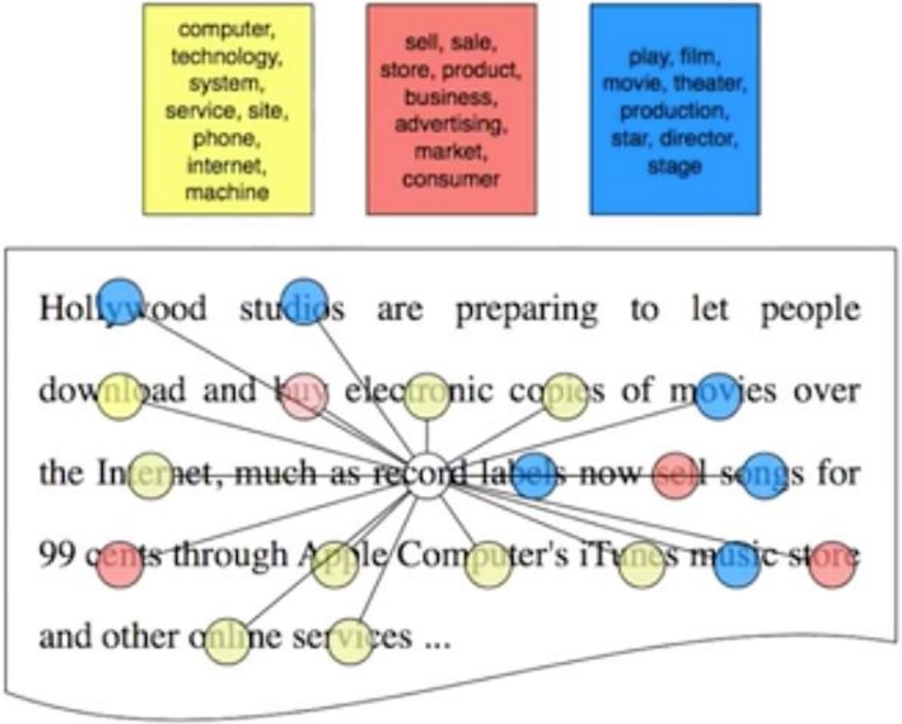


# Algorithm for Extracting Topics

- Directly estimating the topic-word distributions φ and the topic distributions θ
  - Expectation-maximization (Hofmann, 1999) : suffers from local maxima of the likelihood function
- Estimate the posterior distribution over z, the assignment of word tokens to topics, given the observed words w
  - Text collections contain millions of word token, the estimation of the posterior over z requires efficient estimation procedures.
  - Gibbs sampling:
    - Easy to implement, relatively efficient for extracting a set of topics from a large corpus
    - Simulates a high-dimensional distribution by sampling on lower-dimensional subsets of variables

# Gibbs Sampling Intuition

- Considers each word token in the text collection in turn
- Estimates the probability of assigning the current word token to each topic, conditioned on the topic assignments to all other word tokens.
- Initialize randomly
- Sample sequentially until the values approximate target distribution



# Gibbs Sampling Equation

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

- Words are assigned to topics depending on <u>how likely the word is for a topic</u>, as well as <u>how dominant a topic is in a document</u>
- $C^{WT}_{wj}$ number of times word w is assigned to topic j
- $C^{DT}_{dj}$ number of times topic j is assigned to some word token in document d
- α, β hyperparameters, smoothing
- Estimating φ and θ:

$$\phi'^{(j)}_i = \frac{C^{WT}_{ij} + \beta}{\sum_{k=1}^{W} C^{WT}_{kj} + W\beta} \qquad \theta'^{(d)}_j = \frac{C^{DT}_{dj} + \alpha}{\sum_{k=1}^{T} C^{DT}_{dk} + T\alpha}$$

