Statistical Topic Models

Kuan-Yu Chen (陳冠宇)

2020/11/13 @ TR-313, NTUST

Review

- **Latent Semantic Analysis** also called
  - Latent Semantic Indexing (LSI)
  - Latent Semantic Mapping (LSM)
  - Two-Mode Factor Analysis

- SVD is used to decompose a word-by-document matrix

$$\mathrm{A}_{|V|\times|\mathbf{D}|} = \overline{\mathrm{U}}_{|V|\times|V|}\overline{\Sigma}_{|V|\times|\mathbf{D}|}\overline{\mathrm{V}}^{\mathrm{T}}_{|\mathbf{D}|\times|\mathbf{D}|} \approx \mathrm{U}_{|V|\times K}\Sigma_{K\times K}\mathrm{V}^{\mathrm{T}}_{K\times|\mathbf{D}|} = \mathrm{A}'_{|V|\times|\mathbf{D}|}$$

  - New representations
    - For word w_i, the new vector representation is Σu_i^{T}
    - For document d_j, the new vector representation is Σv_j^{T}
  - The fold-in strategy is used to infer the query representation

$$\left(\mathrm{U}_{|V|\times K}\right)^{\mathrm{T}}(\vec{q})_{|V|\times 1} = \Sigma_{K\times K}v_q^{\mathrm{T}}$$

Introduction

- Classic IR might lead to poor retrieval due to:
  - Relevant documents that do not contain at least one index term are not retrieved
  - Synonymy (同義詞) and polysemy (一詞多義) are crucial for IR
    - Car vs. Automobile

      The prevalence of synonyms tends to decrease the **recall** performance of retrieval systems
    - Bank

      Polysemy is one factor underlying poor **precision**
  - Retrieval based on index terms is vague and noisy
    - The user information need is more related to **concepts** and ideas than to **index terms**

From LSA to Probabilistic Topic Models

- Three important claims made for LSA
  - The semantic information can derived from a word-document co-occurrence matrix
  - The dimension reduction is an essential part of its derivation
  - Words and documents can be represented as points in the Euclidean space

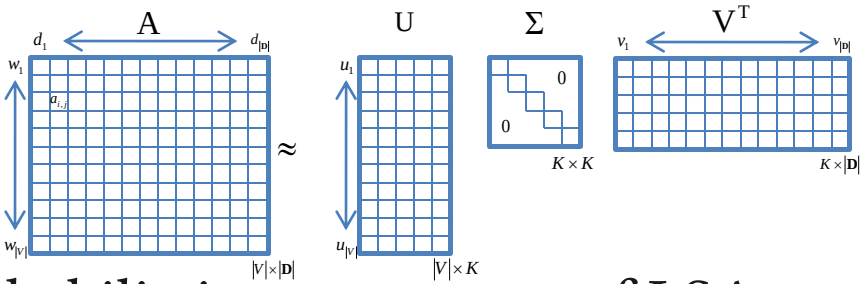
- Probabilistic topic models are consistent with the first two claims, but differs in the third one
  - The semantic properties of words and documents are expressed in terms of probabilistic topics

- **Probabilistic Latent Semantic Analysis** also called
  - Probabilistic Latent Semantic Indexing (PLSI)
  - Aspect Model

Probabilistic Latent Semantic Analysis

- LSA uses SVD to decompose a matrix

$$\mathrm{A}_{|V|\times|\mathbf{D}|} = \bar{\mathrm{U}}_{|V|\times|V|}\bar{\Sigma}_{|V|\times|\mathbf{D}|}\bar{\mathrm{V}}^{\mathrm{T}}_{|\mathbf{D}|\times|\mathbf{D}|} \approx \mathrm{U}_{|V|\times K}\Sigma_{K\times K}\mathrm{V}^{\mathrm{T}}_{K\times|\mathbf{D}|} = \mathrm{A}'_{|V|\times|\mathbf{D}|}$$



- PLSA is a probabilistic counterpart of LSA

$$P(w_i, d_j) = P(d_j)P(w_i|d_j) = P(d_j)\sum_{k=1}^{K} P(w_i|T_k)P(T_k|d_j)$$

  - $P(d_j)$: the probability of selecting document d_j
  - $P(w_i|T_k)$: the probability of word w_i condition on a latent factor/topic T_k
    - **Aspect!**
  - $P(T_k|d_j)$: the probability of a latent factor/topic T_k generated by document d_j

PLSA – 1

- The PLSA model is a latent variable model for co-occurrence data (i.e., each pair of word w_i and document d_j) which associates an unobserved class variable (i.e., latent factor T_k)

$$P(w_i, d_j) = P(d_j)P(w_i|d_j) = P(d_j)\sum_{k=1}^{K} P(w_i|T_k)P(T_k|d_j)$$

$$P(w_i|d_j) = \sum_{k=1}^{K} P(w_i, T_k|d_j) = \sum_{k=1}^{K} \frac{P(w_i, T_k, d_j)}{P(d_j)}$$

$$= \sum_{k=1}^{K} \frac{P(w_i, d_j|T_k)P(T_k)}{P(d_j)}$$

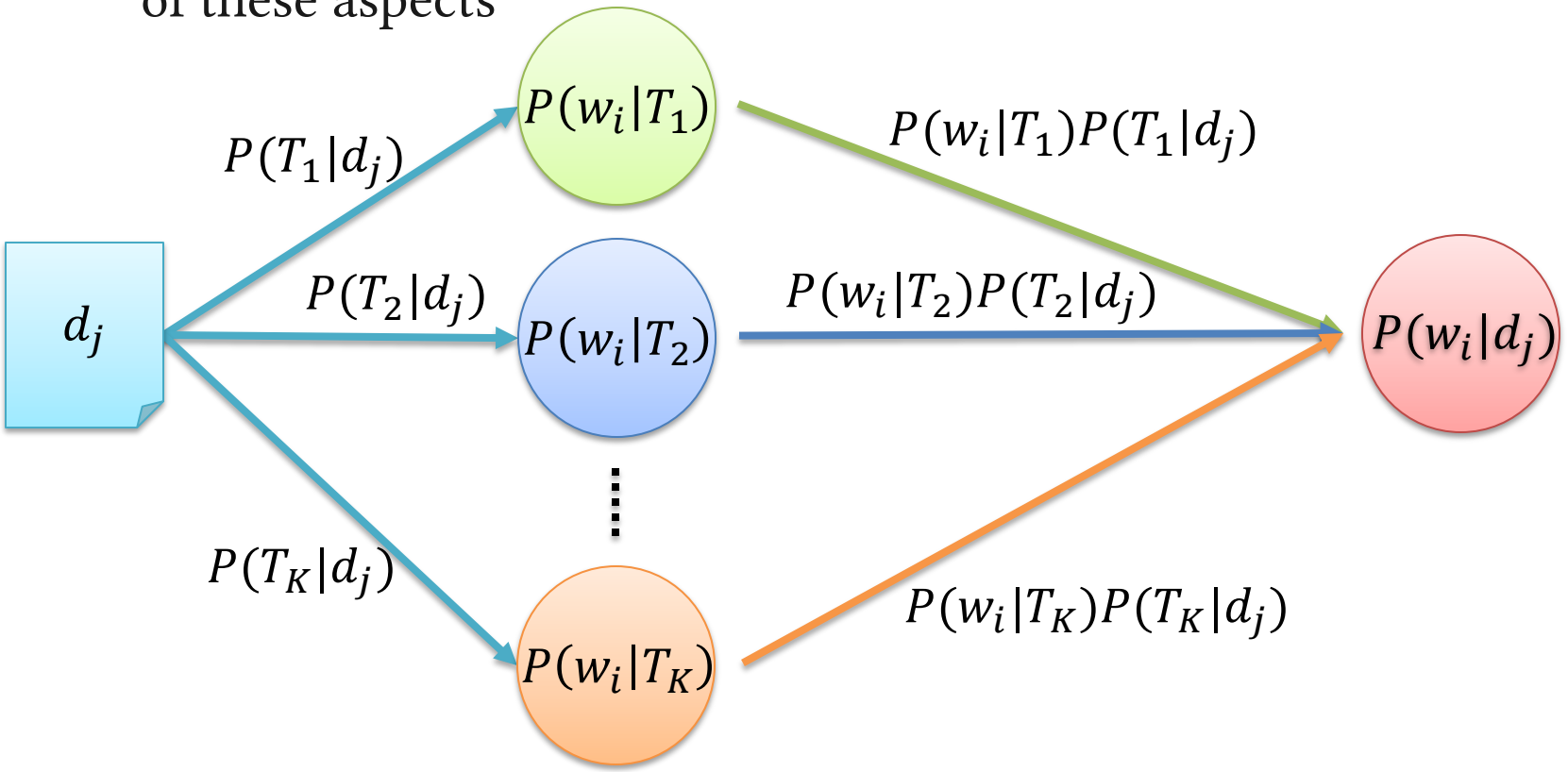
$$= \sum_{k=1}^{K} \frac{P(w_i|T_k)P(d_j|T_k)P(T_k)}{P(d_j)}$$

