CSCI 590 : Machine Learning
Lower Dimensional Latent Semantic Space
LDA,PLSI,NMF

Halid Ziya Yerebakan

March 23 2015

Document-Term Matrices

	Words									
Docs	18	10	1	9	3	0	13	0	0	2
	22	53	9	10	5	0	6	3	7	3
	45	12	3	7	15	0	4	0	5	4
	20	31	30	20	14	2	11	0	8	1
	19	16	2	24	10	2	20	0	18	7
	8	13	1	10	14	3	4	2	1	13
	0	0	30	3	1	8	3	17	0	5
	20	69	2	6	21	0	10	7	5	4
	0	0	13	4	12	3	5	10	11	1
	1	9	0	5	0	4	1	0	3	5

Figure: Document-term matrix

- Document is vector of frequencies of the words.
- Mostly sparse matrices.
- Bag of words assumption, order information ignored.
- Other contexts : Image pixel , haplotypes , movie rating.

Latent Semantic Space

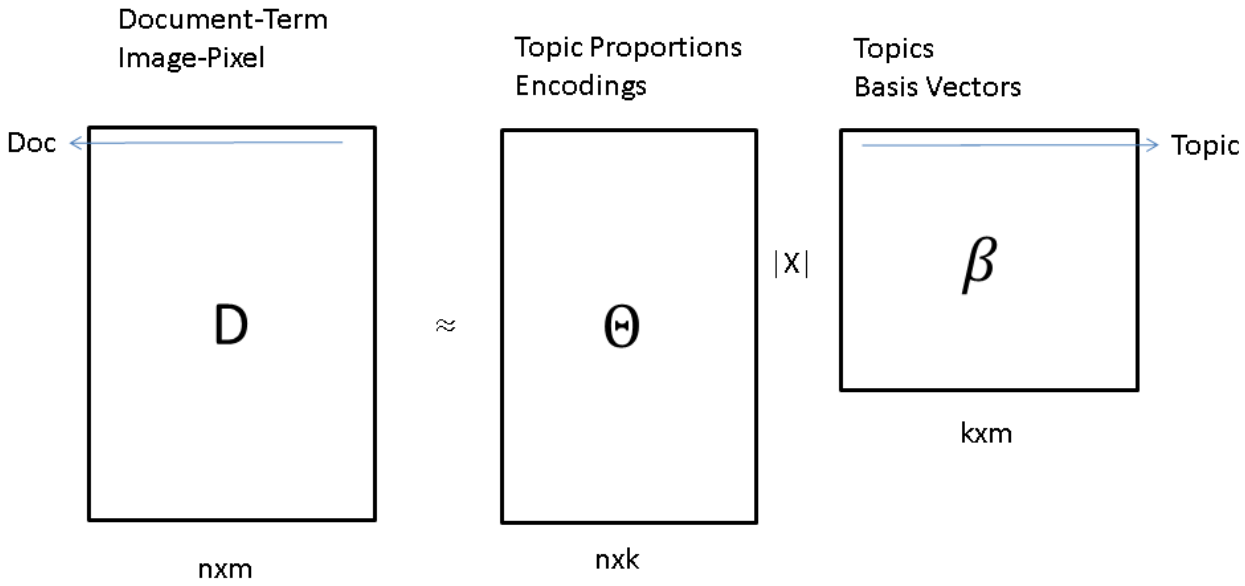


Figure: Latent Semantic Space

Example

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure: LDA on TREC AP Corpus [1]

Definitions

Definition

Topic : Topic is a distribution over words.