# Example

Generate artificial data from a known topic model:

$$\phi^{(1)}_{MONEY} = \phi^{(1)}_{LOAN} = \phi^{(1)}_{BANK} = 1/3$$

$$\phi^{(2)}_{RIVER} = \phi^{(2)}_{STREAM} = \phi^{(2)}_{BANK} = 1/3$$

Randomly assign topics at the start, perform gibbs sampling after 64 internations:

$$\phi'^{(1)}_{MONEY} = .32, \quad \phi'^{(1)}_{LOAN} = .29, \quad \phi'^{(1)}_{BANK} = .39$$

$$\phi'^{(2)}_{RIVER} = .25, \quad \phi'^{(2)}_{STREAM} = .4, \quad \phi'^{(2)}_{BANK} = .35$$

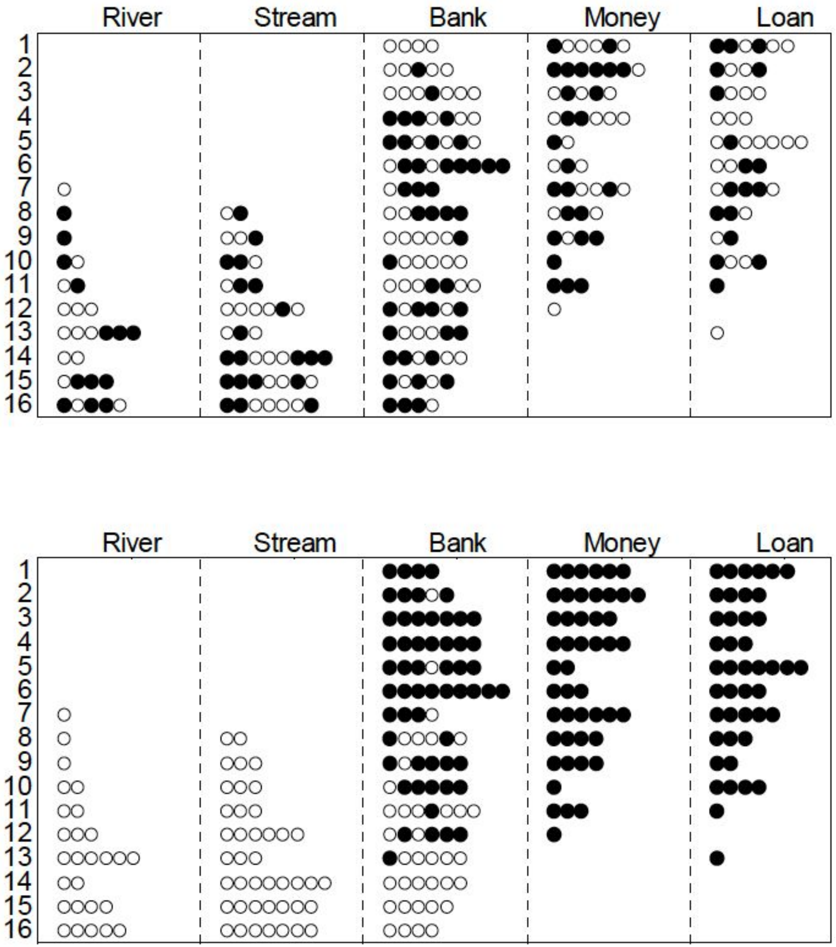


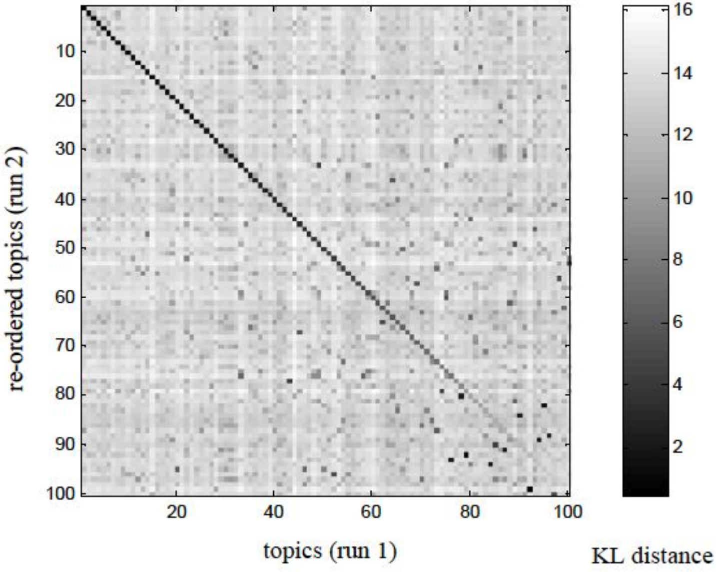

**Figure 7**. An example of the Gibbs sampling procedure.

# Exchangeability of topics

- There is no a priori ordering on the topics that will make the topics identifiable between or even within runs of the algorithm. Therefore, the different samples *cannot* be averaged at the level of topics.
- When topics are used to calculate a statistic which is invariant to the ordering of the topics, it is important to average over different Gibbs samples to improve results
- Model averaging is likely to improve results because it allows sampling from multiple local modes of the posterior.

# Stability

- The solutions from different samples will give different results but that many topics are stable across runs.




| 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 |

**Figure 8**. Stability of topics between different runs.

# Number of Topics

A solution with too few topics will generally result in very broad topics whereas a solution with too many topics will result in uninterpretable topics

- Bayesian model selection:
  - Estimate the posterior probability of the model while integrating over all possible parameter settings (i.e., all ways to assign words to topics)
  - Choose the number of topic that leads to the highest posterior probability.
- Best generalization performance
  - A topic model estimated on a subset of documents should be able to predict word choice in the remaining set of documents
- Non-parametric Bayesian statistics
  - Automatically select number of topics

# Polysemy with Topics

- Probabilistic topic models represent semantic ambiguity through uncertainty over topics

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 |

- Iterative sampling: the assignment of each word token to a topic depends on the assignments of the other words in the context.

