Introduction to Probabilistic Topic Models

Ko, Youngjoong

Computer Engineering
Dong-A University

Contents

Introduction

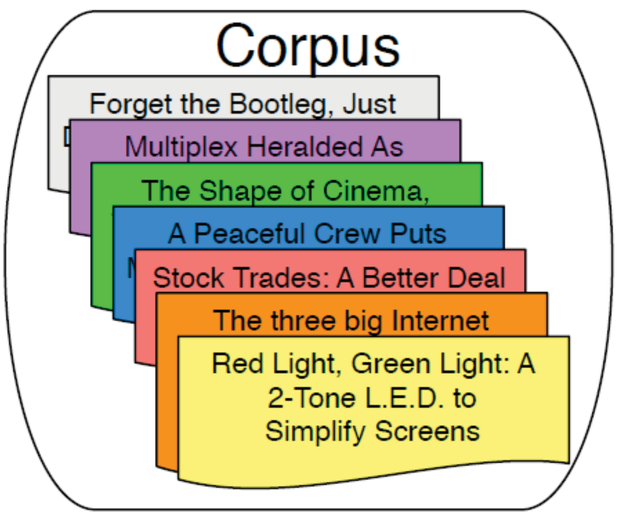
❖ Why topic model?

- Suppose you have a huge number of documents
- You want to know what's going on
- Don't have time to read them (e.g. every New York Times article from the 50's)
- Topic models offer a way to get a corpus-level view of major themes
- Unsupervised

Introduction

❖ Conceptual Approach

From an **input corpus** → words to topics



Introduction

❖ Conceptual Approach

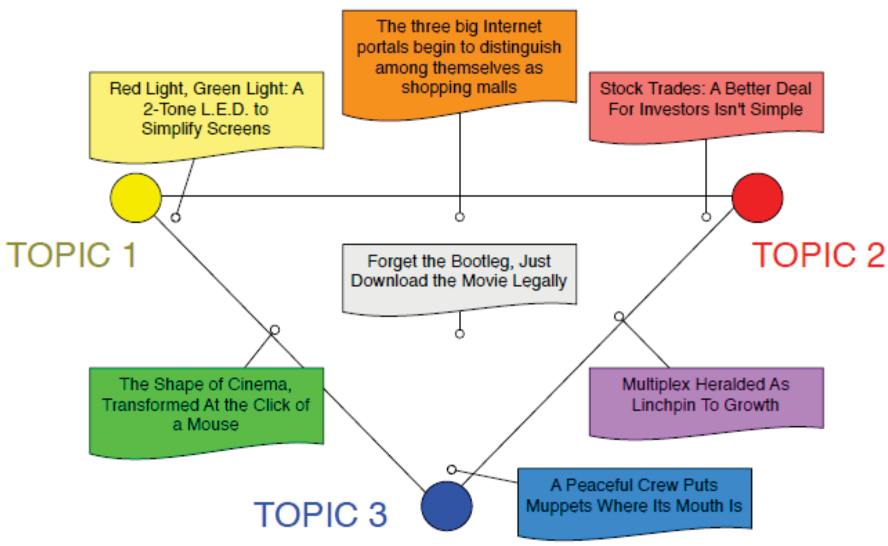
From an input corpus → **words to topics**

TOPIC 1	TOPIC 2	TOPIC 3
computer, technology, system, service, site, phone, internet, machine	sell, sale, store, product, business, advertising, market, consumer	play, film, movie, theater, production, star, director, stage

Introduction

❖ Conceptual Approach

- For each document, what topics are expressed by that document?



Introduction

❖ Topics from Science

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Introduction

❖ Latent Semantic Analysis (LSA)

- People illustrate that applying a statistical method such as LSA to large databases can yield insight into human cognition
- Three claims
  - semantic information can be derived from a word-document co-occurrence matrix
  - dimensionality reduction is an essential part of this derivation
  - words and documents can be represented as points in Euclidean space
- A different approach
  - that is consistent with the first two of these claims,
  - but differs in the third, describing a class of statistical models in which the semantic properties of words and documents are expressed in terms of probabilistic topics.

Introduction

❖ Topic models

- Based upon the idea that documents are mixtures of topics, where a topic is a probability distribution over words.
- A topic model is a generative model for documents
  - It specifies a simple probabilistic procedure by which documents can be generated
  - To make a new document, one chooses a distribution over topics.
  - Then, for each word in that document, one chooses a topic at random according to this distribution, and draws a word from that topic.
- Standard statistical techniques can be used to invert this process
  - Four example topics from the TASA corpus, a collection of over 37,000 text passages
  - The sixteen words that have the highest probability under each topic.
  - Four topics relate to drug use, colors, memory and the mind, and doctor visits.
  - Documents with different content can be generated by choosing different distributions over topics.
  - Ex)
    - ✓ by giving equal probability to the first two topics, one could construct a document about a person that has taken too many drugs, and how that affected color perception.

Introduction

❖ An illustration of four (out of 300) topics

Topic 247 word	prob.	Topic 5 word	prob.	Topic 43 word	prob.	Topic 56 word	prob.
DRUGS	.069	RED	.202	MIND	.081	DOCTOR	.074
DRUG	.060	BLUE	.099	THOUGHT	.066	DR.	.063
MEDICINE	.027	GREEN	.096	REMEMBER	.064	PATIENT	.061
EFFECTS	.026	YELLOW	.073	MEMORY	.037	HOSPITAL	.049
BODY	.023	WHITE	.048	THINKING	.030	CARE	.046
MEDICINES	.019	COLOR	.048	PROFESSOR	.028	MEDICAL	.042
PAIN	.016	BRIGHT	.030	FELT	.025	NURSE	.031
PERSON	.016	COLORS	.029	REMEMBERED	.022	PATIENTS	.029
MARIJUANA	.014	ORANGE	.027	THOUGHTS	.020	DOCTORS	.028
LABEL	.012	BROWN	.027	FORGOTTEN	.020	HEALTH	.025
ALCOHOL	.012	PINK	.017	MOMENT	.020	MEDICINE	.017
DANGEROUS	.011	LOOK	.017	THINK	.019	NURSING	.017
ABUSE	.009	BLACK	.016	THING	.016	DENTAL	.015
EFFECT	.009	PURPLE	.015	WONDER	.014	NURSES	.013
KNOWN	.008	CROSS	.011	FORGET	.012	PHYSICIAN	.012
PILLS	.008	COLORED	.009	RECALL	.012	HOSPITALS	.011