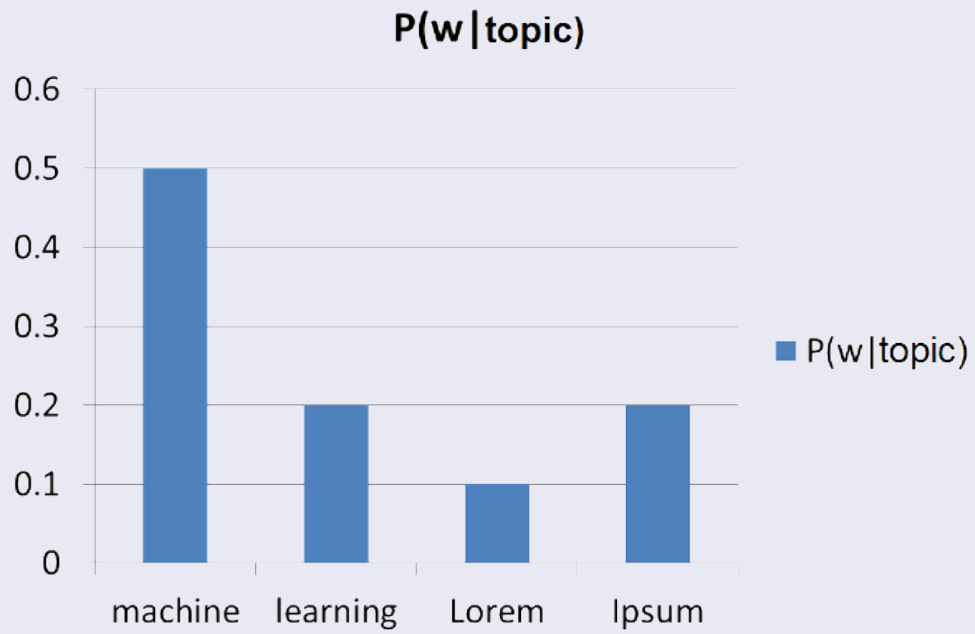


Figure: Topic distribution

Definitions

Definition

Topic proportions : Mixture weights of topics for a document. Documents have multiple topics.

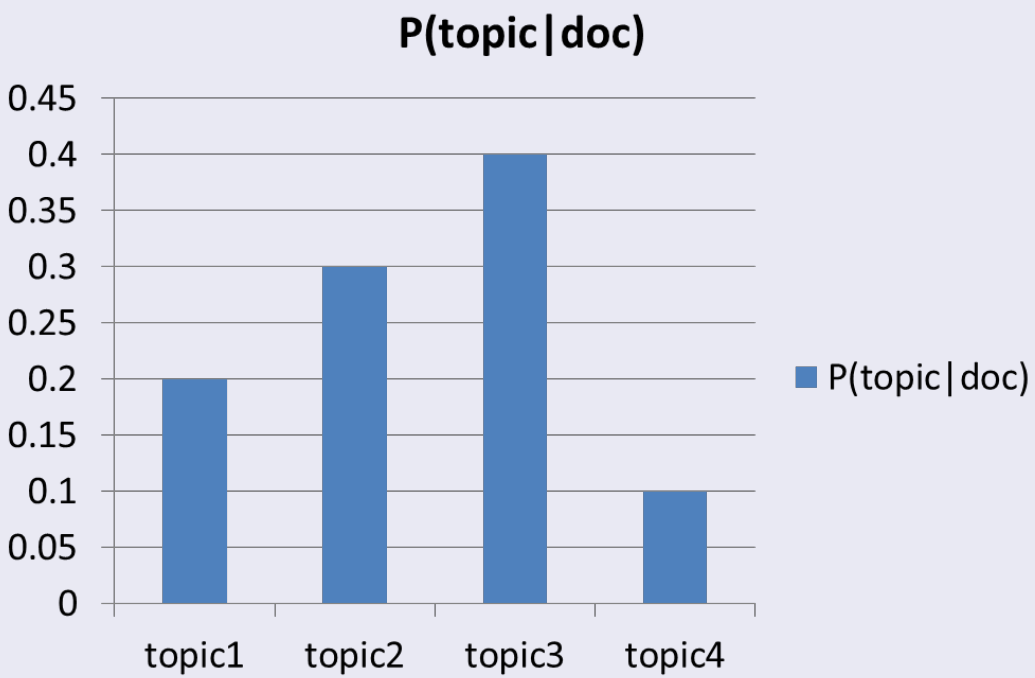


Figure: Topic proportions

Outline

PLSI References


Hofman, T. (1999, August). Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 50-57). ACM.


Hofmann, T. (1999, July). Probabilistic latent semantic analysis. In Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence (pp. 289-296). Morgan Kaufmann Publishers Inc..


Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. Machine learning, 42(1-2), 177-196.

