Latent Semantic Analysis & Topic Models

Kuan-Yu Chen (陳冠宇)

2018/10/19 @ TR-514, NTUST

Review

- Recall
- Precision
- Mean Average Precision (MAP)
- Normalized Discounting Cumulated Gain (NDCG)

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**

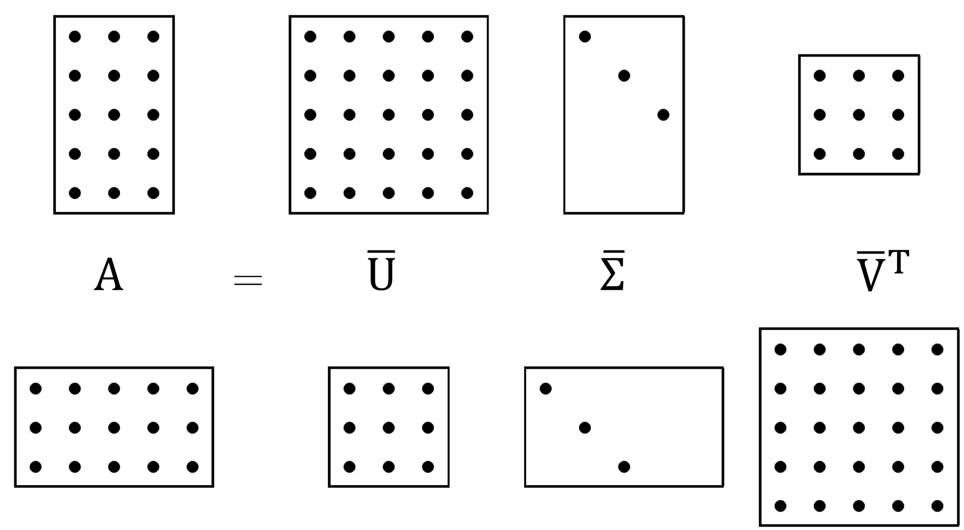
  - Retrieval based on index terms is vague and noisy
    - The user information need is more related to **concepts** and ideas than to **index terms**

Latent Semantic Analysis

Singular Value Decomposition

- In linear algebra, the singular-value decomposition (SVD) is a factorization of a real or complex matrix

- Formally, the SVD of an $m \times n$ matrix A is a factorization of the form $\overline{\mathrm{U}}\overline{\Sigma}\overline{\mathrm{V}}^{\mathrm{T}}$
  – $\overline{\mathrm{U}}$ is an $m \times m$ unitary matrix (i.e.,$\overline{\mathrm{U}}\overline{\mathrm{U}}^{\mathrm{T}} = \mathrm{I} = \overline{\mathrm{U}}^{\mathrm{T}}\overline{\mathrm{U}}$)
  – $\overline{\Sigma}$ is an $m \times n$ rectangular diagonal matrix with non-negative real numbers on the diagonal
  – $\overline{\mathrm{V}}$ is an $n \times n$ unitary matrix



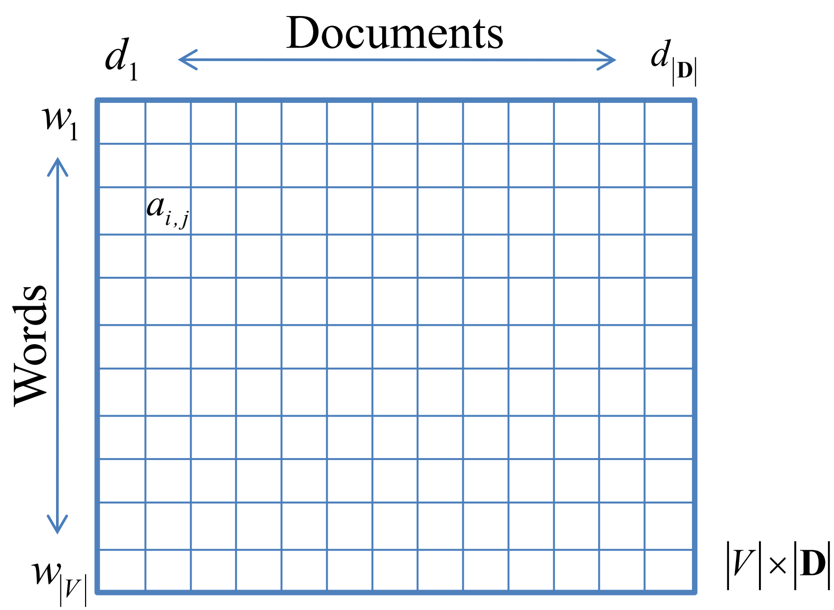
Introduction - LSA

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

- The LSA paradigm operates under the assumption that there is some underlying **latent semantic structure** in the data
  - Algebraic and/or statistical techniques are brought to bear to estimate this latent structure and get rid of the obscuring "noise"

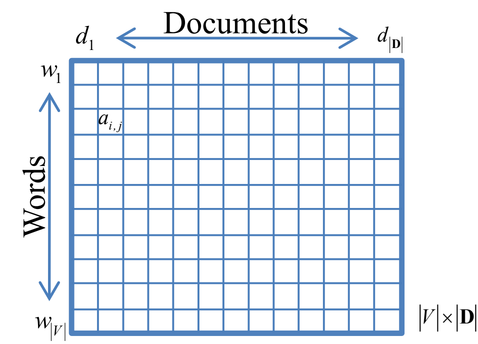
Latent Semantic Analysis – 1

- A given document collection can be represented as a word-document matrix
  - Row: composed of **words** (terms), which are the individual components making up a document
  - Column: composed of **documents** which are of a predetermined size of text such as paragraphs, collections of paragraphs, sentences, etc.