$$= \sum_{k=1}^{K} \frac{P(w_i|T_k)P(d_j, T_k)}{P(d_j)} = \sum_{k=1}^{K} P(w_i|T_k)P(T_k|d_j)$$

Conditional Independence Assumption
document and word are independent conditioned on the state of the associated latent variable

PLSA – 2

- The goal of PLSA is to
  - identify conditional probability mass functions $P(w_i|T_k)$
  - the document-specific word distributions $P(w_i|d_j)$ are as faithfully as possible approximated by **convex combinations** of these aspects



PLSA – 3

- The training objective is defined to maximize the total log-likelihood of a given training collection
  - The model parameters are $P(d_j)$, $P(w_i|T_k)$, and $P(T_k|d_j)$

$$P(w_i, d_j) = P(d_j)P(w_i|d_j) = P(d_j)\sum_{k=1}^{K} P(w_i|T_k)P(T_k|d_j)$$

$$\mathcal{L} = \sum_{w_i \in V}\sum_{d_j \in \mathbf{D}} c(w_i, d_j)\log P(w_i, d_j)$$

$$= \sum_{w_i \in V}\sum_{d_j \in \mathbf{D}} c(w_i, d_j)\log\left(P(d_j)\sum_{k=1}^{K} P(w_i|T_k)P(T_k|d_j)\right)$$

PLSA – 4.

- By using the **Expectation-Maximization algorithm**
  - E-step

$$P(T_k|w_i, d_j) = \frac{P(w_i|T_k)P(T_k|d_j)}{\sum_{k=1}^{K} P(w_i|T_k)P(T_k|d_j)}$$

  - M-step

$$P(w_i|T_k) = \frac{\sum_{d_j \in \mathbf{D}} c(w_i, d_j)P(T_k|w_i, d_j)}{\sum_{i'=1}^{|V|} \sum_{d_j \in \mathbf{D}} c(w_{i'}, d_j)P(T_k|w_{i'}, d_j)}$$

$$P(T_k|d_j) = \frac{\sum_{i=1}^{|V|} c(w_i, d_j)P(T_k|w_i, d_j)}{\sum_{i'=1}^{|V|} c(w_{i'}, d_j)} = \frac{\sum_{i=1}^{|V|} c(w_i, d_j)P(T_k|w_i, d_j)}{|d_j|}$$

PLSA – 4..

- About the E-step:

$$P(T_k|w_i,d_j) = \frac{P(w_i|T_k)P(T_k|d_j)}{\sum_{k=1}^{K} P(w_i|T_k)P(T_k|d_j)}$$

$$= \frac{P(w_i|T_k)\frac{P(T_k,d_j)}{P(d_j)}\frac{P(T_k)}{P(T_k)}}{\sum_{k=1}^{K} P(w_i|T_k)P(T_k|d_j)} = \frac{P(w_i|T_k)P(d_j|T_k)\frac{P(T_k)}{P(d_j)}}{\sum_{k=1}^{K} P(w_i|T_k)P(T_k|d_j)}$$

$$= \frac{P(w_i,d_j|T_k)\frac{P(T_k)}{P(d_j)}}{\sum_{k=1}^{K} P(w_i|T_k)P(T_k|d_j)} = \frac{\frac{P(w_i,d_j,T_k)}{P(d_j)}}{\sum_{k=1}^{K} P(w_i|T_k)P(T_k|d_j)}$$

$$= \frac{\frac{P(w_i,d_j,T_k)}{P(d_j)}}{\sum_{k=1}^{K}\frac{P(w_i,d_j,T_k)}{P(d_j)}} = \frac{P(w_i,d_j,T_k)}{P(w_i,d_j)}$$

PLSA – 5

- Consequently, for a given pair of query and document, the relevance degree can be determined by combining unigram model and PLSA model

$$P(w_i|d_j) = \frac{c(w_i, d_j)}{|d_j|}$$

$$P(q|d_j) \equiv \prod_{i=1}^{|q|} P(w_i|d_j)$$

$$= \prod_{i=1}^{|q|} \left(\alpha \cdot P(w_i|d_j) + (1-\alpha) \cdot P_{PLSA}(w_i|d_j)\right)$$

$$= \prod_{i=1}^{|q|} \left[\alpha \cdot P(w_i|d_j) + (1-\alpha) \cdot \left(\sum_{k=1}^{K} P(w_i|T_k) P(T_k|d_j)\right)\right]$$

  - In order to incorporate the general information, the background model can also be integrated

$$P(q|d_j) = \prod_{i=1}^{|q|} \left[\alpha \cdot P(w_i|d_j) + \beta \cdot \left(\sum_{k=1}^{K} P(w_i|T_k) P(T_k|d_j)\right) + (1-\alpha-\beta) \cdot P_{BG}(w_i)\right]$$

PLSA – 6

- For a new document d_m, the **fold-in** strategy can be perform to obtain the topic distribution $P(T_k|d_m)$ for the document
  - E-step

$$P(T_k|w_i, d_m) = \frac{P(w_i|T_k)P(T_k|d_m)}{\sum_{k=1}^{K} P(w_i|T_k)P(T_k|d_m)}$$

  - M-step
    - The word distribution for each topic $P(w_i|T_k)$ is fixed

$$P(T_k|d_m) = \frac{\sum_{i=1}^{|V|} c(w_i, d_m)P(T_k|w_i, d_m)}{\sum_{i'=1}^{|V|} c(w_{i'}, d_m)}$$

PLSA – 7

- In addition to the query likelihood measure, we can combine PLSA with the **vector space model**
  - Each document has its own distribution over topic $P(T_k|d_j)$
  - Query can be treated as a document, thus the fold-in strategy can be perform to obtain $P(T_k|q)$
  - The topic distributions for document and query are vector representations
  - The similarity degree can be estimated under the semantic space

$$sim(q, d_j) = \cos\left(\vec{q}, \vec{d}_j\right) = \frac{\sum_{k=1}^{K} P(T_k|q)P(T_k|d_j)}{\sqrt{\sum_{k=1}^{K} P(T_k|q)^2}\sqrt{\sum_{k=1}^{K} P(T_k|d_j)^2}}$$