PLSI Neighbors

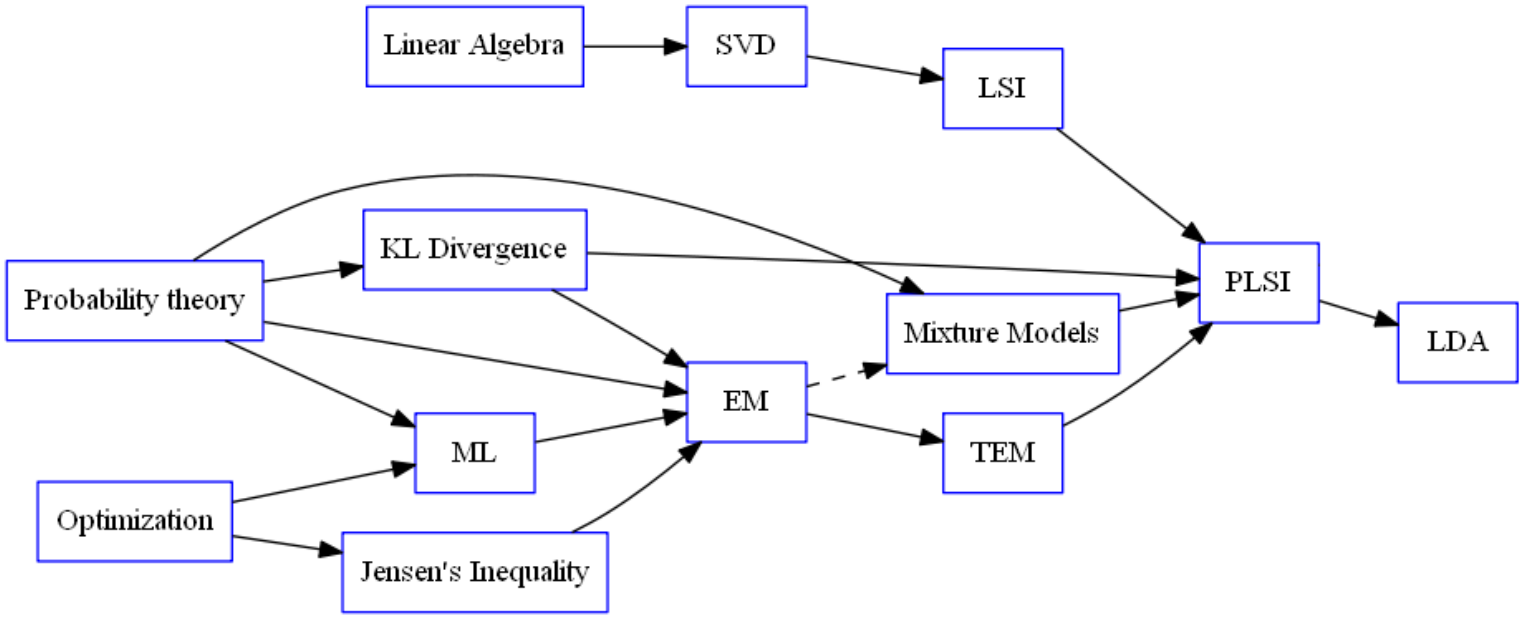


Figure: PLSI Concept Graph

Graphical Generative Model

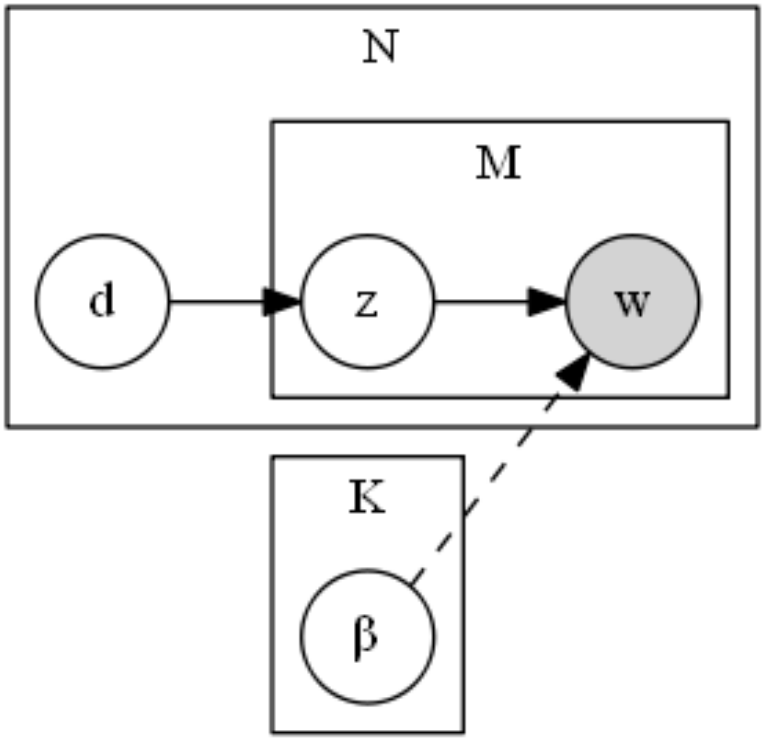


Figure: PLSI Model

Generative Model

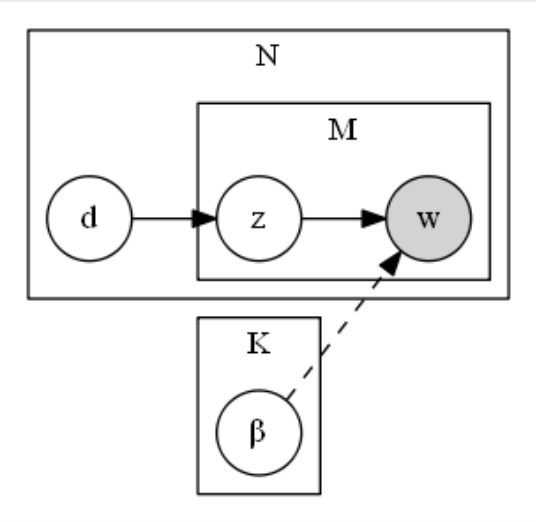


Figure: PLSI Model

Generative Process

1. Select document with probability P(d)
2. Pick a latent class (topic) z with probability P(z|d)
3. Generate a word w with probability P(w|z)

Results

Topics

"plane"	"space shuttle"	"family"	"Hollywood"
plane	space	home	film
airport	shuttle	family	movie
crash	mission	like	music
flight	astronauts	love	new
safety	launch	kids	best
aircraft	station	mother	hollywood
air	crew	life	love
passenger	nasa	happy	actor
board	satellite	friends	entertainment
airline	earth	cnn	star

Figure: Few Topics in PLSI[1]

IR Performance

	MED		CRAN		CACM		CISI	
	precision	improvement	precision	improvement	precision	improvement	precision	improvement
cos+tf	44.3	-	29.9	-	17.9	-	12.7	-
LSI	51.7	+16.7	*28.7	-4.0	*16.0	-11.6	12.7	±0.0
PLSI-U	63.1	+42.4	32.8	+9.7	19.2	+7.2	14.0	+10.2