Latent Semantic Analysis – 2.

- In the word-by-document, each element $a_{i,j}$ is represented the importance of word w_i in document d_j
  - The TF-IDF score
  - The Entropy-based method



$$a_{i,j} = (1 - \varepsilon_i) \frac{c(w_i, d_j)}{|d_j|}$$

$$\varepsilon_i = -\frac{1}{log\,|\mathbf{D}|} \sum_{j=1}^{|\mathbf{D}|} \left(\frac{c(w_i, d_j)}{\sum_{j'=1}^{|\mathbf{D}|} c(w_i, d_{j'})} log \frac{c(w_i, d_j)}{\sum_{j'=1}^{|\mathbf{D}|} c(w_i, d_{j'})} \right)$$

- $0 \leq \varepsilon_i \leq 1$
  - $\varepsilon_i = 1 \Rightarrow \forall d_j,\ \ c\left(w_i, d_j\right) = \frac{\sum_{j'=1}^{|\mathbf{D}|} c(w_i, d_{j'})}{|\mathbf{D}|}$: the word distributed across many documents throughout the corpus
  - $\varepsilon_i = 0 \Rightarrow \exists d_j,\ \ c\left(w_i, d_j\right) \approx \sum_{j'=1}^{|\mathbf{D}|} c(w_i, d_{j'})$: the word is present only in a few specific documents

$$a_{i,j} = (1-\varepsilon_i)\frac{c(w_i,d_j)}{|d_j|}$$

Latent Semantic Analysis – 2..

$$\varepsilon_i = -\frac{1}{log\,|\mathbf{D}|}\sum_{j=1}^{|\mathbf{D}|}\left(\frac{c(w_i,d_j)}{\sum_{j'=1}^{|\mathbf{D}|}c(w_i,d_{j'})}\,log\,\frac{c(w_i,d_j)}{\sum_{j'=1}^{|\mathbf{D}|}c(w_i,d_{j'})}\right)$$

- $\varepsilon_i = 1 \Rightarrow \forall d_j,\ \ c\left(w_i,d_j\right) = \frac{\sum_{j'=1}^{|\mathbf{D}|}c(w_i,d_{j'})}{|\mathbf{D}|}$: the word distributed across many documents throughout the corpus

$$\begin{aligned}\varepsilon_i &= -\frac{1}{log|\mathbf{D}|}\sum_{j=1}^{|\mathbf{D}|}\left(\frac{c\left(w_i,d_j\right)}{\sum_{j'=1}^{|\mathbf{D}|}c\left(w_i,d_{j'}\right)}\,log\,\frac{c\left(w_i,d_j\right)}{\sum_{j'=1}^{|\mathbf{D}|}c\left(w_i,d_{j'}\right)}\right)\\ &= -\frac{1}{log|\mathbf{D}|}\sum_{j=1}^{|\mathbf{D}|}\left(\frac{\frac{\sum_{j'=1}^{|\mathbf{D}|}c\left(w_i,d_{j'}\right)}{|\mathbf{D}|}}{\sum_{j'=1}^{|\mathbf{D}|}c\left(w_i,d_{j'}\right)}\,log\,\frac{\frac{\sum_{j'=1}^{|\mathbf{D}|}c\left(w_i,d_{j'}\right)}{|\mathbf{D}|}}{\sum_{j'=1}^{|\mathbf{D}|}c\left(w_i,d_{j'}\right)}\right)\\ &= -\frac{1}{log|\mathbf{D}|}\sum_{j=1}^{|\mathbf{D}|}\left(\frac{1}{|\mathbf{D}|}\,log\,\frac{1}{|\mathbf{D}|}\right) = -\frac{1}{log|\mathbf{D}|}\left(log\,\frac{1}{|\mathbf{D}|}\right) = -\frac{1}{log|\mathbf{D}|}(-\,log|\mathbf{D}|) = 1\end{aligned}$$

$$a_{i,j} = (1-\varepsilon_i)\frac{c(w_i,d_j)}{|d_j|}$$

Latent Semantic Analysis – 2…

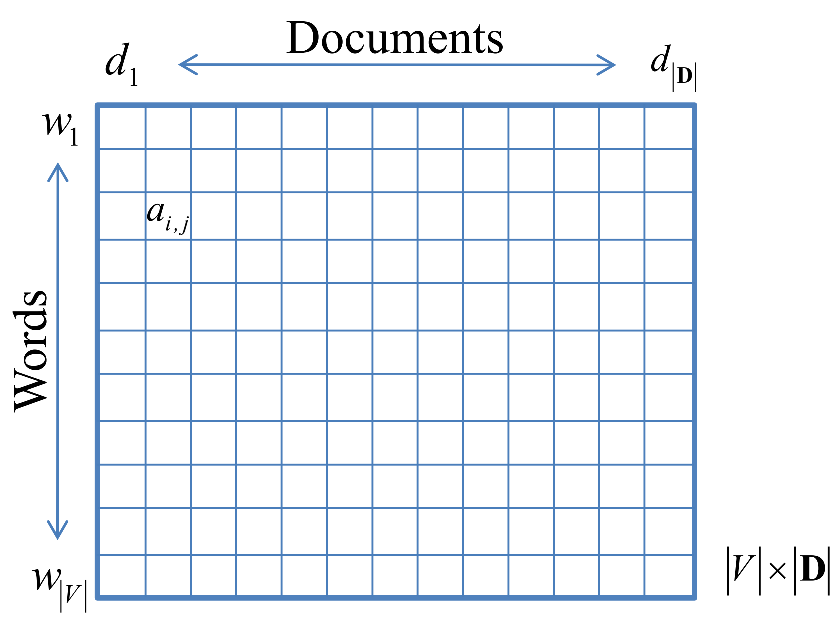
$$\varepsilon_i = -\frac{1}{log\,|\mathbf{D}|}\sum_{j=1}^{|\mathbf{D}|}\left(\frac{c(w_i,d_j)}{\sum_{j'=1}^{|\mathbf{D}|}c(w_i,d_{j'})}\,log\,\frac{c(w_i,d_j)}{\sum_{j'=1}^{|\mathbf{D}|}c(w_i,d_{j'})}\right)$$