- Representing the content of words and documents with probabilistic topics has one distinct advantage over spatial representation
  - Each topic is individually interpretable,
    - ✓ providing a probability distribution over words that picks out a coherent cluster of correlated terms.
    - ✓ the topics are typically as interpretable as the ones shown here.
  - This contrasts with the arbitrary axes of a spatial representation, and can be extremely useful in many applications

Introduction

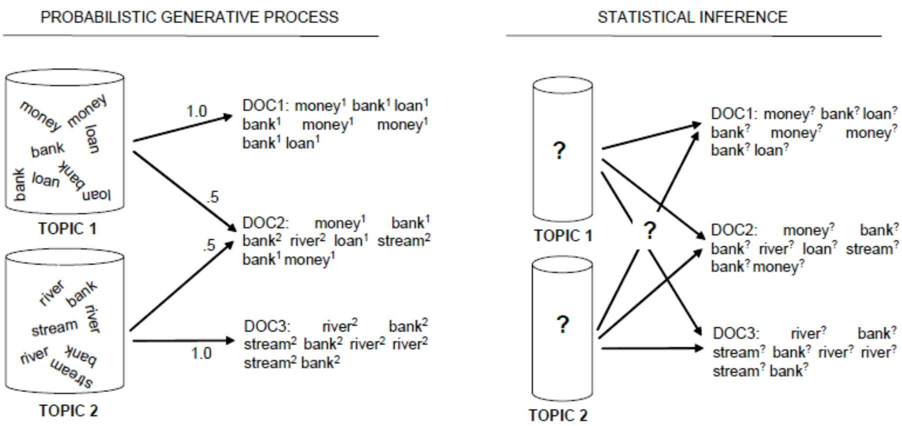
❖ Why should you care?

- Neat way to explore/understand corpus collections
- NLP Applications
  - POS Tagging [Toutanova and Johnson 2008]
  - Word Sense Disambiguation [Boyd-Graber et al. 2007]
  - Word Sense Induction [Brody and Lapata 2009]
  - Discourse Segmentation [Purver et al. 2006]
- Psychology [Griths et al.2007b]: word meaning, polysemy
- Inference is (relatively) simple

Generative Models

❖ A Generative Model for Documents

- Based on simple probabilistic sampling rules that describe how words in documents might be generated on the basis of latent (random) variables
  - The goal is to find the best set of latent variables that can explain the observed data (i.e., observed words in documents), assuming that the model actually generated the data.
  - Ex) Illustration of the topic modeling approach in two distinct ways:



Generative Models

❖ A Generative Model for Documents

- ➢ Ex) Illustration of the topic modeling approach in two distinct ways:
  - In the generative model (Left panel)
    - ✓ With two topics (money and rivers)
    - ✓ Bags containing different distributions over words
    - ✓ Different documents can be produced by picking words from a topic depending on the weight given to the topic.
    - ✓ topic models to capture polysemy (eg. Bank)
    - ✓ there is no notion of mutual exclusivity that restricts words to be part of one topic only.
    - ✓ Bag-of-words assumption : common to many statistical models of language with LSA
  - In the statistical inference (Right Panel)
    - ✓ Given the observed words in a set of documents, we would like to know what topic model is most likely to have generated the data.
    - ✓ This involves inferring the probability distribution over words associated with each topic, the distribution over topics for each document, and, often, the topic responsible for generating each word

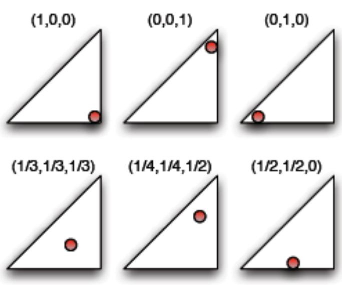
Generative Models

❖ A Generative Model for Documents

- ➢ How your data came to be
- ➢ Sequence of Probabilistic Steps
- ➢ Posterior Inference

❖ Multinomial Distribution

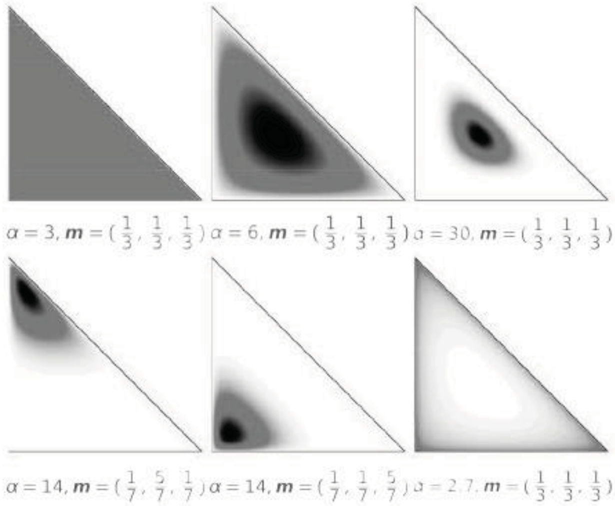
- ➢ Distribution over discrete outcomes
- ➢ Represented by non-negative vector that sums to one
- ➢ Picture representation
- ➢ Come from a Dirichlet distribution



Generative Models

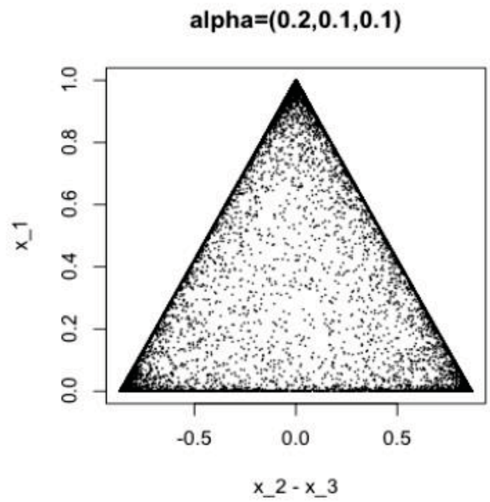
❖ Dirichlet Distribution

$$P(\boldsymbol{p} \mid \alpha \boldsymbol{m}) = \frac{\Gamma(\sum_k \alpha m_k)}{\prod_k \Gamma(\alpha m_k)} \prod_k p_k^{\alpha m_k - 1}$$



Generative Models

❖ Dirichlet Distribution



Generative Models

❖ Dirichlet Distribution

- If $\phi \sim \mathrm{Dir}(\alpha)$, $\mathbf{w} \sim \mathrm{Mult}(\phi)$, and $n_k = |\{w_i : w_i = k\}|$ then

$$p(\phi|\alpha, \mathbf{w}) \propto p(\mathbf{w}|\phi)p(\phi|\alpha) \quad (1)$$
$$\propto \prod_k \phi^{n_k} \prod_k \phi^{\alpha_k - 1} \quad (2)$$
$$\propto \prod_k \phi^{\alpha_k + n_k - 1} \quad (3)$$