Link PLSA and LSA

- Another derivation of PLSA model

$$P(w_i, d_j) = \sum_{k=1}^{K} P(w_i, d_j, T_k)$$

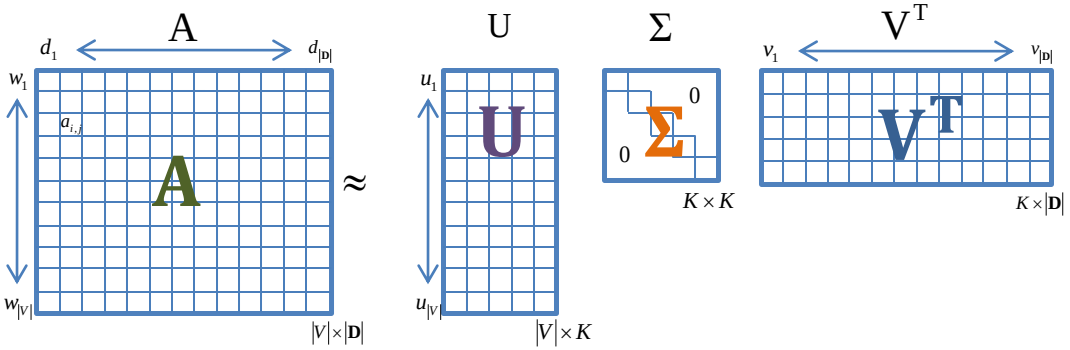
$$= \sum_{k=1}^{K} P(w_i|d_j, T_k)P(d_j, T_k)$$

$$= \sum_{k=1}^{K} P(w_i|T_k)P(d_j, T_k)$$

$$= \sum_{k=1}^{K} P(w_i|T_k)P(T_k)P(d_j|T_k)$$

Conditional Independence Assumption
document and word are independent conditioned on the state of the associated latent variable

$$P(T_k)P(d_j|T_k) = P(T_k)\frac{P(d_j, T_k)}{P(T_k)}$$
$$= P(d_j, T_k) = P(T_k|d_j)P(d_j)$$



Comparisons – PLSA & LSA

- Decomposition/Approximation
  - LSA: least-squares criterion measured on the L2- or Frobenius norms of the word-by-document matrix
  - PLSA: maximization of the collection likelihood, which implies to minimize the cross-entropy loss

- Computational complexity
  - LSA: SVD decomposition
  - PLSA: EM training
  - The model complexity of Both LSA and PLSA grows linearly with the number of training documents
  - There is no general way to estimate or predict the vector representation (of LSA) or the model parameters (of PLSA) for a newly observed document
    - Fold-in strategy

Revisiting the Objective Function

$$\mathcal{L} = -\sum_{w_i \in V}\sum_{d_j \in \mathbf{D}} c(w_i, d_j)\log P(w_i, d_j)$$

$$H(T,E) = -\sum_{x \in \mathbf{X}} T(x) log E(x)$$

$$= -\sum_{w_i \in V}\sum_{d_j \in \mathbf{D}} c(w_i, d_j)\log\left(P(d_j)\sum_{k=1}^{K} P(w_i|T_k)P(T_k|d_j)\right)$$

$$= -\sum_{d_j \in \mathbf{D}}\sum_{w_i \in V} c(w_i, d_j)\left[\log P(d_j) + \log\left(\sum_{k=1}^{K} P(w_i|T_k)P(T_k|d_j)\right)\right]$$

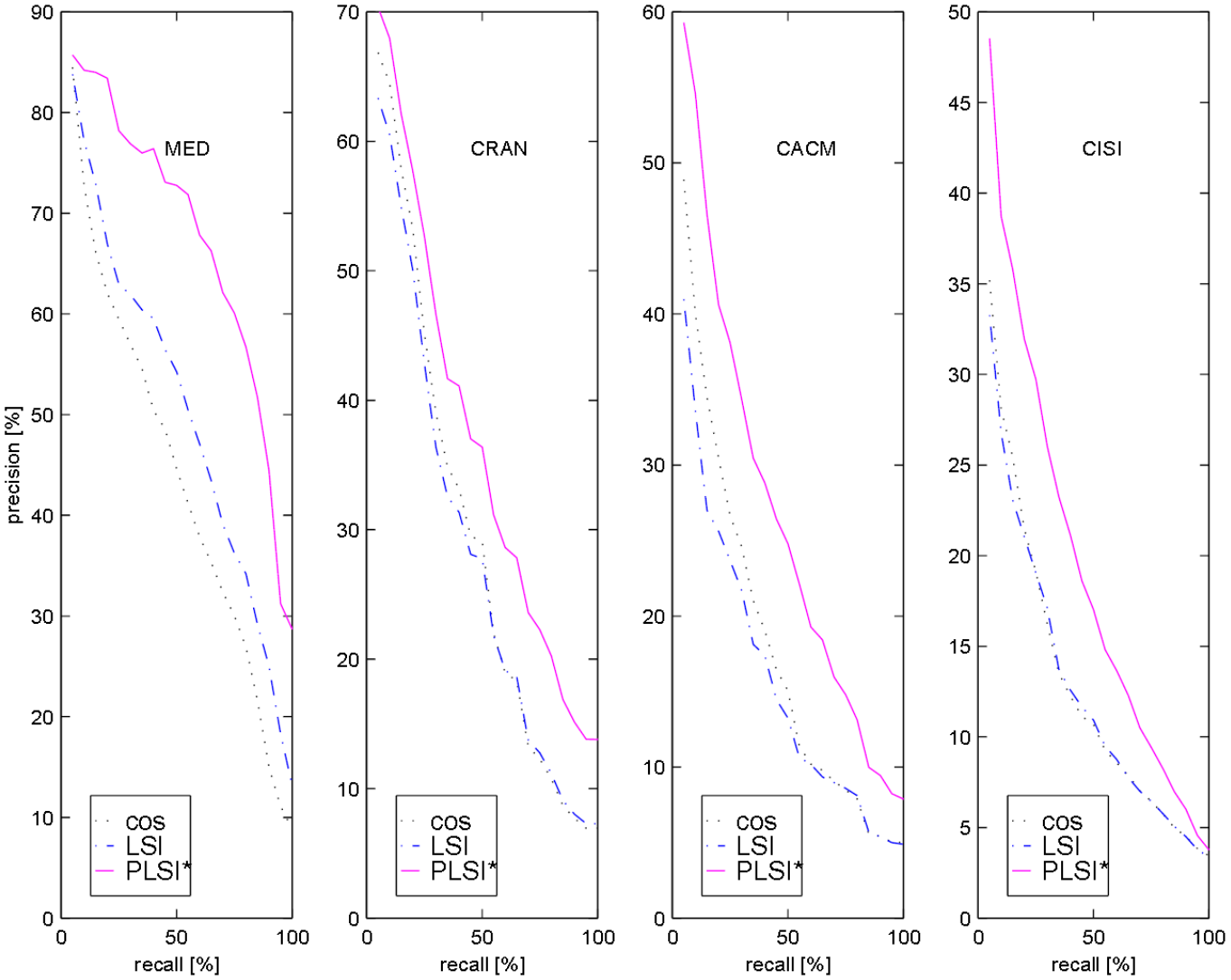
$$= -\sum_{d_j \in \mathbf{D}}\sum_{w_i \in V} |d_j|\frac{c(w_i, d_j)}{|d_j|}\left[\log P(d_j) + \log\left(\sum_{k=1}^{K} P(w_i|T_k)P(T_k|d_j)\right)\right]$$

$$= -\sum_{d_j \in \mathbf{D}} |d_j| \sum_{w_i \in V} P(w_i|d_j)\left[\log P(d_j) + \log P_{PLSA}(w_i|d_j)\right]$$

$$= \sum_{d_j \in \mathbf{D}} |d_j| \sum_{w_i \in V} \left(\underbrace{-P(w_i|d_j)\log P(d_j)}_{\text{Constant}} \underbrace{- P(w_i|d_j)\log P_{PLSA}(w_i|d_j)}_{\text{Cross Entropy}}\right)$$