- $\varepsilon_i = 0 \Rightarrow \exists d_j,\ \ c(w_i,d_j) \approx \sum_{j'=1}^{|\mathbf{D}|}c(w_i,d_{j'})$: the word is present only in a few specific documents

$$\begin{aligned}
\varepsilon_i &= -\frac{1}{\log|\mathbf{D}|}\sum_{j=1}^{|\mathbf{D}|}\left(\frac{c(w_i,d_j)}{\sum_{j'=1}^{|\mathbf{D}|}c(w_i,d_{j'})}\,log\,\frac{c(w_i,d_j)}{\sum_{j'=1}^{|\mathbf{D}|}c(w_i,d_{j'})}\right)\\
&= -\frac{1}{\log|\mathbf{D}|}\times(|\mathbf{D}|-1)\times\left(\frac{0}{\sum_{j'=1}^{|\mathbf{D}|}c(w_i,d_{j'})}\,log\,\frac{0}{\sum_{j'=1}^{|\mathbf{D}|}c(w_i,d_{j'})}\right)\\
&\quad -\frac{1}{\log|\mathbf{D}|}\times\left(\frac{\sum_{j'=1}^{|\mathbf{D}|}c(w_i,d_{j'})}{\sum_{j'=1}^{|\mathbf{D}|}c(w_i,d_{j'})}\,log\,\frac{\sum_{j'=1}^{|\mathbf{D}|}c(w_i,d_{j'})}{\sum_{j'=1}^{|\mathbf{D}|}c(w_i,d_{j'})}\right)\\
&= 0
\end{aligned}$$

Latent Semantic Analysis – 3

- For the word-by-document matrix, it should be noted that
  - the dimensions and can be extremely large
  - the vectors and are typically very sparse
  - the two spaces (for words and documents) are distinct from one other

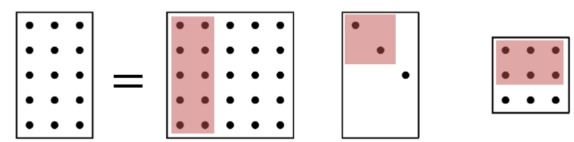


Latent Semantic Analysis – 4

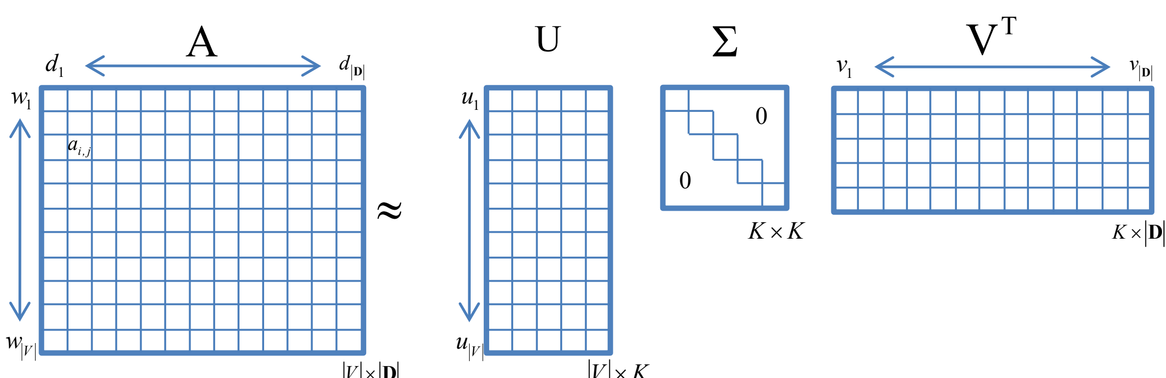
- In order to explore the latent semantic space, to project word and document vectors in the space, and to reduce the size of the vectors, the **Singular Value Decomposition** (SVD) can be employed
  - $K \leq \min(|V|, |\mathbf{D}|)$: **low-rank approximation**

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

  - The objective function is



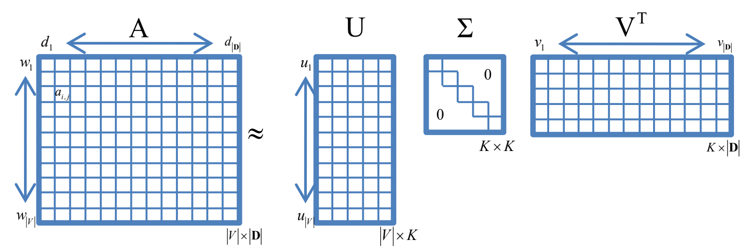
$$min\|\mathrm{A} - \mathrm{A}'\|_F^2, \text{for a given } K$$