- Conjugacy: this **posterior** has the same form as the **prior**

Probabilistic Topic Models

❖ The Fundamental Idea

- A document is a mixture of topics
- To introduce notation
  - P(z) : the distribution over topics *z* in a particular document
  - P(w | z) : the probability distribution over words *w* given topic *z*
- Generative Process
  - Each word w_i in a document (where the index refers to the *i*-th word token) is generated by first sampling a topic from the topic distribution,
  - then choosing a word from the topic-word distribution

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

    - ✓ $P(z_i = j)$: the probability that the *j*-th topic was sampled for the *i*-th word token
    - ✓ $P(w_i \mid z_i = j)$ as the probability of word w_i under topic *j*.
    - ✓ T : the number of topics.
  - To simplify notation,
    - ✓ φ(j) = P(w | z=j) : the multinomial distribution over words for topic *j*
    - ✓ θ(d) = P(z) : the multinomial distribution over topics for document *d*

Probabilistic Topic Models

❖ The Fundamental Idea

➢ Generative Process

- the text collection : D documents
- each document *d* : N_d word tokens
- N : the total number of word tokens (i.e., $N = \Sigma N_d$).
- The parameters φ and θ indicate which words are important for which topic and which topics are important for a particular document, respectively.

➢ From probabilistic Latent Semantic Indexing method (pLSI)

- Hofmann (1999; 2001) introduced the probabilistic topic approach to document modeling in his Probabilistic Latent Semantic Indexing method
- The pLSI model does not make any assumptions about how the mixture weights θ are generated, making it difficult to test the generalizability of the model to new documents.
- Blei et al. (2003) extended this model by introducing a Dirichlet prior on θ, calling the resulting generative model Latent Dirichlet Allocation (LDA)
- As a conjugate prior for the multinomial, the Dirichlet distribution is a convenient choice as prior, simplifying the problem of statistical inference

Probabilistic Topic Models

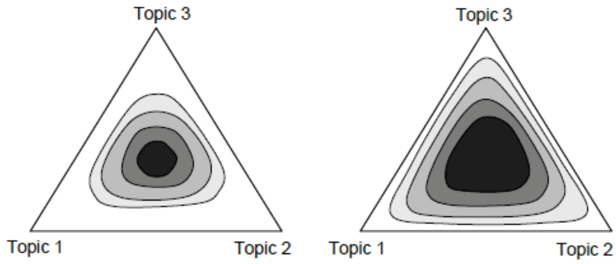
➢ The probability density of a *T* dimensional Dirichlet distribution over the multinomial distribution p=(p1, …, pT)

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

- parameters : α1 … αT
- each hyperparameter α_j can be interpreted as a prior observation count for the number of times topic *j* is sampled in a document, before having observed any actual words from that document.

Probabilistic Topic Models

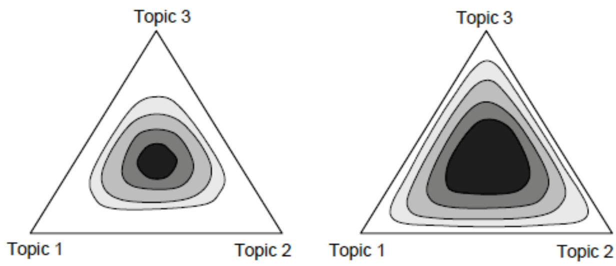
- Dirichlet distribution over the multinomial distribution p=(p1, …, pT)
  - Ex) Dirichlet distribution for three topics in a two-dimensional simplex (Left: α = 4. Right: α = 2)
    - ✓ a smoothed topic distribution, with the amount of smoothing determined by the α parameter



  - Dirichlet prior on the topic distributions can be interpreted as forces on the topic combinations with higher α moving the topics away from the corners of the simplex, leading to more smoothing (compare the left and right panel).
  - For α < 1, the modes of the Dirichlet distribution are located at the corners of the simplex.
  - In this regime (often used in practice), there is a bias towards sparsity, and the pressure is to pick topic distributions favoring just a few topics.

Probabilistic Topic Models

- Dirichlet distribution over the multinomial distribution p=(p1, …, pT)
  - Ex) Dirichlet distribution for three topics in a two-dimensional simplex (Left: α = 4. Right: α = 2)
    - ✓ a smoothed topic distribution, with the amount of smoothing determined by the α parameter



  - Dirichlet prior on the topic distributions can be interpreted as forces on the topic combinations with higher α moving the topics away from the corners of the simplex, leading to more smoothing (compare the left and right panel).
  - For α < 1, the modes of the Dirichlet distribution are located at the corners of the simplex.
  - In this regime (often used in practice), there is a bias towards sparsity, and the pressure is to pick topic distributions favoring just a few topics.

Probabilistic Topic Models

➢ The hyperparameter β

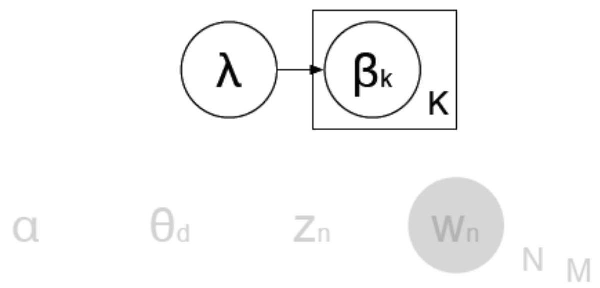
- placing a symmetric Dirichlet(β) prior on φ
- interpreted as the prior observation count on the number of times words are sampled from a topic before any word from the corpus is observed
- This smoothes the word distribution in every topic, with the amount of smoothing determined by β

➢ Good choices for the hyperparameters α and β

- depend on number of topics and vocabulary size.
- From previous research, we have found α =50/T and β = 0.01 to work well with many different text collections.