PLSI-Q	63.9	+44.2	35.1	+17.4	22.9	+27.9	18.8	+48.0
PLSI-U*	**67.5**	**+52.4**	33.3	+11.4	19.5	+8.9	14.7	+15.7
PLSI-Q*	66.3	+49.7	**37.5**	**+25.4**	**26.8**	**+49.7**	**20.1**	**+58.3**
cos+tfidf	49.0	-	35.2	-	21.9	-	20.2	-
LSI	64.6	+31.8	38.7	+9.9	23.8	+8.7	21.9	+8.4
PLSI-U	69.5	+41.8	38.9	+10.5	25.3	+15.5	23.3	+15.3
PLSI-Q	63.2	+29.0	38.6	+9.7	26.6	+21.5	23.1	+14.4
PLSI-U*	**72.1**	**+47.1**	**40.4**	**+14.8**	27.6	+26.0	**24.6**	**+21.8**
PLSI-Q*	66.3	+35.3	40.1	+13.9	**28.3**	**+29.2**	24.4	+20.8

Figure: PLSI IR performance[1]

Advantages - Disadvantages

Advantages

- Defines proper probability distributions on words
- More realistic document model, interpretable topics.

Disadvantages

- K is fixed
- Local Maximum , Overfitting
- PLSI is not nested like LSA (n dimensional solution includes n-1 dimensional solution).
- Slow training speed compared to LSA

Advantages - Disadvantages

Advantages

- Defines proper probability distributions on words
- More realistic document model, interpretable topics.