$$\|\mathrm{B}\|_F^2 = \sum_i \sum_j b_{i,j}^2$$

Latent Semantic Analysis – 5

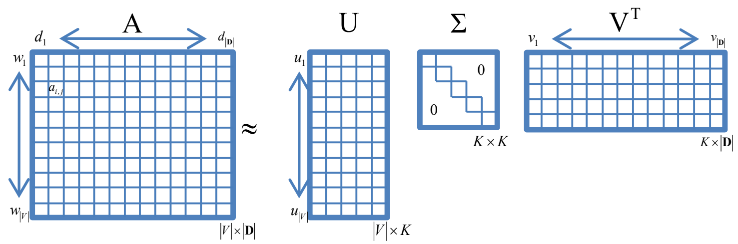
- Properties of SVD decomposition
  - Both left and right singular matrices and are column-orthonormal
    - $U^T U = V^T V = I$
  - Values (nonnegative real numbers) in diagonal matrix are square roots of the eigenvalues of $A^T A$
    - $\Sigma^2 = \mathrm{diag}\{\lambda_1, \lambda_2, \dots, \lambda_K\}$
    - $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K \geq 0$



  - The column vectors of U define an orthonormal basis for d_j
    - $A \approx U\Sigma V^T \Rightarrow A^T U \approx \left(U\Sigma V^T\right)^T U = V\Sigma U^T U = V\Sigma \Rightarrow U^T A = \Sigma V^T$
  - The column vectors of V define an orthonormal basis for w_i
    - $A \approx U\Sigma V^T \Rightarrow AV \approx \left(U\Sigma V^T\right) V = U\Sigma V^T V = U\Sigma \Rightarrow V^T A^T = \Sigma U^T$

Latent Semantic Analysis – 6

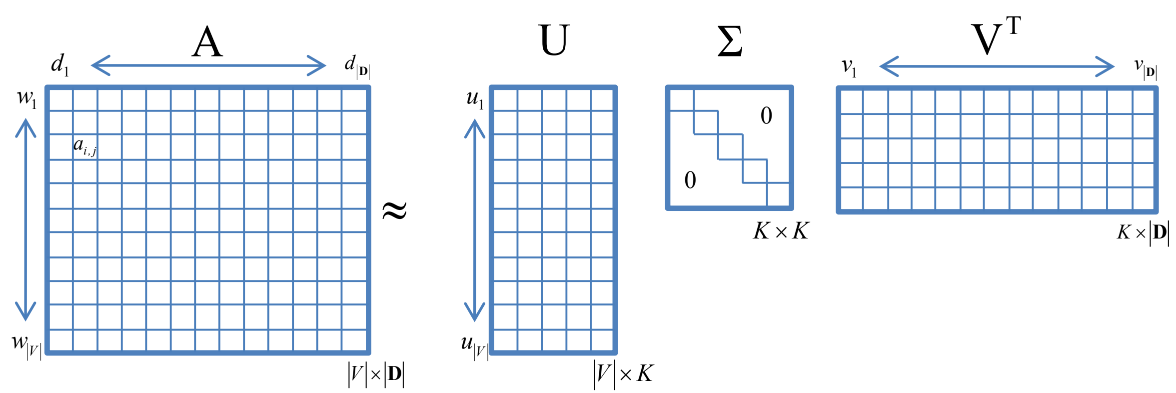
- New representations
  - For each words, the new vector representation is Σu_i^{T}
  - For each document, the new vector representation is Σv_i^{T}



- While the original high-dimensional vectors are sparse, the corresponding low-dimensional latent vectors will typically not be sparse
  - It is possible to compute meaningful association values between pairs of documents, even if the documents do not have any terms in common
    - The hope is that terms having a common meaning (synonyms), are roughly mapped to the same direction in the latent space

Latent Semantic Analysis – 7

- By using the decomposition
  - Compare two words
    - $A \approx U\Sigma V^T \Rightarrow AA^T \approx U\Sigma V^T(U\Sigma V^T)^T = U\Sigma V^T V\Sigma^T U^T = U\Sigma(U\Sigma)^T$
  - Compare two documents
    - $A \approx U\Sigma V^T \Rightarrow A^T A \approx (U\Sigma V^T)^T U\Sigma V^T = V\Sigma^T U^T U\Sigma V^T = V\Sigma(V\Sigma)^T$
  - Compare words and documents
    - $A \approx U\Sigma V^T$



Latent Semantic Analysis – 8

- For a given query (as a document), a low-dimensional representation should be inferred
  - The low-dimensional representation can be obtained by using the **fold-in** strategy

$$(\vec{q}'^{\mathrm{T}})_{1\times K} = \underbrace{(\vec{q}^{\mathrm{T}})_{1\times|V|}\mathrm{U}_{|V|\times K}}_{\text{Weighted sum of the word vectors}}\underbrace{\Sigma^{-1}_{K\times K}}_{\text{Each dimension has its own weight}}$$

$\mathrm{B} = \mathrm{U}\Sigma\mathrm{V}^{\mathrm{T}}$
$\Rightarrow \mathrm{B}^{\mathrm{T}} = (\mathrm{U}\Sigma\mathrm{V}^{\mathrm{T}})^{\mathrm{T}} = \mathrm{V}\Sigma^{\mathrm{T}}\mathrm{U}^{\mathrm{T}}$
$\Rightarrow \mathrm{B}^{\mathrm{T}}\mathrm{U} = \mathrm{V}\Sigma^{\mathrm{T}}$
$\Rightarrow \mathrm{B}^{\mathrm{T}}\mathrm{U}\Sigma^{-1} = \mathrm{V}\Sigma\Sigma^{-1} = \mathrm{V}$

  - For a new document, the representation can also be derived by the fold-in strategy
  - Consequently, the relevance degree can be computed:

$$sim(q, d_j) = cos(\Sigma\vec{q}', \Sigma\vec{d}_j') = \frac{\Sigma\vec{q}' \cdot \Sigma\vec{d}_j'}{|\Sigma\vec{q}||\Sigma\vec{d}_j'|}$$

$$sim(q, d_j) = cos(\vec{q}', \vec{d}_j')$$

Example – 1.