Probabilistic Topic Models

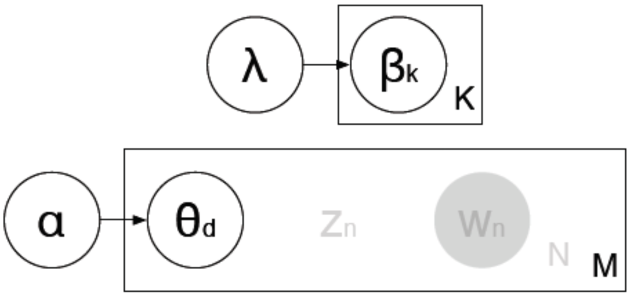
❖ Graphical Model for Generative Model Approach



- For each topic $k \in \{1, \ldots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ

Probabilistic Topic Models

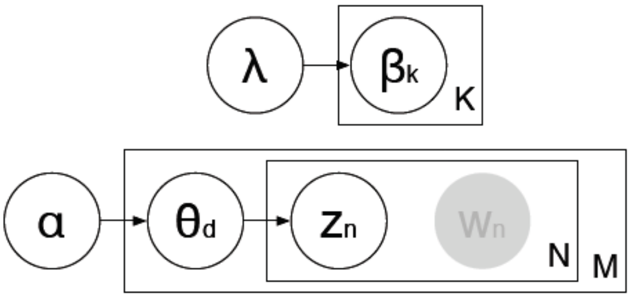
❖ Graphical Model for Generative Model Approach



- For each topic $k \in \{1, \ldots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ
- For each document $d \in \{1, \ldots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α

Probabilistic Topic Models

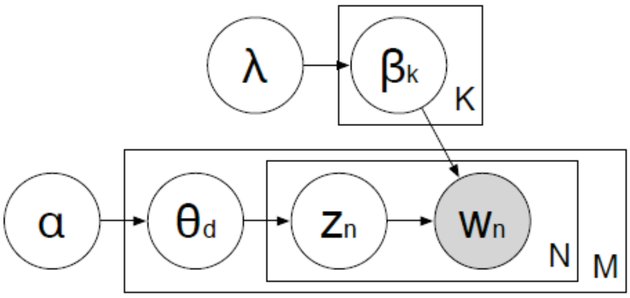
❖ Graphical Model for Generative Model Approach



- For each topic $k \in \{1, \ldots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ
- For each document $d \in \{1, \ldots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α
- For each word position $n \in \{1, \ldots, N\}$, select a hidden topic z_n from the multinomial distribution parameterized by θ.

Probabilistic Topic Models

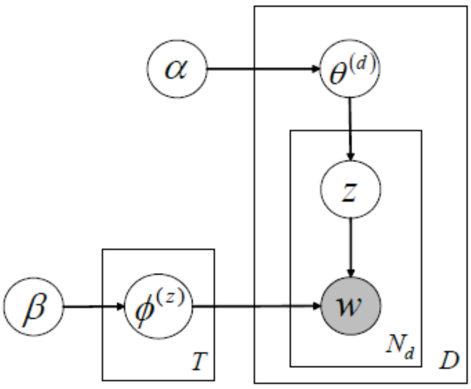
❖ Graphical Model for Generative Model Approach



- For each topic $k \in \{1, \dots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ
- For each document $d \in \{1, \dots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α
- For each word position $n \in \{1, \dots, N\}$, select a hidden topic z_n from the multinomial distribution parameterized by θ.
- Choose the observed word w_n from the distribution β_{z_n}.

Probabilistic Topic Models

❖ Graphical Model



- the inner plate over z and w illustrates the repeated sampling of topics and words until N_d words have been generated for document d.
- The plate surrounding θ(d) illustrates the sampling of a distribution over topics for each document d for a total of D documents.
- The plate surrounding φ(z) illustrates the repeated sampling of word distributions for each topic z until T topics have been generated

Probabilistic Topic Models

❖ Graphical Model

➢ Plate Notation

- Probabilistic generative models with repeated sampling steps can be conveniently illustrated
- shaded variables : observed variable, unshaded variables : latent (i.e., unobserved) variables
- The variables φ and θ, as well as z (the assignment of word tokens to topics) are the three sets of latent variables that we would like to infer.
- To treat the hyperparameters α and β as constants in the model

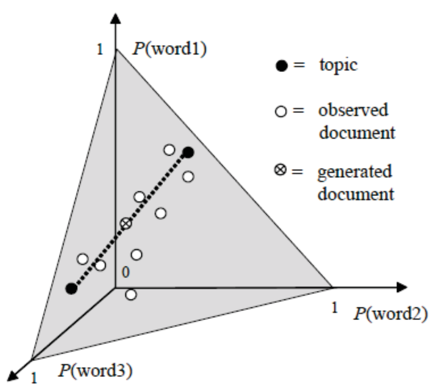
- Arrows indicate conditional dependencies between variables
- plates (the boxes in the figure) refer to repetitions of sampling steps with the variable in the lower right corner referring to the number of samples.

Probabilistic Topic Models

❖ Geometric Interpretation

➢ The probabilistic topic model has an elegant geometric interpretation

- With a vocabulary of W distinct words, a W dimensional space can be constructed where each axis represents the probability of observing a particular word type.
- The W-1 dimensional simplex represents all probability distributions over words
- the shaded region is the two-dimensional simplex that represents all probability distributions over three words
- the topics span a low-dimensional subsimplex and the projection of each document onto the low-dimensional subsimplex can be thought of as dimensionality reduction.

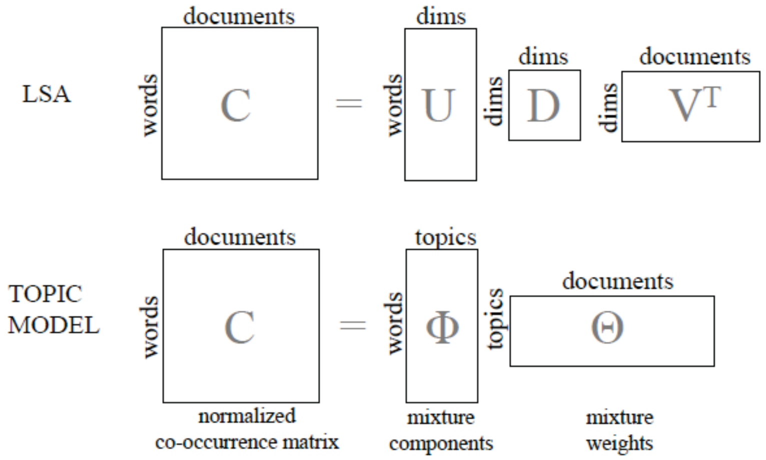


Probabilistic Topic Models

❖ Matrix Factorization Interpretation

➢ LSA vs. Topic Modeling

- In LSA, a word document co-occurrence matrix can be decomposed into three matrices: a matrix of word vectors, a diagonal matrix with singular values and a matrix with document vectors.
- In the topic model, the word-document co-occurrence matrix is split into two parts: a topic matrix Φ and a document matrix.