Disadvantages

- K is fixed
- Local Maximum , Overfitting
- PLSI is not nested like LSA (n dimensional solution includes n-1 dimensional solution).
- Slow training speed compared to LSA

Outline

LDA References

 Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.

 Blei, D. M., Ng, A. Y., & Jordan, M. I. (2001). Latent dirichlet allocation. In Advances in neural information processing systems.

LDA Concept Graph

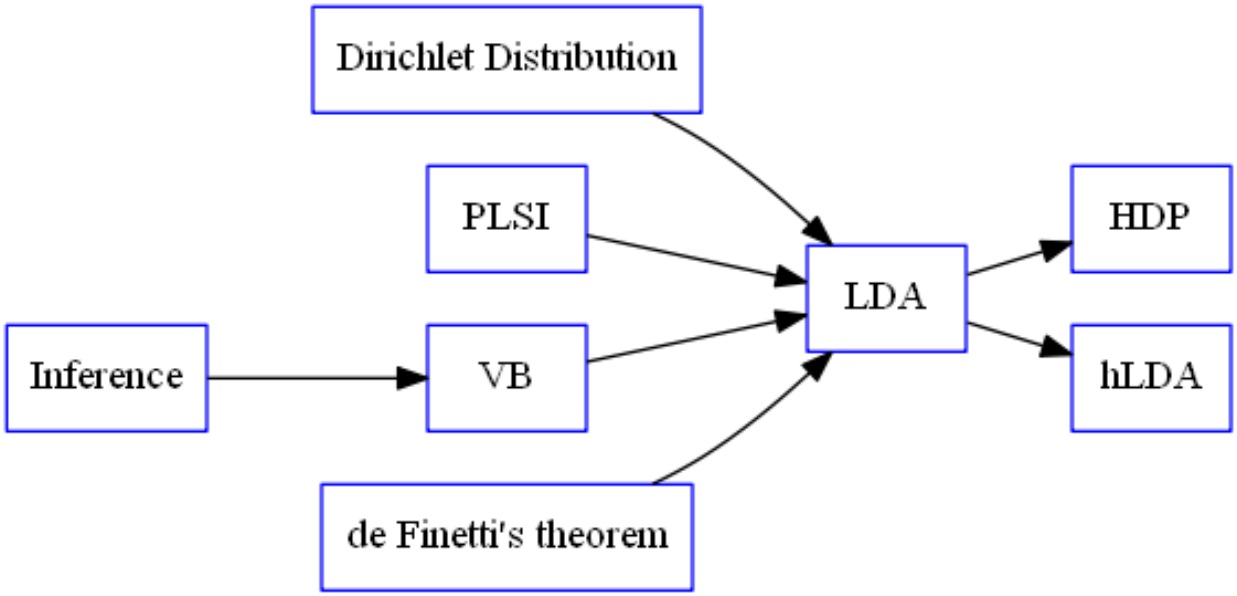


Figure: LDA Concept Graph

LDA Model

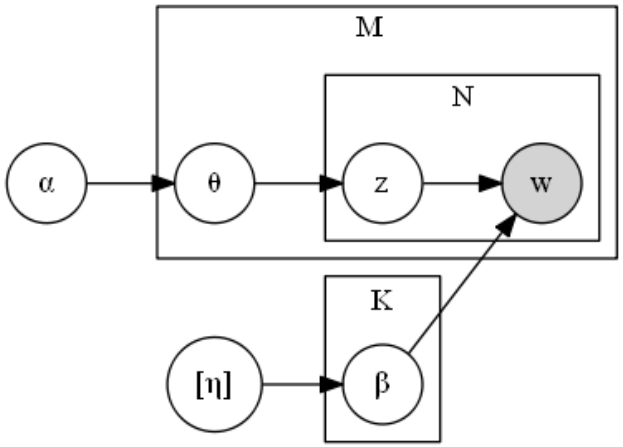


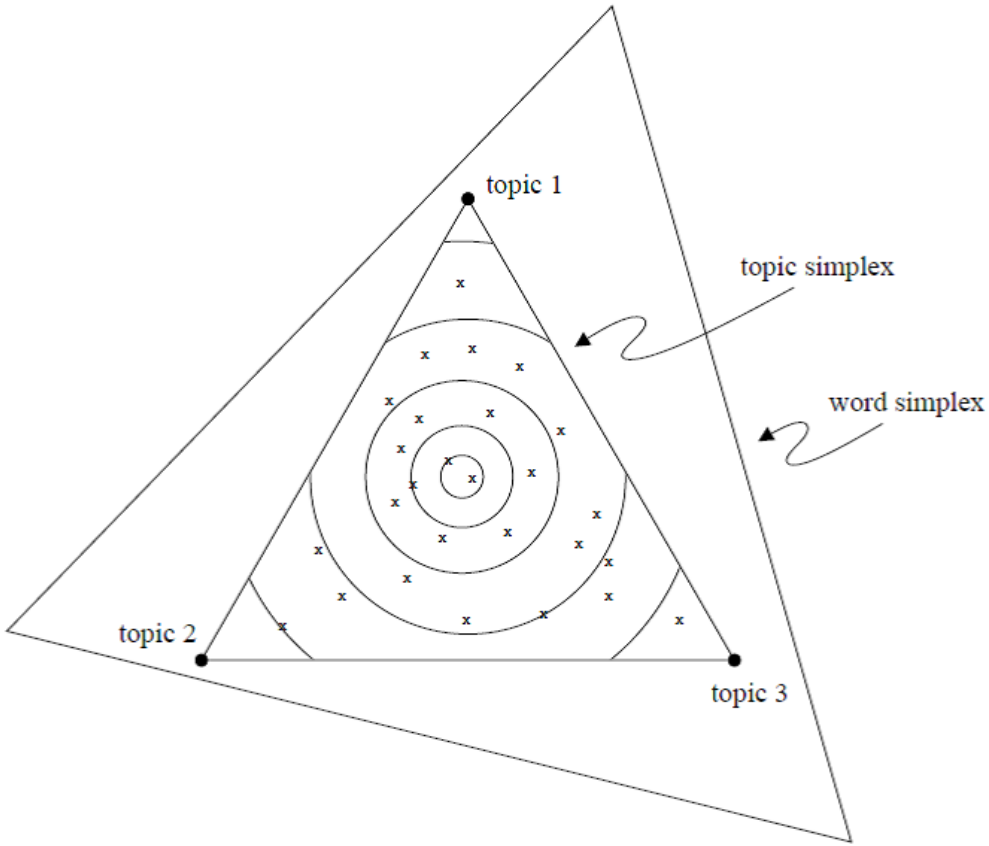
Figure: LDA Model

LDA Model

1. For each document
2. Chose N$\sim$Poisson(.)
    1. Chose $\theta \sim \text{Dir}(\alpha)$
    2. For each word w_n
        1. Chose topic $z_n \sim Multinomial(\theta)$
        2. Chose a word from $w_n \sim p(w_n|\beta_{z_n})$

$$P(\theta, \mathbf{z}, \mathbf{w}|\alpha.\beta) = P(\theta|\alpha) \prod_{n=1}^{N} P(z_n|\theta) P(w_n|z_n, \beta)$$

LDA Geometry



[1]

Experiments

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure: LDA on TREC AP Corpus [1]

Advantages-Disadvantages

Advantages

- Defines a proper generative model on documents.
- It can be incorporated in more complex models

Disadvantages

- K is fixed
- IID generative assumption of topics.

Advantages-Disadvantages

Advantages

- Defines a proper generative model on documents.
- It can be incorporated in more complex models

Disadvantages

- K is fixed
- IID generative assumption of topics.

Outline

References

 Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788-791.

 Seung, D., & Lee, L. (2001). Algorithms for non-negative matrix factorization. Advances in neural information processing systems, 13, 556-562.