$A =$

	d_1	d_2	d_3	d_4	d_5	d_6
ship	1	0	1	0	0	0
boat	0	1	0	0	0	0
ocean	1	1	0	0	0	0
voyage	1	0	0	1	1	0
trip	0	0	0	1	0	1

$U =$

	1	2
ship	−0.44	−0.30
boat	−0.13	−0.33
ocean	−0.48	−0.51
voyage	−0.70	0.35
trip	−0.26	0.65

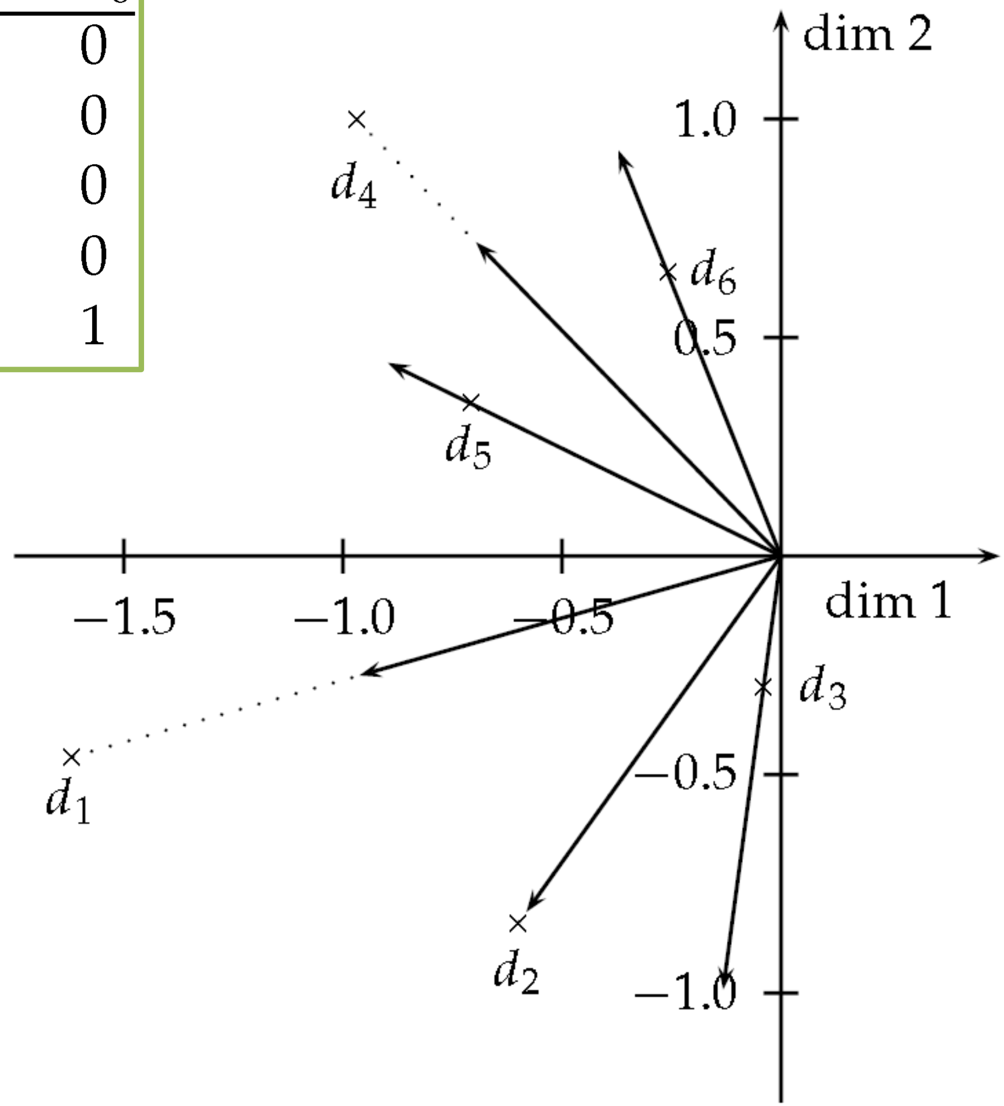
$$\Sigma = \begin{matrix} 2.16 & 0.00 \\ 0.00 & 1.59 \end{matrix}$$

$V^T =$

	d_1	d_2	d_3	d_4	d_5	d_6
1	−1.62	−0.60	−0.44	−0.97	−0.70	−0.26
2	−0.46	−0.84	−0.30	1.00	0.35	0.65

Example – 1..

	d_1	d_2	d_3	d_4	d_5	d_6
ship	1	0	1	0	0	0
boat	0	1	0	0	0	0
ocean	1	1	0	0	0	0
voyage	1	0	0	1	1	0
trip	0	0	0	1	0	1



Example – 2.

c1: *Human* machine *interface* for Lab ABC *computer* applications
c2: A *survey* of *user* opinion of *computer system response time*
c3: The *EPS user interface* management *system*
c4: *System* and *human system* engineering testing of *EPS*
c5: Relation of *user*-perceived *response time* to error measurement

m1: The generation of random, binary, unordered *trees*
m2: The intersection *graph* of paths in *trees*
m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4: *Graph minors*: A *survey*

	Terms	**Documents**								
		c1	c2	c3	c4	c5	m1	m2	m3	m4
1	*human*	1	0	0	1	0	0	0	0	0
2	*interface*	1	0	1	0	0	0	0	0	0
3	*computer*	1	1	0	0	0	0	0	0	0
4	*user*	0	1	1	0	1	0	0	0	0
5	*system*	0	1	1	2	0	0	0	0	0
6	*response*	0	1	0	0	1	0	0	0	0
7	*time*	0	1	0	0	1	0	0	0	0
8	*EPS*	0	0	1	1	0	0	0	0	0
9	*survey*	0	1	0	0	0	0	0	0	1
10	*trees*	0	0	0	0	0	1	1	1	0
11	*graph*	0	0	0	0	0	0	1	1	1
12	*minors*	0	0	0	0	0	0	0	1	1

Example – 2..