Probabilistic Topic Models

❖ Matrix Factorization Interpretation

➢ LSA vs. Topic Modeling

- To find a low-dimensional representation for the content of a set of documents
- In topic models, the word and document vectors of the two decomposed matrices are probability distributions with the accompanying constraint that the feature values are non-negative and sum up to one.
- In the LDA model, additional a priori constraints are placed on the word and topic distributions

Probabilistic Topic Models

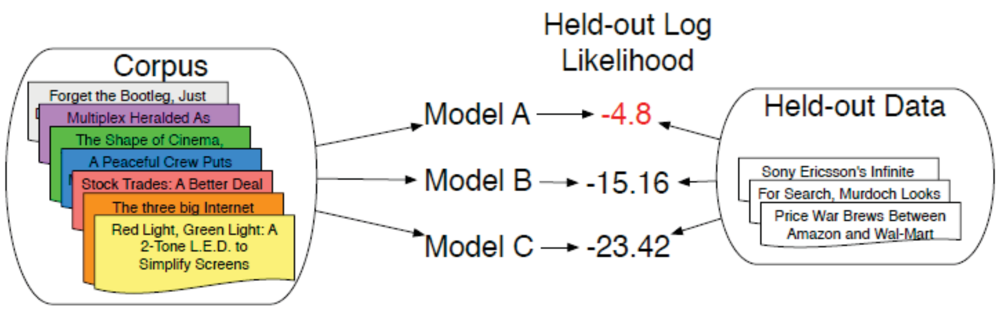
❖ Topic Models: What's Important

- Topic Models
  - Topics to words : multinomial distribution
  - Documents to topics : multinomial distribution
- Statistical structure inferred from data
- Have semantic coherence because of language use
- We use latent Dirichlet allocation (LDA) [Bleietal.2003], a fully Bayesian version of pLSI [Hofmann1999], probabilistic version of LSA [LandauerandDumais1997]

Probabilistic Topic Models

❖ Evaluation

- Likelihood



Measures predictive power, not what the topics are

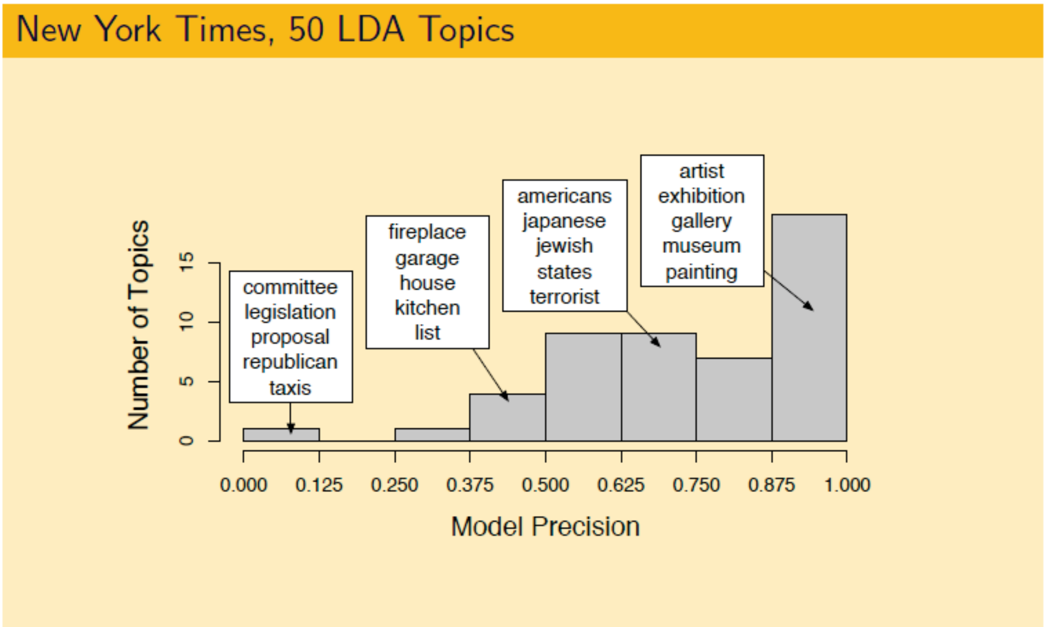
$$P(\mathbf{w} \mid \mathbf{w}', \mathbf{z}', \alpha\mathbf{m}, \beta\mathbf{u}) = \sum_{\mathbf{z}} P(\mathbf{w}, \mathbf{z} \mid \mathbf{w}', \mathbf{z}', \alpha\mathbf{m}, \beta\mathbf{u})$$

How you compute it is important too [?]

Probabilistic Topic Models

❖ Evaluation

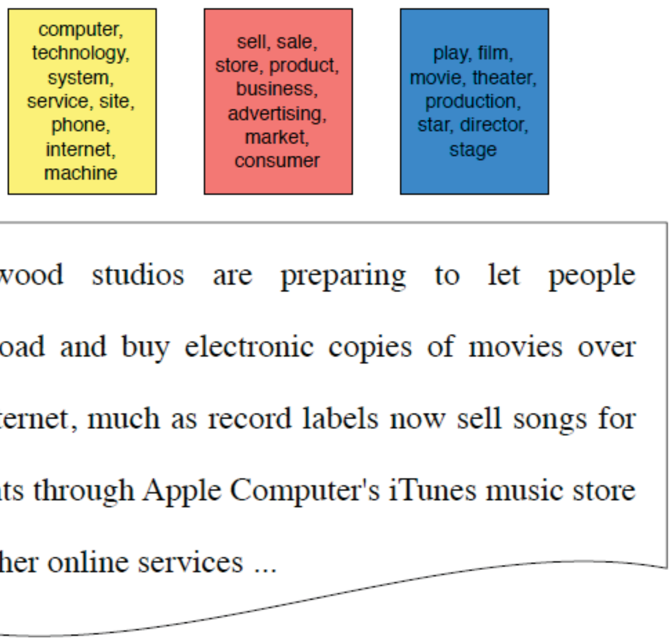
➢ Which Topics are Interpretable?