Concept Graph

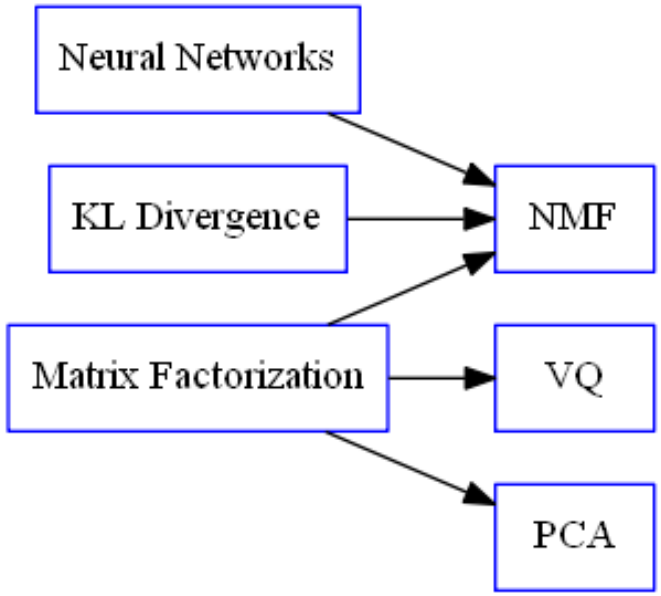


Figure: NMF Concept Graph

Non Negative Matrix Factorization

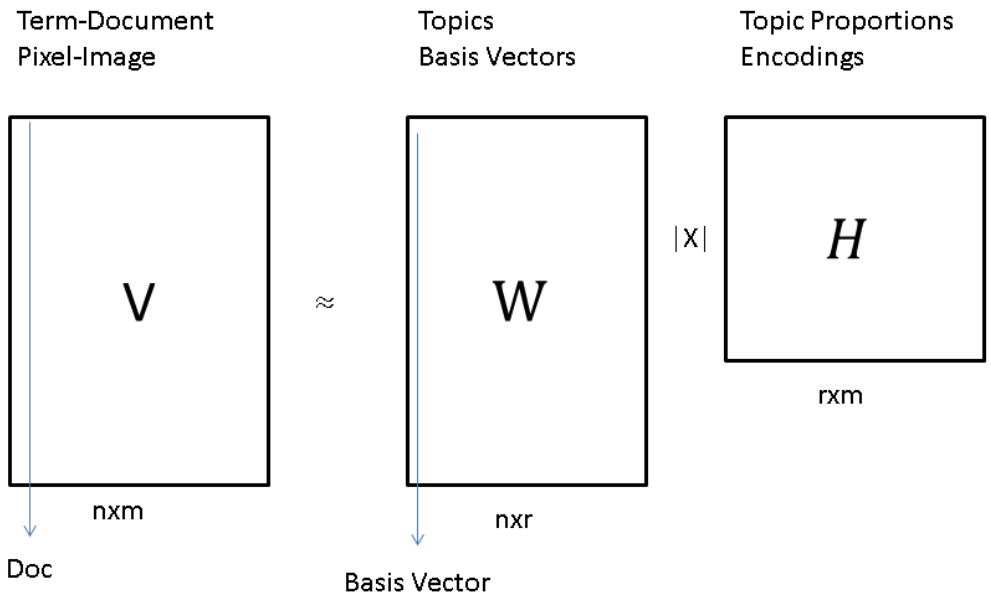


Figure: NMF

Motivation

Matrix Factorization

$$V_{i\mu} \approx (WH)_{i\mu} = \sum_{a=1}^{r} W_{ia} H_{a\mu}$$

Designed for

- Part based representation , no cancellation.
- Compressed form of data (n+m)r<nm
- Intuitive notions of basis vectors

Motivation

Matrix Factorization

$$V_{i\mu} \approx (WH)_{i\mu} = \sum_{a=1}^{r} W_{ia} H_{a\mu}$$

Designed for

- Part based representation , no cancellation.
- Compressed form of data (n+m)r<nm
- Intuitive notions of basis vectors

Algorithm

Objective Function

$$
\begin{aligned}
F &= \sum_{i=1}^{n}\sum_{\mu=1}^{m}[V_{i\mu}log(WH)_{i\mu} - (WH)_{i\mu}] \\
s.t &\quad W_{ia} \geq 0, H_{a\mu} \geq 0
\end{aligned}
$$