Query="human computer interaction"

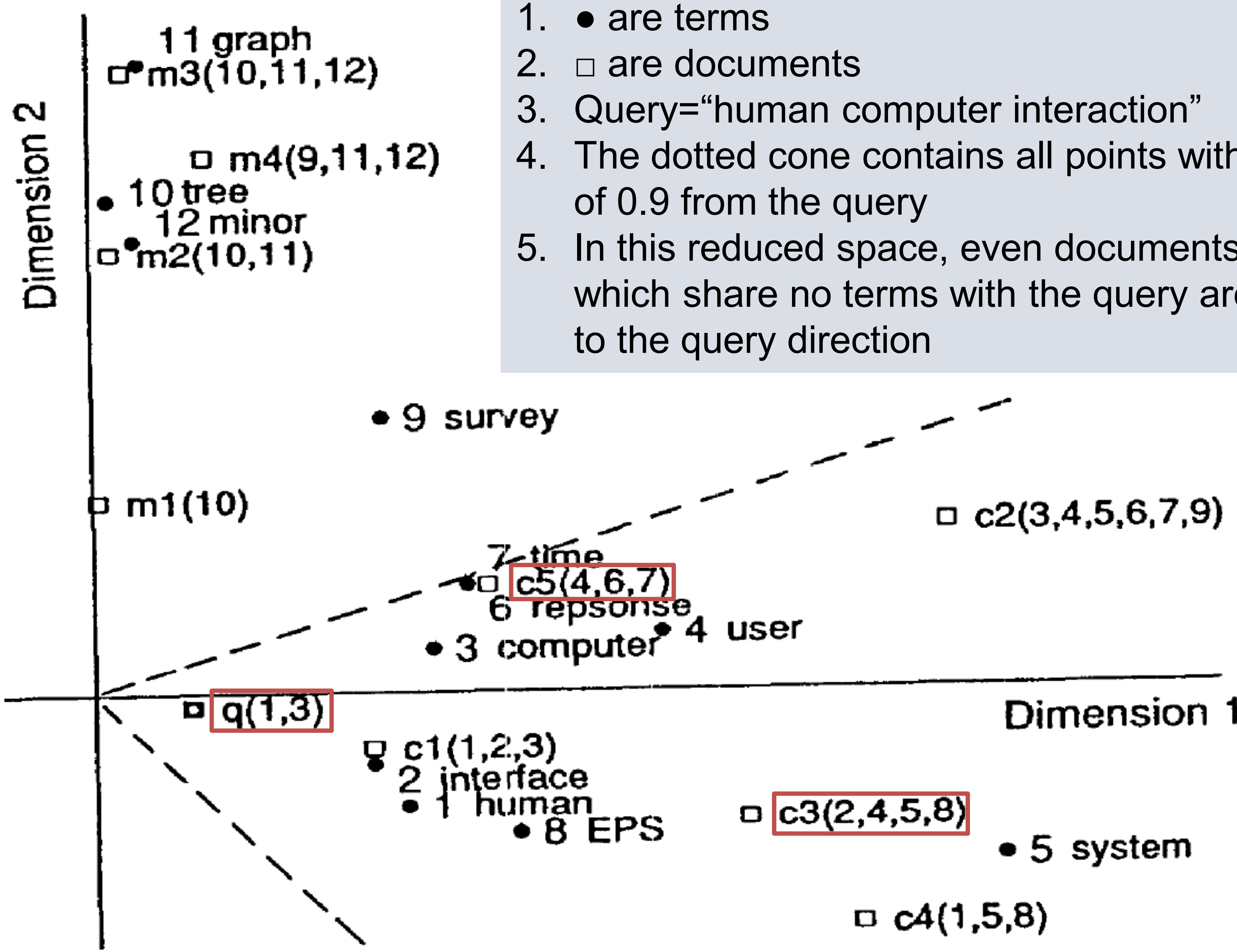
c1: *Human* machine *interface* for Lab ABC *computer* applications
c2: A *survey* of *user* opinion of *computer system response time*
c3: The *EPS user interface* management *system*
c4: *System* and *human system* engineering testing of *EPS*
c5: Relation of *user*-perceived *response time* to error measurement

m1: The generation of random, binary, unordered *trees*
m2: The intersection *graph* of paths in *trees*
m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4: *Graph minors*: A *survey*

	Terms	Documents								
		c1	c2	c3	c4	c5	m1	m2	m3	m4
1	*human*	1	0	0	1	0	0	0	0	0
2	*interface*	1	0	1	0	0	0	0	0	0
3	*computer*	1	1	0	0	0	0	0	0	0
4	*user*	0	1	1	0	1	0	0	0	0
5	*system*	0	1	1	2	0	0	0	0	0
6	*response*	0	1	0	0	1	0	0	0	0
7	*time*	0	1	0	0	1	0	0	0	0
8	*EPS*	0	0	1	1	0	0	0	0	0
9	*survey*	0	1	0	0	0	0	0	0	1
10	*trees*	0	0	0	0	0	1	1	1	0
11	*graph*	0	0	0	0	0	0	1	1	1
12	*minors*	0	0	0	0	0	0	0	1	1

Example – 2..