Algorithm for Extracting Topics

❖ Gibbs Sampling

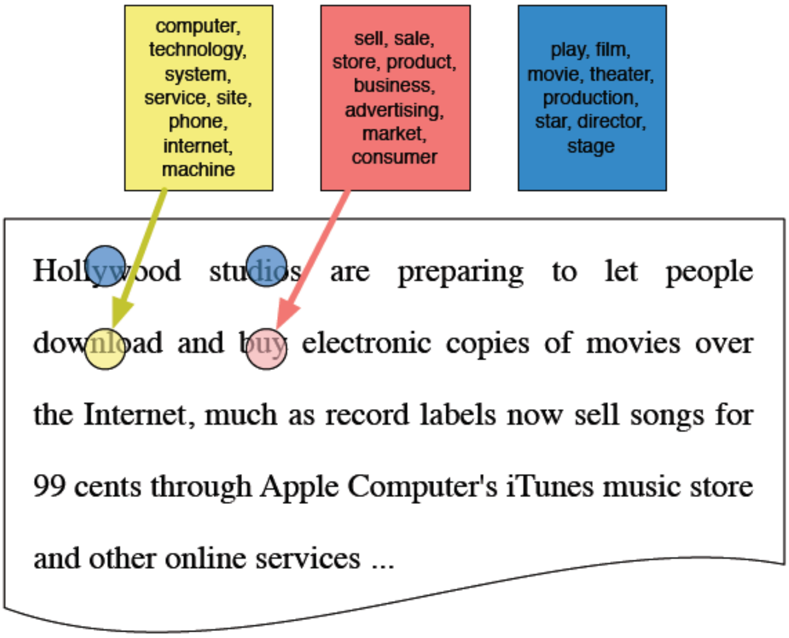
➢ Inference



Algorithm for Extracting Topics

❖ Gibbs Sampling

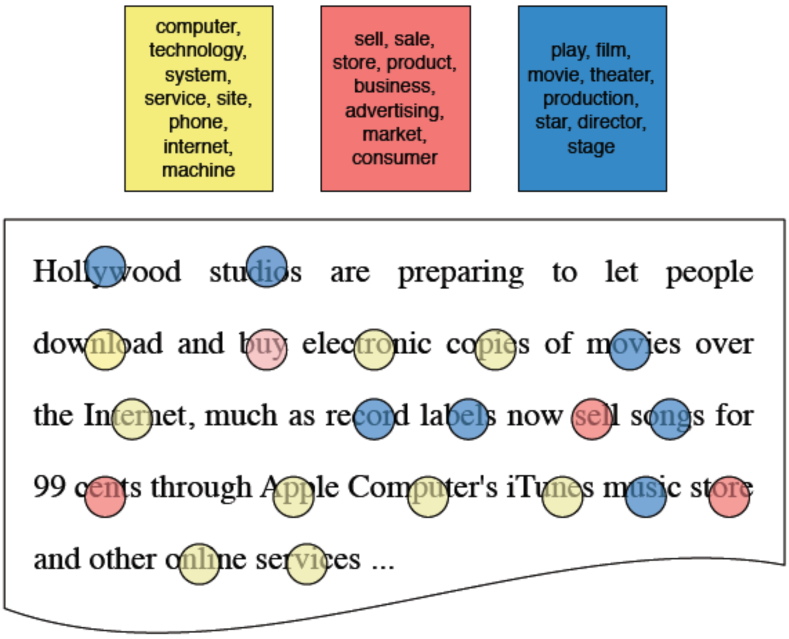
➢ Inference



Algorithm for Extracting Topics

❖ Gibbs Sampling

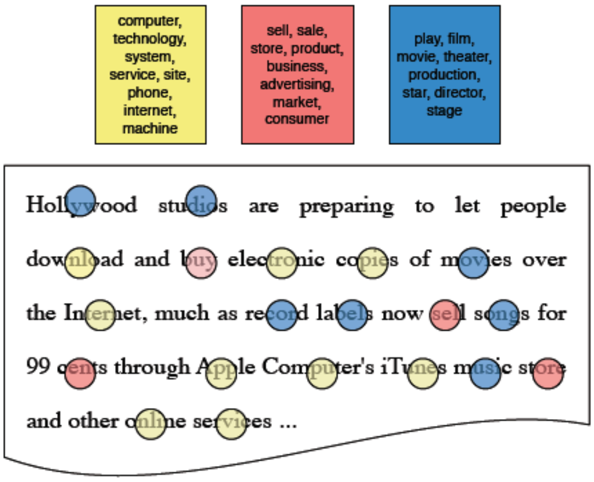
➢ Inference



Algorithm for Extracting Topics

❖ Gibbs Sampling

➢ Inference



And repeat, conditioning $z_{d,n}$ on all of the other assignments

Algorithm for Extracting Topics

❖ Gibbs Sampling

➢ A form of Markov chain Monte Carlo (MCMC)

- Hofmann (1999) used the expectation-maximization (EM) algorithm to obtain direct estimates of φ and θ; suffers from problems involving local maxima.
- Markov chain Monte Carlo (MCMC) refers to a set of approximate iterative techniques designed to sample values from complex (often high-dimensional) distributions
- To simulates a high-dimensional distribution by sampling on lower-dimensional subsets of variables where each subset is conditioned on the value of all others.
- The sampling is done sequentially and proceeds until the sampled values approximate the target distribution.
- While the Gibbs procedure we will describe does not provide direct estimates of φ and θ, we will show how φ and θ can be approximated using posterior estimates of z.

Algorithm for Extracting Topics

❖ Gibbs Sampling

➢ Algorithm

- To considers each word token in the text collection in turn, and estimates the probability of assigning the current word token to each topic, conditioned on the topic assignments to all other word tokens.
- From this conditional distribution, a topic is sampled and stored as the new topic assignment for this word token.
- P(z_i =j | z-i, w_i, d_i, ·),
  - ✓ z_i = j represents the topic assignment of token i to topic j, z-i refers to the topic assignments of all other word tokens
  - ✓ "·" refers to all other known or observed information such as all other word and document indices w-i and d-i, and hyperparameters α, and β.

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

Algorithm for Extracting Topics

❖ Gibbs Sampling

➢ Algorithm

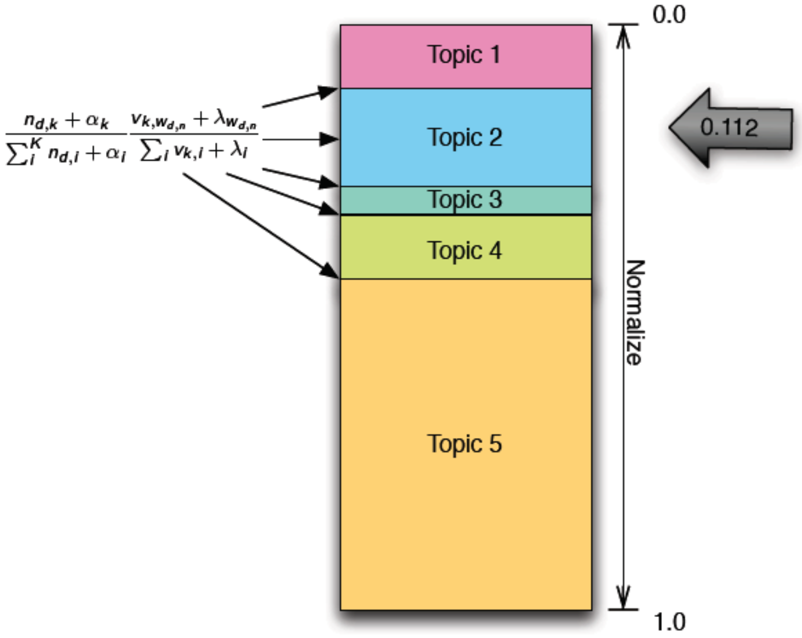
$$P(z_i = j \mid \mathbf{z}_{-i}, w_i, d_i, \cdot) \propto \frac{C^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha}$$