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 play[077] the cornet. And he wanted[268] to 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 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 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] play[166] the game[166]. The boys[020] 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] play[166] the game[166]. They play[166]...

# Similarities between documents

- The similarity between documents d1 and d2 can be measured by the similarity between their corresponding topic distributions
- Distribution similarity function: Kullback Leibler (KL) divergence

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

- Equal to zero when for all j, $p_j = q_j$
- Symmetric measure based on KL divergence:

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

  - Jensen-Shannon (JS) divergence:

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

# Similarities between documents

- Find similar documents to the given document (information retrieval application)
  - Assess the similarity between the topic distributions
  - Model information retrieval as a probabilistic query to the topic model

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

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

- Important to obtain stable estimates for the topic distributions
  - Average the similarity function over multiple Gibbs samples

# Similarity between words

- Measured by the extent that two words share the same topics
  - Similarity between conditional topic distributions for two words w1 and w2

$$\theta^{(1)} = P\left(z \mid w_i = w_1\right) \text{ and } \theta^{(2)} = P\left(z \mid w_i = w_2\right)$$

  - Measured by symmetrized KL or JS distance
- Associative relations between words

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

| HUMANS | | TOPICS | |
|---|---|---|---|
| FUN | .141 | BALL | .036 |
| BALL | .134 | GAME | .024 |
| GAME | .074 | CHILDREN | .016 |
| WORK | .067 | TEAM | .011 |
| GROUND | .060 | WANT | .010 |
| MATE | .027 | MUSIC | .010 |
| CHILD | .020 | SHOW | .009 |
| ENJOY | .020 | HIT | .009 |
| WIN | .020 | CHILD | .008 |
| ACTOR | .013 | BASEBALL | .008 |
| FIGHT | .013 | GAMES | .007 |
| HORSE | .013 | FUN | .007 |
| KID | .013 | STAGE | .007 |
| MUSIC | .013 | FIELD | .006 |

**Figure 9**. Observed and predicted response distributions for the word PLAY.

The balance between the influence of word frequency and semantic relatedness found by the topic model can result in better performance than LSA on this task.