Comparisons – Experiments

- All of the results are based on cosine similarity measure



Comparisons – Factors/Topics

- Factors from a 128 factor decomposition of the TDT-1 corpus
  - Factors are represented by their 10 most probable words, i.e., the words are ordered according to $P(w_i|T_k)$

- There is no obvious interpretation of the directions in the LSA latent space, while the directions in the PLSA space are interpretable as multinomial word distributions

"plane"	"space shuttle"	"family"	"Hollywood"	"Bosnia"	"Iraq"	"Rwanda"	"Kobe"
plane	space	home	film	un	iraq	refugees	building
airport	shuttle	family	movie	bosnian	iraqi	aid	city
crash	mission	like	music	serbs	sanctions	rwanda	people
flight	astronauts	love	new	bosnia	kuwait	relief	rescue
safety	launch	kids	best	serb	un	people	buildings
aircraft	station	mother	hollywood	sarajevo	council	camps	workers
air	crew	life	love	nato	gulf	zaire	kobe
passenger	nasa	happy	actor	peacekeepers	saddam	camp	victims
board	satellite	friends	entertainment	nations	baghdad	food	area
airline	earth	cnn	star	peace	hussein	rwandan	earthquake

Comparisons – Polysemy

- Many words in natural language are polysemous, having multiple senses; their semantic ambiguity can only be resolved by other words in the context
  - For example, 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

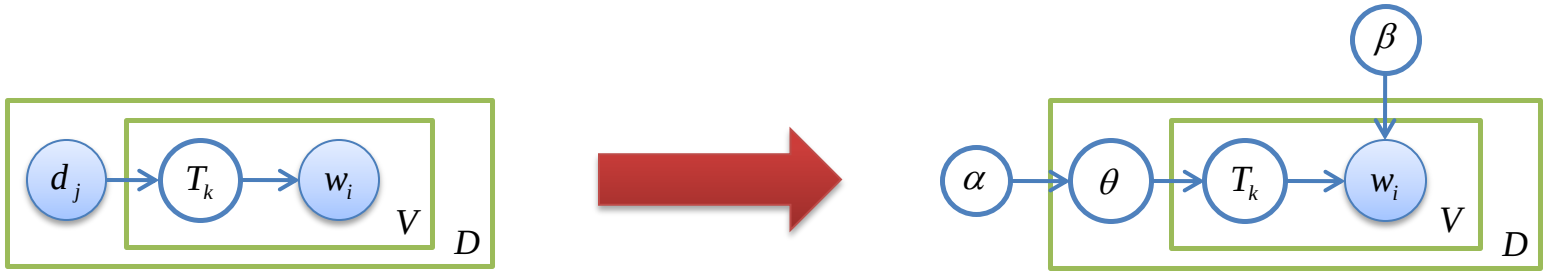
word	prob.
LITERATURE	.031
POEM	.028
POETRY	.027
POET	.020
PLAYS	.019
POEMS	.019
PLAY	.015
LITERARY	.013
WRITERS	.013
DRAMA	.012
WROTE	.012
POETS	.011
WRITER	.011
SHAKESPEARE	.010
WRITTEN	.009
STAGE	.009

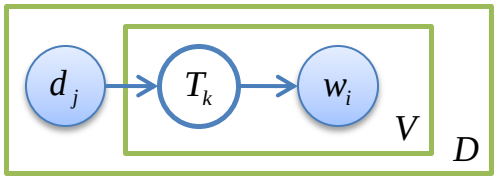
Topic 166

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

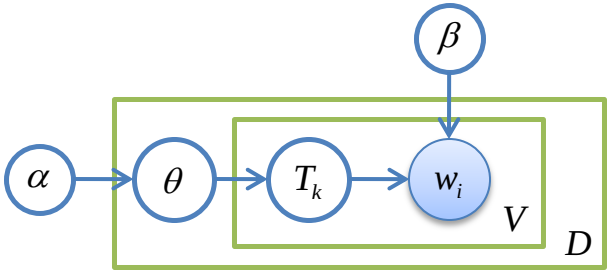
From PLSA to Latent Dirichelet Allocation

- In traditional topic models, there are several problems:
  - The model parameters grow linearly with the size of the corpus
    - EM is time-consuming
  - It is not clear how to assign probability to a document outside of the training set
    - Fold-in is a compromising strategy
    - Retrain the model is time-consuming





PLSA & LDA – 1



- PLSA
  - PLSA assumes that the model parameters are fixed and unknown

$$\mathcal{L} = \prod_{w_i \in V} \prod_{d_j \in \mathbf{D}} P(w_i, d_j)^{c(w_i, d_j)} = \prod_{d_j \in \mathbf{D}} \prod_{i=1}^{|d_j|} P(w_i, d_j)$$

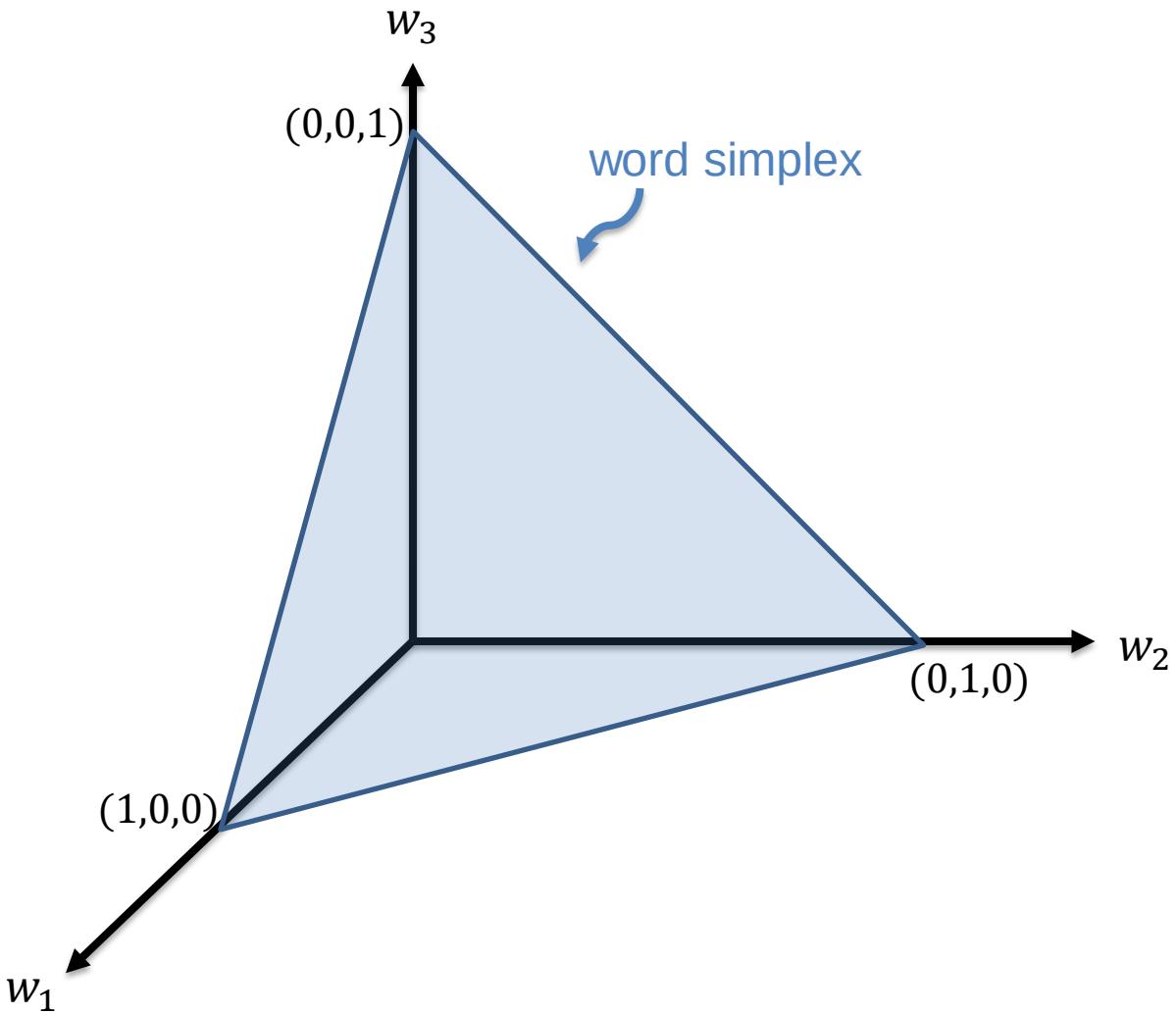
$$= \prod_{d_j \in \mathbf{D}} \prod_{i=1}^{|d_j|} \left(P(d_j) \sum_{k=1}^{K} P(w_i|T_k) P(T_k|d_j) \right)$$