Algorithm

Update Equations

$$W_{ia} \leftarrow W_{ia} \sum_{\mu} \frac{V_{i\mu}}{(WH)_{i\mu}} H_{a\mu}$$

$$W_{ia} \leftarrow \frac{W_{ia}}{\sum_j W_{ja}}$$

$$H_{a\mu} \leftarrow H_{a\mu} \sum_{i} W_{ia} \frac{V_{i\mu}}{(WH)_{i\mu}}$$

Face Images

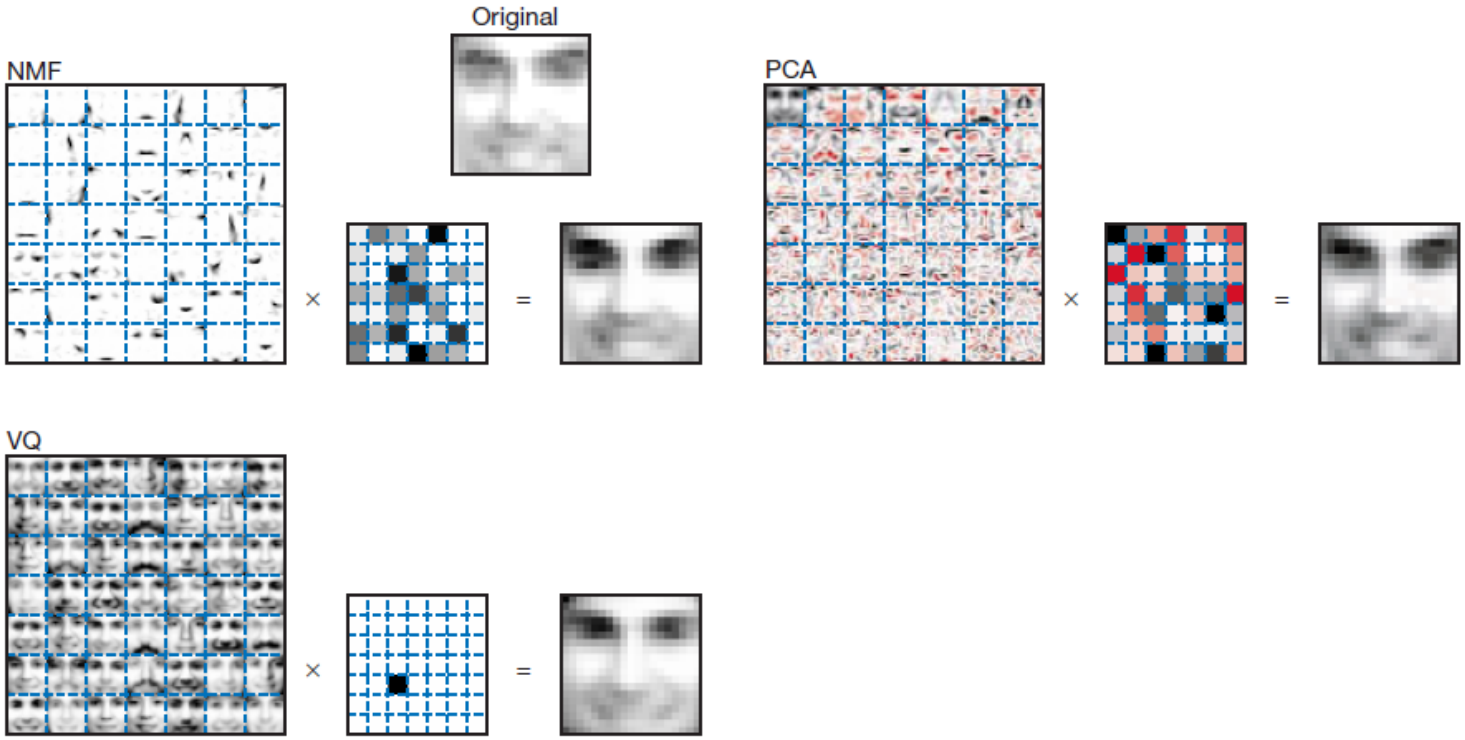


Figure: NMF,VQ,PCA on face images [1]

Advantages-Disadvantages

Advantages

- Part based representations , intuitive
- Sparse encoding

Disadvantages

- Different viewpoints or articulated objects cannot be learned.
- No learning of syntactic relation between parts.
- K is fixed.

Advantages-Disadvantages

Advantages

- Part based representations , intuitive
- Sparse encoding

Disadvantages

- Different viewpoints or articulated objects cannot be learned.
- No learning of syntactic relation between parts.
- K is fixed.

Outline

Code

- http://www.nltk.org/
- Topic Model Toolbox
- Mallet
- Matlab nnmf
- Blei's implementations