- C^{WT} and C^{DT} are matrices of counts with dimensions W x T and D x T respectively;
- C^{WT}_{wj} contains the number of times word w is assigned to topic j, not including the current instance i
- C^{DT}_{dj} contains the number of times topic j is assigned to some word token in document *d*, not including the current instance i.
- Note that this equation gives the unnormalized probability.
- The actual probability of assigning a word token to topic j is calculated by dividing the quantity in this equation for topic t by the sum over all topics T.
- The left part is the probability of word w under topic j
- The right part is the probability that topic j has under the current topic distribution for document d.
- Words are assigned to topics depending on how likely the word is for a topic, as well as how dominant a topic is in a document

Algorithm for Extracting Topics

❖ Gibbs Sampling

➢ How to sample from a distribution?



Algorithm for Extracting Topics

❖ Gibbs Sampling

➢ Implementation

Algorithm

1. For each iteration i:
    1. For each document d and word n currently assigned to z_{old}:
        1. Decrement $n_{d,z_{old}}$ and $v_{z_{old},w_{d,n}}$
        2. Sample $z_{new} = k$ with probability proportional to $\frac{n_{d,k}+\alpha_k}{\sum_i^K n_{d,i}+\alpha_i} \frac{v_{k,w_{d,n}}+\lambda_{w_{d,n}}}{\sum_i v_{k,i}+\lambda_i}$
        3. Increment $n_{d,z_{new}}$ and $v_{z_{new},w_{d,n}}$

Algorithm for Extracting Topics

❖ Gibbs Sampling

➢ Algorithm

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

- starts by assigning each word token to a random topic in [1..T].
- For each word token, the count matrices C^{WT} and C^{DT} are first decremented by one for the entries that correspond to the current topic assignment.
- Then, a new topic is sampled from the distribution in this equation and the count matrices C^{WT} and C^{DT} are incremented with the new topic assignment.
- Each Gibbs sample consists the set of topic assignments to all N word tokens in the corpus, achieved by a single pass through all documents.
- During the initial stage of the sampling process (also known as the burnin period), the Gibbs samples have to be discarded because they are poor estimates of the posterior.
- After the burnin period, the successive Gibbs samples start to approximate the target distribution (i.e., the posterior distribution over topic assignments)

Algorithm for Extracting Topics

❖ Gibbs Sampling

➢ Estimating parameters

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

➢ An Example

- The Gibbs sampling algorithm can be illustrated by generating artificial data from a known topic model and applying the algorithm to check whether it is able to infer the original generative structure
- Topic 1 gives equal probability to words MONEY, LOAN, and BANK

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

- Topic 2 gives equal probability to words RIVER, STREAM, and BANK

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

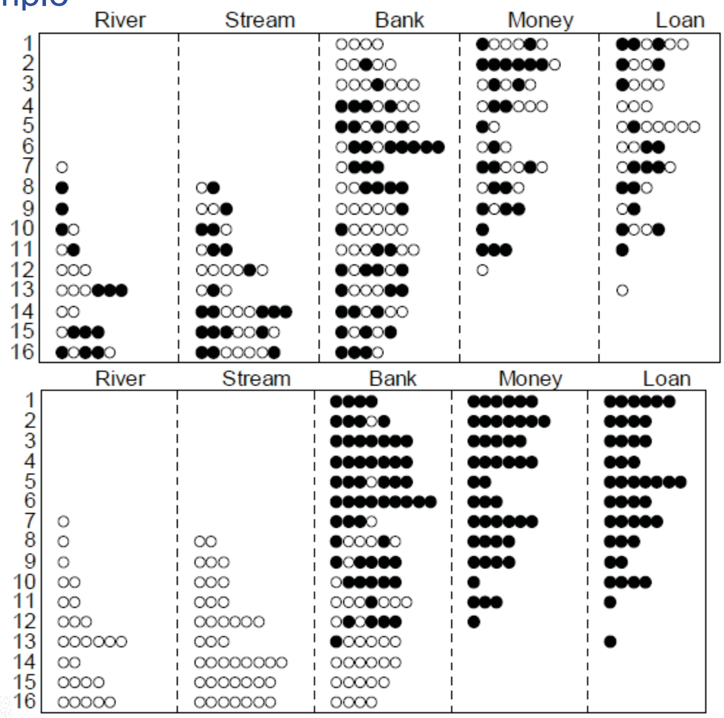
- 16 documents generated by arbitrarily mixing the two topics
- the color of the circles indicate the topic assignments (black=topic 1; white=topic 2).
- after 64 iterations: Topic 1 and Topic 2

$$\phi'^{(1)}_{MONEY} = .32,\ \phi'^{(1)}_{LOAN} = .29,\ \phi'^{(1)}_{BANK} = .39 \qquad \phi'^{(2)}_{RIVER} = .25,\ \phi'^{(2)}_{STREAM} = .4,\ \phi'^{(2)}_{BANK} = .35.$$

Algorithm for Extracting Topics

❖ Gibbs Sampling

➢ An Example



Algorithm for Extracting Topics

❖ Gibbs Sampling

➢ Exchangeability of topics

- when topics are used to calculate a statistic which is invariant to the ordering of the topics, it becomes possible and even important to average over different Gibbs samples
- Model averaging is likely to improve results because it allows sampling from multiple local modes of the posterior.

Algorithm for Extracting Topics

❖ Gibbs Sampling

- Hyperparameters: Sample them too (slice sampling)
- Initialization: Random
- Sampling: Until likelihood converges
- Lag / burn-in: Difference of opinion on this
- Number of chains: Should do more than one

➢ Available implementations

- Mallet (http://mallet.cs.umass.edu)
- LDAC (http://www.cs.princeton.edu/ blei/lda-c)
- Topicmod (http://code.google.com/p/topicmod)

Polysemy with Topics

❖ Probabilistic topic models represent semantic ambiguity through uncertainty over topics.