- LDA
  - LDA places a priori constraints on the model parameters
    - Dirichelet distribution

$$\mathcal{L} = \prod_{d_j \in \mathbf{D}} \int P(\theta_{d_j}|\alpha) \left(\prod_{i=1}^{|d_j|} \left(\sum_{k=1}^{K} P(w_i|T_k, \beta) P\left(T_k \middle| \theta_{d_j}\right) \right) \right) d\theta_{d_j}$$

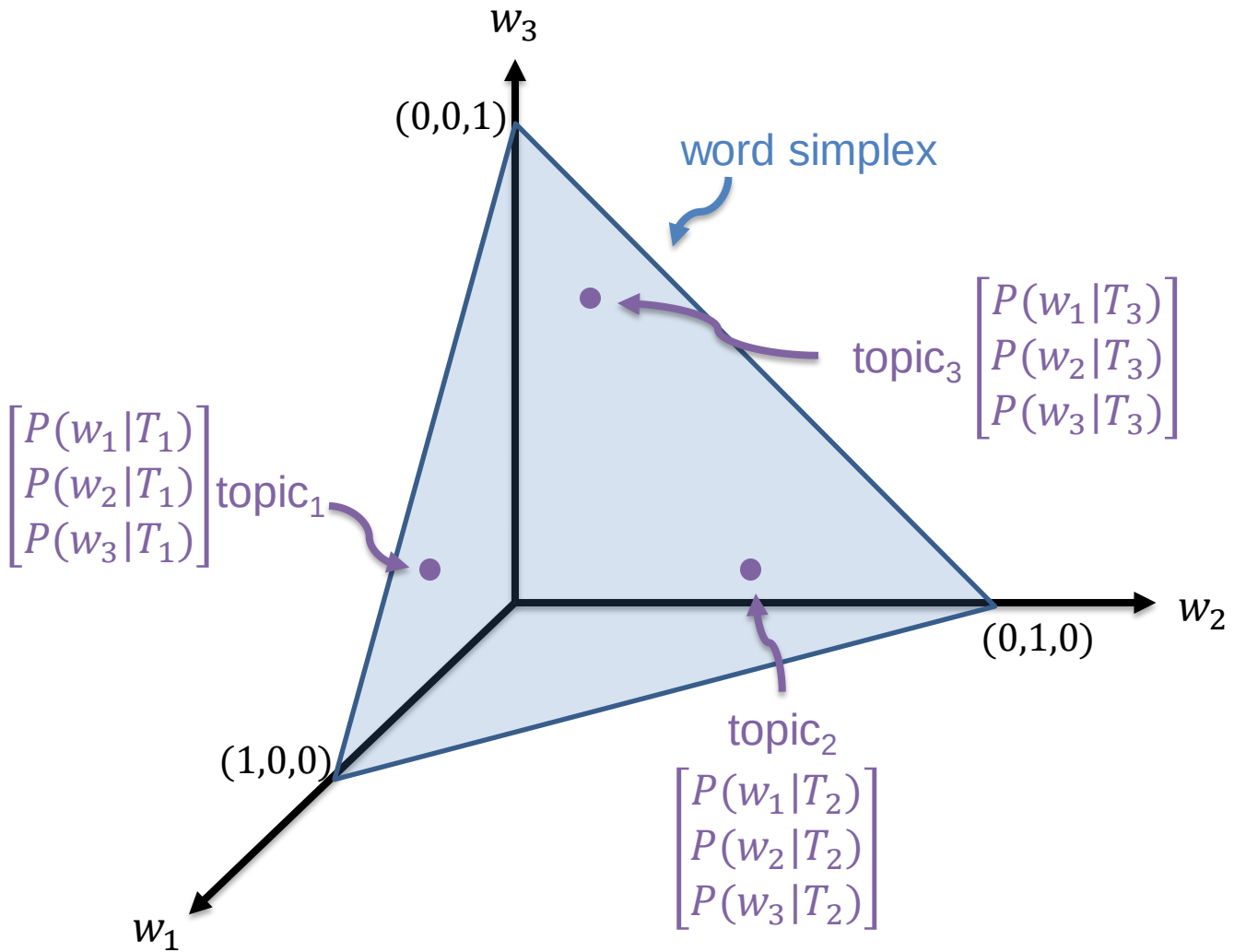
PLSA & LDA – 2

- The topic simplex for three topics embedded in the word simplex for three words