# Canvas Questions...

- Why are the Gibbs samples poor estimates of the posterior during the initial stage of the sampling process (burn-in period)?
    - Random initialization
    - Ignore samples at the beginning, keeping every kth sample, averaging ...
- How to determine how many iterations you would run for the Gibbs Sampling algorithm? Efficiency (time consumption) of Gibbs Sampling?
    - Guaranteed to converge. A good initialization might help?

# Canvas Questions...

- Downside of Gibbs Sampling?

1. Long convergence time especially with the dimensionality of the data growing.
2. Convergence time also depends on the shape of the distribution. When there are islands of high-probability states with no paths between them, Gibbs sampling will become trapped in one of the two high-probability vectors, and will never reach the other one.

# Canvas Questions...

- What is the difference between exchangeability and stability of topics?

  Exchangeability: Does topics have same ordering, between or within runs

  Stability: Does same topics reappear across different runs
- Some of the topics are unstable across runs, what is the reason behind it?

  Sampling

- Why is it more important to average over different Gibbs samples when topics are used to calculate a statistic which is invariant to the ordering of the topics?

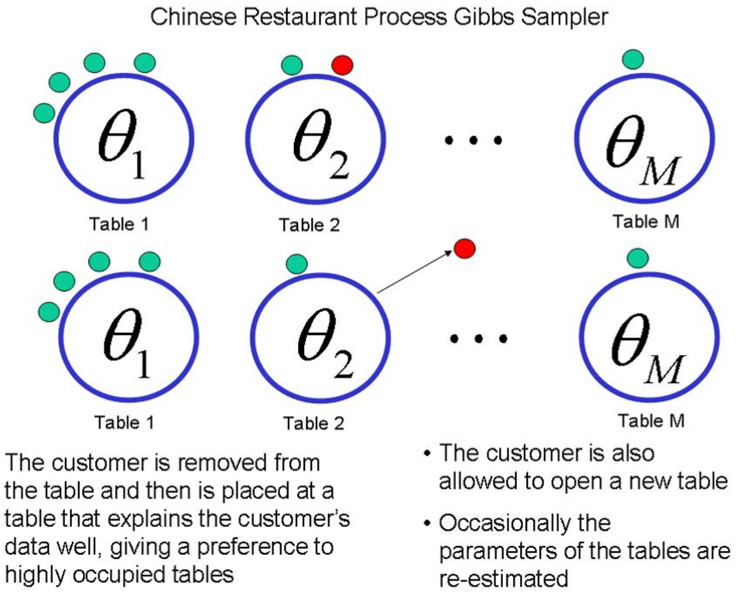
  Allows sampling from multiple local modes of the posterior.

# Canvas Questions...

- Considering that automatic mechanisms do not meet the users' needs in advance. What strategies can be helpful in those cases? Is hierarchical topic modeling a good approach?
- Is there any preferred/best method for determining the Number of Topics?

Bayesian Nonparametrics

- Chinese Restaurant Process (Dirichlet Process)



# Canvas Questions...

- KL vs JS: Is one approach better than the other? Can anything be said about the assumptions/performance of these approaches? What other topic similarity metrics exist?

  - KL is not symmetric, which can be a feature in some applications
  - It is also possible to consider the topic distributions as vectors and apply geometrically motivated functions such as Euclidean distance, dot product or cosine.

# Canvas Questions...

- How are topic models evaluated?
  - Likelihood of held-out data
    - Since they are probabilistic models, likelihood of a new document can be calculated.
  - Word intrusion
    - Insert one random word inside and ask humans to identify the random word.

# Canvas Questions...

- What are some algorithms for extracting topics not mentioned in the paper? Is the state-of-the-art any of the mentioned approaches?
  - Latent Dirichlet Allocation
  - Gibbs Sampling
  - Variational Inference: Minimizes KL(q||p) where q is a simpler graphical model than the original p
  - Structural Topic Model (STM) : Incorporates metadata into the model and uncover how different documents might talk about the same underlying topic using different word choices.