1. ● are terms
2. □ are documents
3. Query="human computer interaction"
4. The dotted cone contains all points within a cosine of 0.9 from the query
5. In this reduced space, even documents c3 and c5, which share no terms with the query are very close to the query direction



Pros and Cons

- Advantages
  - As we reduce K, **recall tends to increase**, as expected
  - Most surprisingly, a value of K in the low hundreds can actually **increase precision** on some query benchmarks
  - Finding a new space for words and documents

- Disadvantages
  - The computational cost of the SVD is significant
  - Irrelevant or Antonymous
  - The reconstruction has negative entities

LSA-based Language Modeling – 1

- A goal of statistical language modeling is to learn the joint probability function of sequences of words in a language
  - By using n-gram model

$$P(w_1, w_2, \dots, w_T) \approx \prod_{t=1}^{T} P(w_t | w_{t-n+1}, \dots, w_{t-1})$$

  - By incorporating n-gram model and LSA-based model

$$P(w_1, w_2, \dots, w_T) \approx \prod_{t=1}^{T} P\left(w_t | H_{t-1}^{n,l}\right) = \prod_{t=1}^{T} P\left(w_t | \underbrace{H_{t-1}^{n}}_{\text{Lexical Information}}, \underbrace{H_{t-1}^{l}}_{\text{Semantic Information}}\right)$$

LSA-based Language Modeling – 2

- The probability can further be decomposed:

$$P\left(w_t|H_{t-1}^n, H_{t-1}^l\right) = \frac{P\left(w_t, H_{t-1}^l|H_{t-1}^n\right)}{\sum_{w_i \in V} P\left(w_i, H_{t-1}^l|H_{t-1}^n\right)}$$

- Expanding and rearranging, the numerator is seen to be:

$$\begin{aligned}
P\left(w_t, H_{t-1}^l|H_{t-1}^n\right) &= \frac{P\left(w_t, H_{t-1}^l, H_{t-1}^n\right)}{P(H_{t-1}^n)} \\
&= \frac{P\left(w_t, H_{t-1}^l, H_{t-1}^n\right)P(w_t, H_{t-1}^n)}{P(H_{t-1}^n)P(w_t, H_{t-1}^n)} \\
&= P(w_t|H_{t-1}^n)P\left(H_{t-1}^l|w_t, H_{t-1}^n\right) \\
&= P(w_t|w_{t-n+1}, \dots, w_{t-1})P\left(H_{t-1}^l\middle|w_{t-n+1}, \dots, w_{t-1}, w_t\right) \\
&= P(w_t|w_{t-n+1}, \dots, w_{t-1})P\left(H_{t-1}^l\middle|w_t\right)
\end{aligned}$$

We assume the probability of the document history given the current word is not affected by other context words

LSA-based Language Modeling – 3

- Consequently, we can obtain:

$$P\left(w_t|H_{t-1}^{n,l}\right) = P\left(w_t|H_{t-1}^{n}, H_{t-1}^{l}\right) = \frac{P(w_t|w_{t-n+1}, \dots, w_{t-1})P\left(H_{t-1}^{l}|w_t\right)}{\sum_{w_i \in V} P(w_i|w_{t-n+1}, \dots, w_{t-1})P\left(H_{t-1}^{l}|w_i\right)}$$

$$= \frac{P(w_t|w_{t-n+1}, \dots, w_{t-1})\frac{P(w_t|H_{t-1}^{l})}{P(w_t)}}{\sum_{w_i \in V} P(w_i|w_{t-n+1}, \dots, w_{t-1})\frac{P(w_i|H_{t-1}^{l})}{P(w_i)}}$$

- H_{t-1}^{l} can be represented by a vector in the semantic space

$$\left(\overrightarrow{H_{t-1}^{l}}'^{\mathrm{T}}\right)_{1\times K} = \left(\overrightarrow{H_{t-1}^{l}}^{\mathrm{T}}\right)_{1\times|V|} \mathrm{U}_{|V|\times K}\Sigma_{K\times K}^{-1}$$

- Thus, the semantic smoothing factor can be estimated by:

$$P(w_t|H_{t-1}^{l}) \propto \cos(\Sigma^{\frac{1}{2}}\overrightarrow{H_{t-1}^{l}}', \Sigma^{\frac{1}{2}}u_{w_t}^{\mathrm{T}})$$

LSA-based Language Modeling – Appendix

- By using the entropy-based method to score each element in the vector, a fast strategy can be derived for sequential data

$$\overrightarrow{H_t^l} = \frac{|H_t^l| - 1}{|H_t^l|} \overrightarrow{H_{t-1}^l} + \frac{1 - \varepsilon_{w_t}}{|H_t^l|} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$a_{i,j} = (1 - \varepsilon_i) \frac{c(w_i, d_j)}{|d_j|}$$

Statistical Topic Models

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

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

- PLSA is a probabilistic counterpart of LSA
  - $P(d_j)$: 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)