➢ A 300 topic solution for the TASA corpus

- The word PLAY is given relatively high probability related to the different senses of the word (playing music, theater play, playing games).

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

- Fragments of three documents are shown from TASA that use PLAY in three different senses.
- The presence of other less ambiguous words (e.g., MUSIC in the first document) builds up evidence for a particular topic in the document.
- When a word has uncertainty over topics, the topic distribution developed for the document context is the primary factor for disambiguating the word.

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}...

Computing Similarities

❖ The similarity of words and documents

➢ Two words are similar to the extent that they appear in the same topics, and two documents are similar to the extent that the same topics appear in those documents

➢ Similarity between documents

- The similarity between documents d1 and d2 can be measured by the similarity between their corresponding topic distributions and $\theta^{(d_1)}$ and $\theta^{(d_2)}$.
- The Kullback Leibler (KL) divergence

$$D(p,q)=\sum_{j=1}^{T} p_j \log_2 \frac{p_j}{q_j} \qquad KL(p,q)=\frac{1}{2}\left[D(p,q)+D(q,p)\right]$$

  ✓ This non-negative function is equal to zero when for all j, pj = qj.

- The symmetrized Jensen-Shannon (JS) divergence

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

  ✓ measures similarity between p and q through the average of p and q

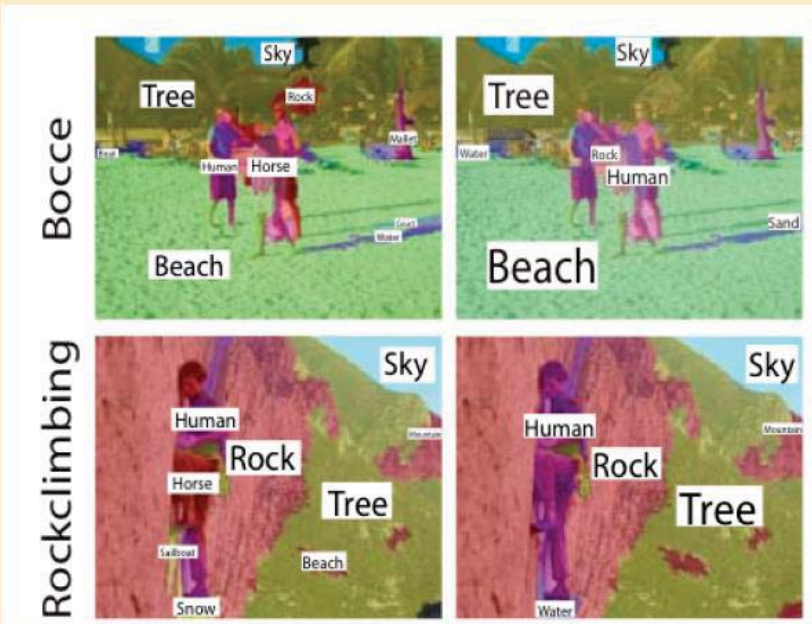
  ✓ two distributions p and q will be similar if they are similar to their average (p+q)/2

Applications

- What's a document?
- What's a word?
- What's your vocabulary?
- How do you evaluate?

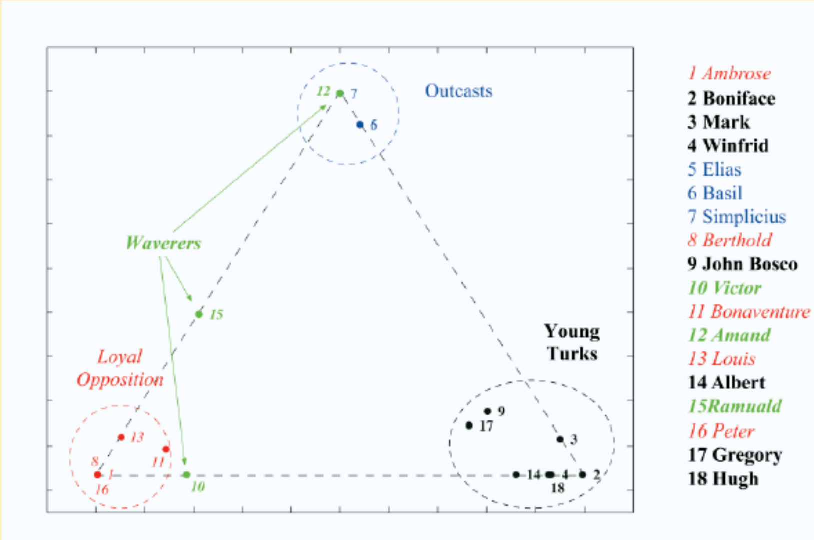
Applications

Computer Vision [*Li Fei-Fei and Perona* 2005]



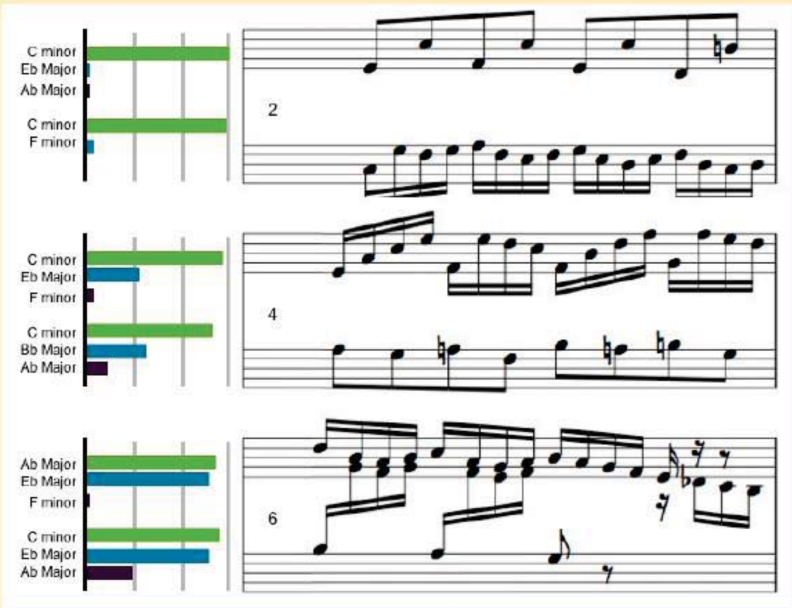
Applications

Social Networks [*Airoldi et al.* 2008]



Applications

Music [*Hu and Saul* 2009]



Q & A

Thank you for your attention !!

My websites:

web.donga.ac.kr/yjko/

Islab.donga.ac.kr

Email address:

youngjoong.ko@gmail.com