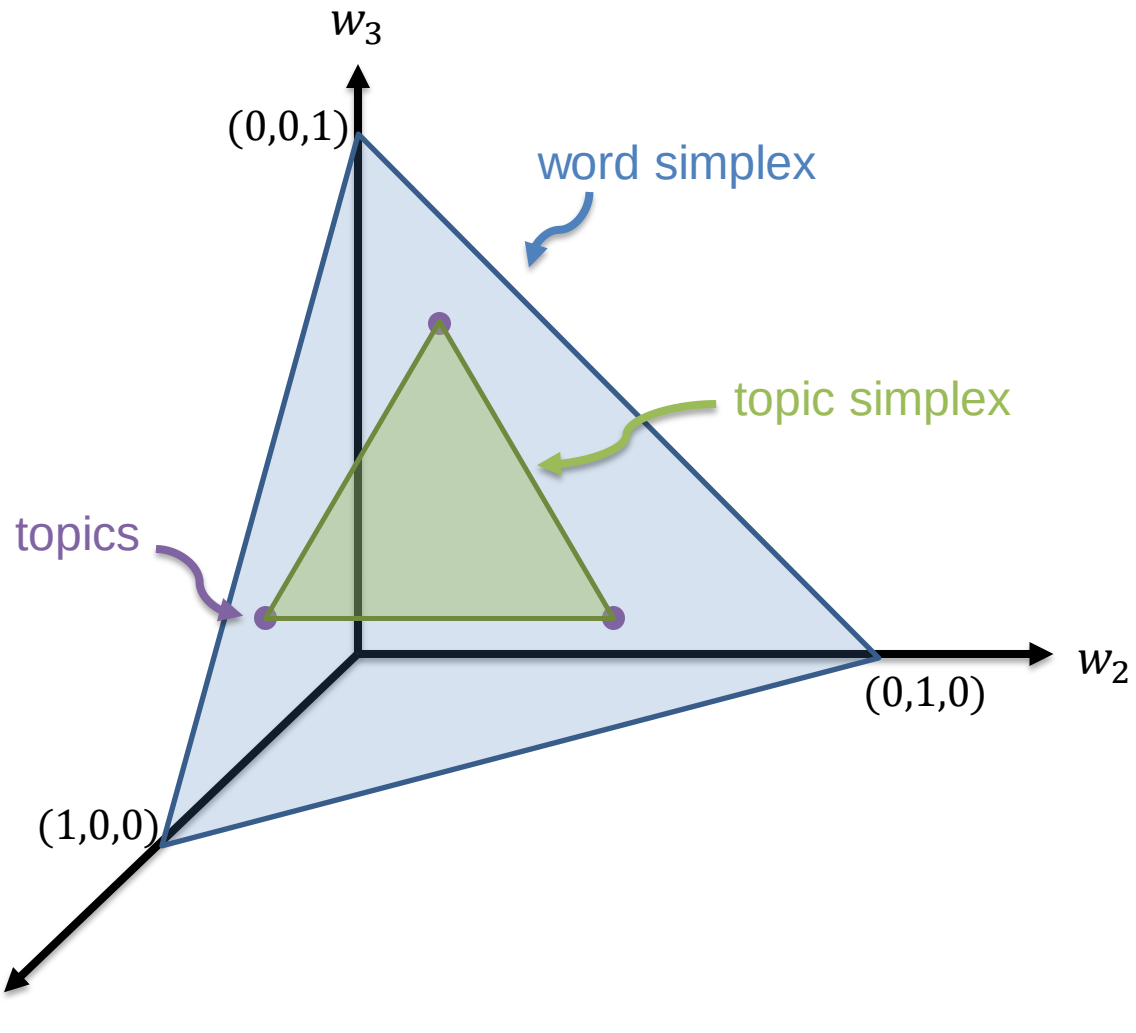
PLSA & LDA – 3

- The topic simplex for three topics embedded in the word simplex for three words



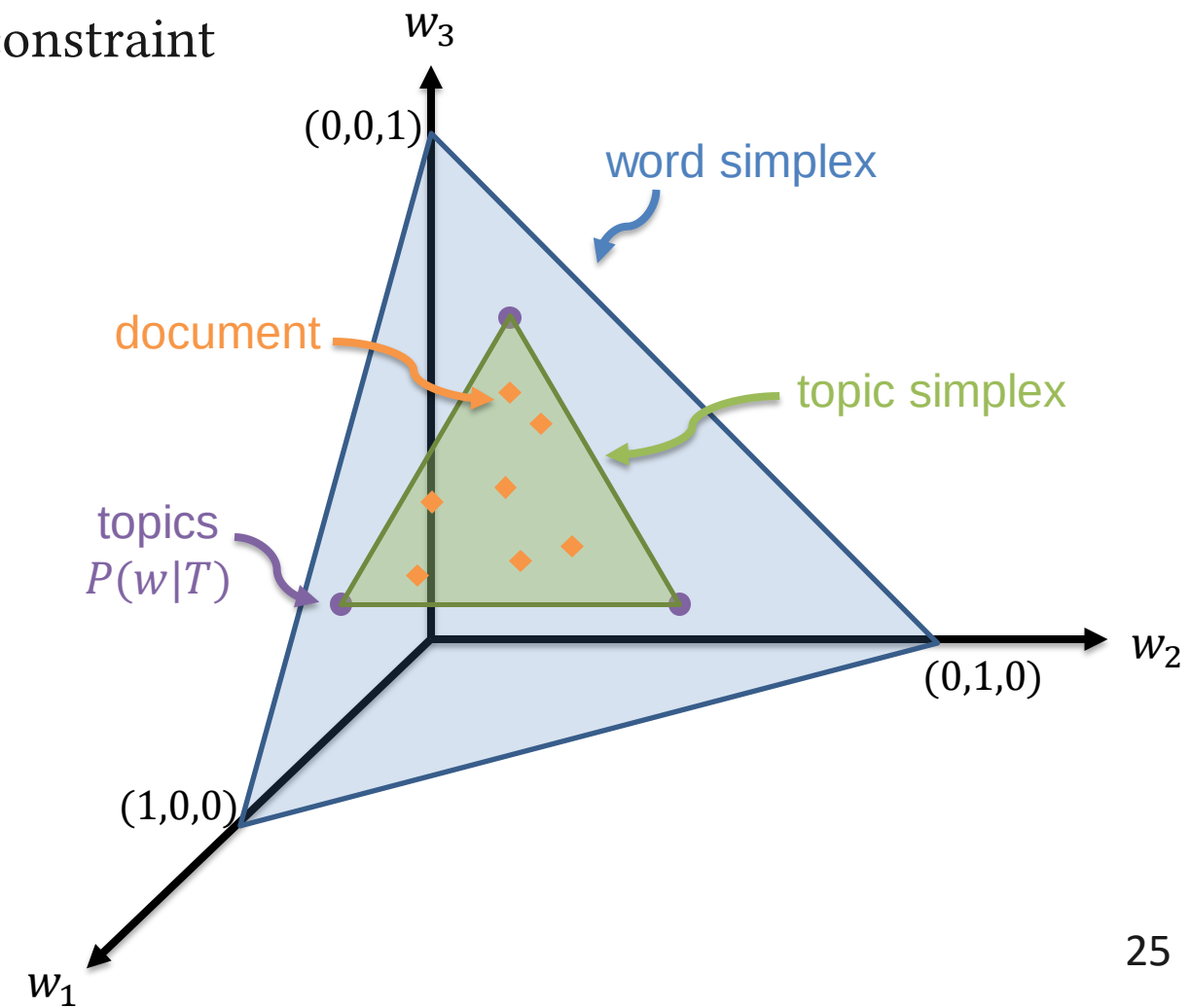
PLSA & LDA – 4

- The topic simplex for three topics embedded in the word simplex for three words



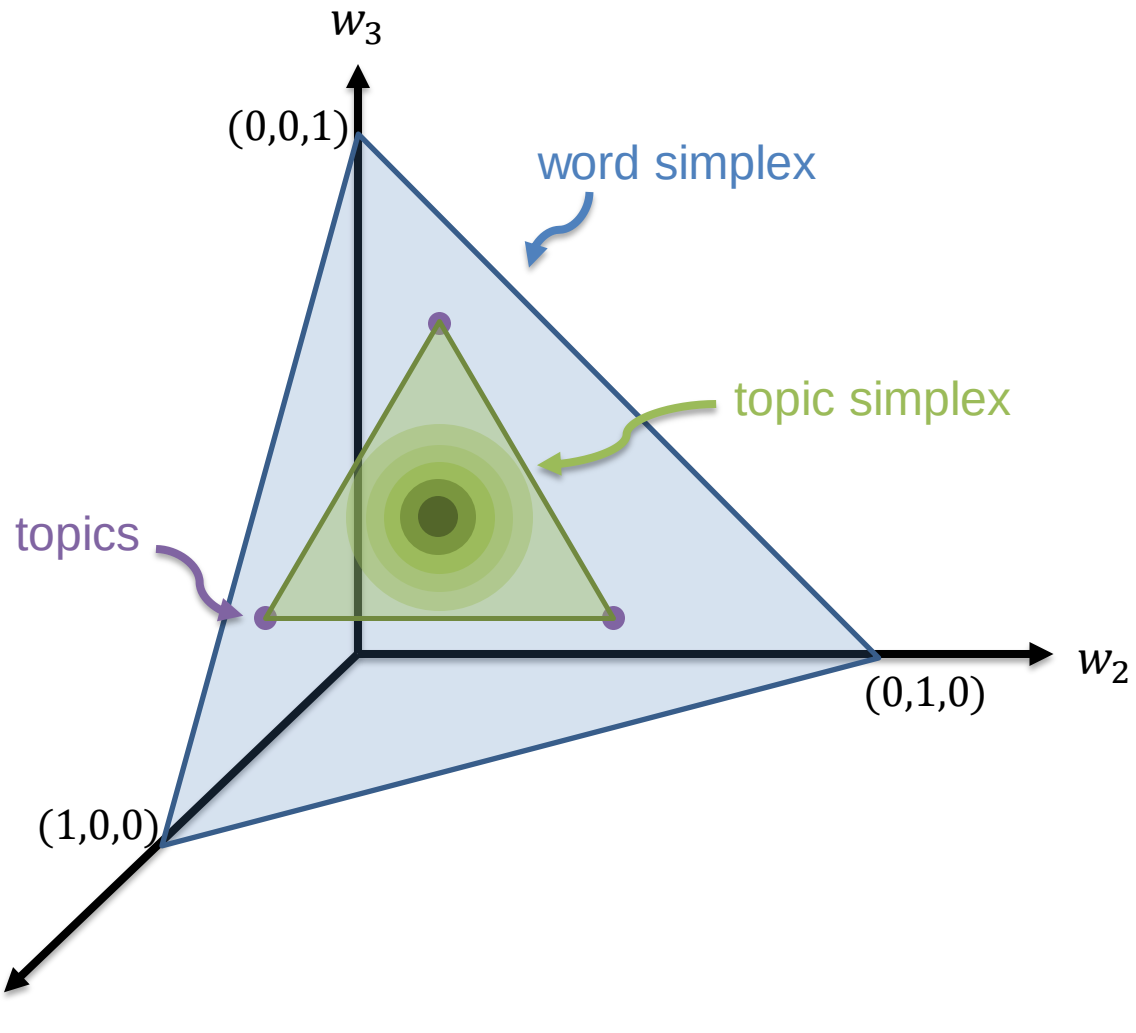
PLSA & LDA – 5

- The topic simplex for three topics embedded in the word simplex for three words
  - PLSA: no prior constraint



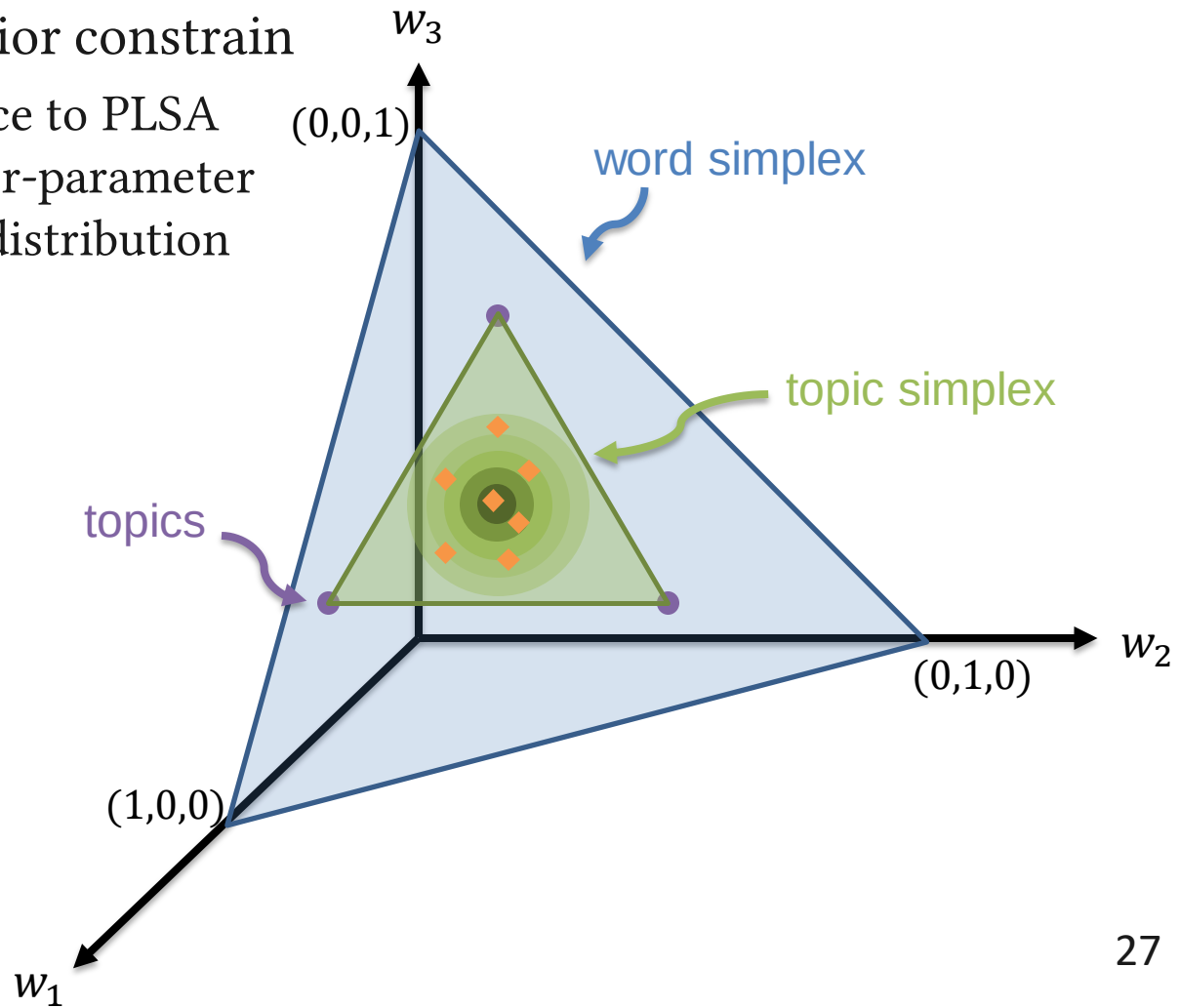
PLSA & LDA – 6

- The topic simplex for three topics embedded in the word simplex for three words

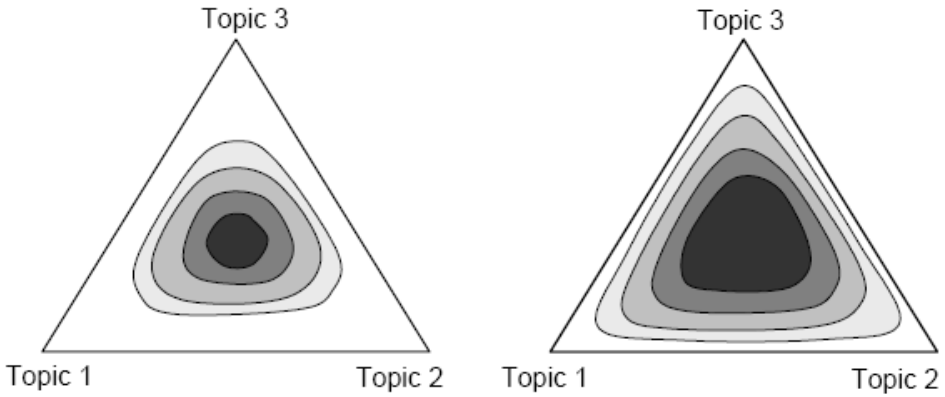


PLSA & LDA – 7

- The topic simplex for three topics embedded in the word simplex for three words
  - LDA: follow a prior constrain
    - LDA will reduce to PLSA when the hyper-parameter for Dirichelet distribution sets to 1



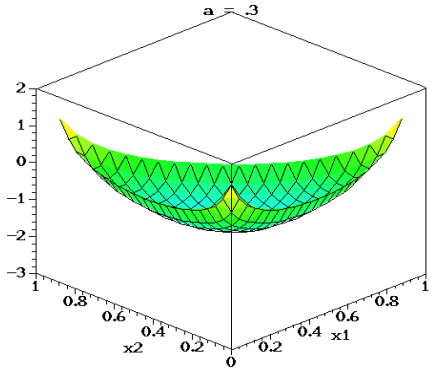
PLSA & LDA – 8



$\alpha = 4.0$ $\alpha = 2.0$

- Dirichlet priors 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

$$P(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})}\prod_{k=1}^{K}\theta_k^{\alpha_k-1} = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_k^{\alpha_k-1}$$



LDA – Experiments

- QL: query likelihood measure
- CBDM: cluster-based model (simplified variant of PLSA)
- LBDM: LDA model

Collection	QL	CBDM	LBDM	%chg over QL	%chg over CBDM
AP	0.2179	0.2326	0.2651	+21.64*	+13.97*
FT	0.2589	0.2713	0.2807	+7.54*	+3.46*
SJMN	0.2032	0.2171	0.2307	+13.57*	+6.26*
LA	0.2468	0.2590	0.2666	+8.02[2]	+2.93
WSJ	0.2958	0.2984	0.3253	+9.97*	+9.01*

The Evolution

David M. Blei
Columbia University, USA

Thomas Hofmann
ETH Zurich, Switzerland

Scott Deerwester

2003 Latent Dirichlet Allocation

1999 Probabilistic Latent Semantic Analysis

1998 Language Modeling Approaches

1994 Best Match Models (Okapi Systems)

1988 Latent Semantic Analysis

1976 Probabilistic Model

1975 Vector Space Model

1973 Boolean Model

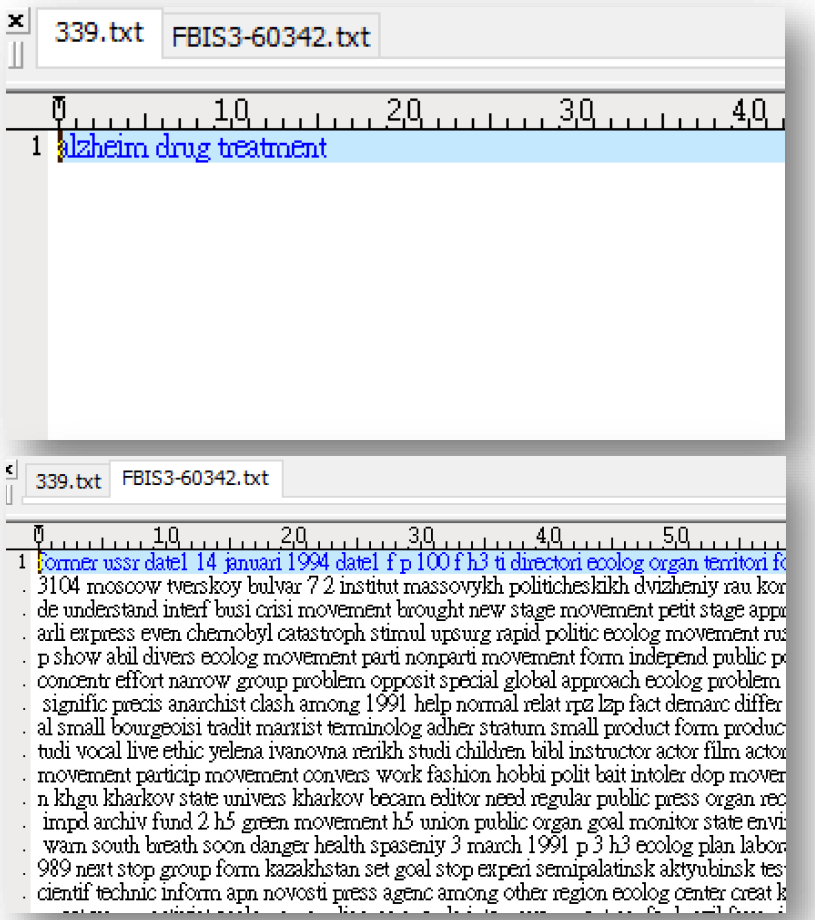
1972 Inverse Document Frequency

1957 Term Frequency

Homework 4 – PLSA

Homework 4 - Description.

- In this project, we have
  - 100 Queries
  - 14995 Documents
  - Our goal is to implement the PLSA model, and incorporate the PLSA and query likelihood measure for retrieval
    - The ultimate goal is to enhance the estimation of each document language model



$$P(q|d_j) \approx \prod_{i=1}^{|q|} P'(w_i|d_j)$$

$$P'(w_i|d_j) = \alpha \cdot P(w_i|d_j) + \beta \cdot \sum_{k=1}^{K} P(w_i|T_k)P(T_k|d_j) + (1-\alpha-\beta) \cdot P(w_i|BG)$$

Homework 4 - Description..

$$P(q|d_j) \approx \prod_{i=1}^{|q|} P'(w_i|d_j)$$

$$P'(w_i|d_j) = \alpha \cdot P(w_i|d_j) + \beta \cdot \sum_{k=1}^{K} P(w_i|T_k)P(T_k|d_j) + (1-\alpha-\beta) \cdot P(w_i|BG)$$

- The background language model can be estimated by referring to the document collection

$$P(w_i|BG) = \frac{\sum_{d_j \in \mathbf{D}} c(w_i, d_j)}{\sum_{d_k \in \mathbf{D}} |d_k|}$$

- The document unigram model can be obtained in the same manner

$$P(w_i|d_j) = \frac{c(w_i, d_j)}{|d_j|}$$

- The PLSA model is trained on the whole document collection

Homework 4 – Description...

- The evaluation measure is MAP@1000
  - The **hard** deadline is 11/26 23:59
  - You point is depended on your performance on the **private** leaderboard!
    - 50 public queries and 50 private queries

$$YourScore = 5 + \frac{YourMAP - BaselineMAP}{HighestMAP - BaselineMAP} \times 8$$

  - Please submit a **report** and your **source codes** to the Moodle system, otherwise you will get 0 point
    - The report will be judged by TA, and the score is either 1 or 2

Homework 4 – Description....

- You should
  - upload your answer file to kaggle
    - https://www.kaggle.com/t/7de0b487e0d2451a95a2b033fb46c262
    - The maximum number of daily submissions is 20
    - **Your team name is ID_Name**

      M123456_陳冠宇

- Please follow our rules
  - Don't cheat!
  - Don't create multiple accounts!
  - Implement the language model-based IR system!
    - You can only leverage ULM and PLSA to do retrieval
    - Enjoy the language models

Questions?

kychen@mail.ntust.edu.tw