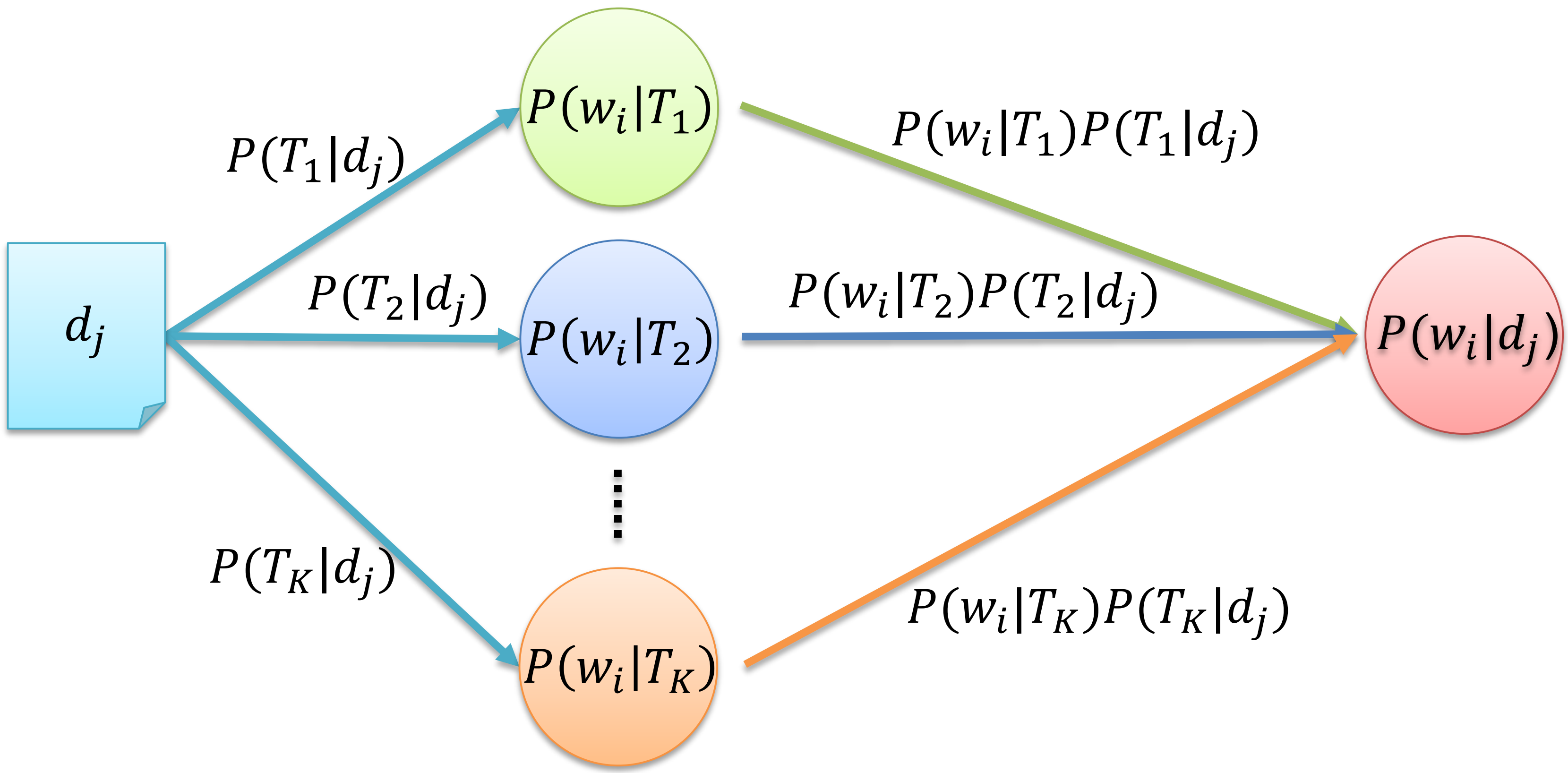
$$P(w_i, d_j) = P(d_j)P(w_i|d_j) = P(d_j)\sum_{k=1}^{K} P(w_i|T_k)P(T_k|d_j)$$

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

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

PLSA – 2

- Thus, the modeling goal is to identify conditional probability mass functions $P(w_i|T_k)$ such that 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 $P(T_k|d_j)$

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

PLSA – 4

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

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

  - M-step

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

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

PLSA – 5

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

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

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

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

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

PLSA – 7

- In addition to the query likelihood measure, we can combine PLSA with the vector-space model
  - 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

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

Link PLSA and LSA

- Another derivation of PLSA model

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

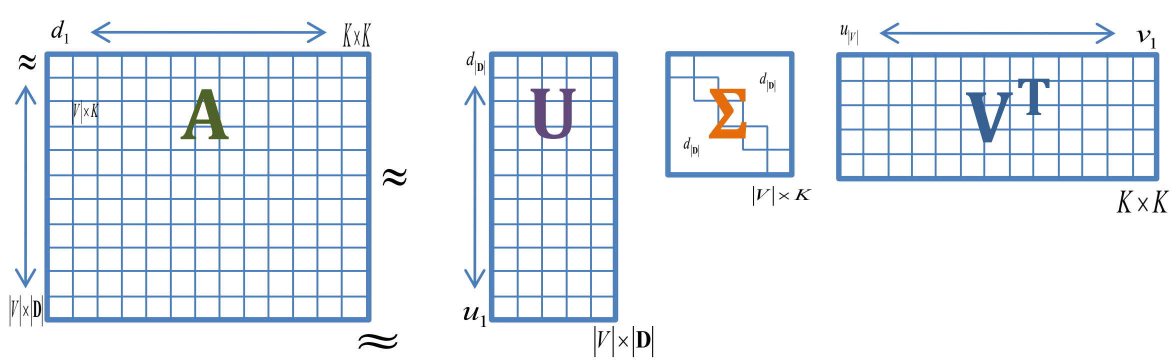
$$= \sum_{k=1}^{K} P(w_i | d_j, T_k) P(d_j, T_k)$$

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

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

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

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



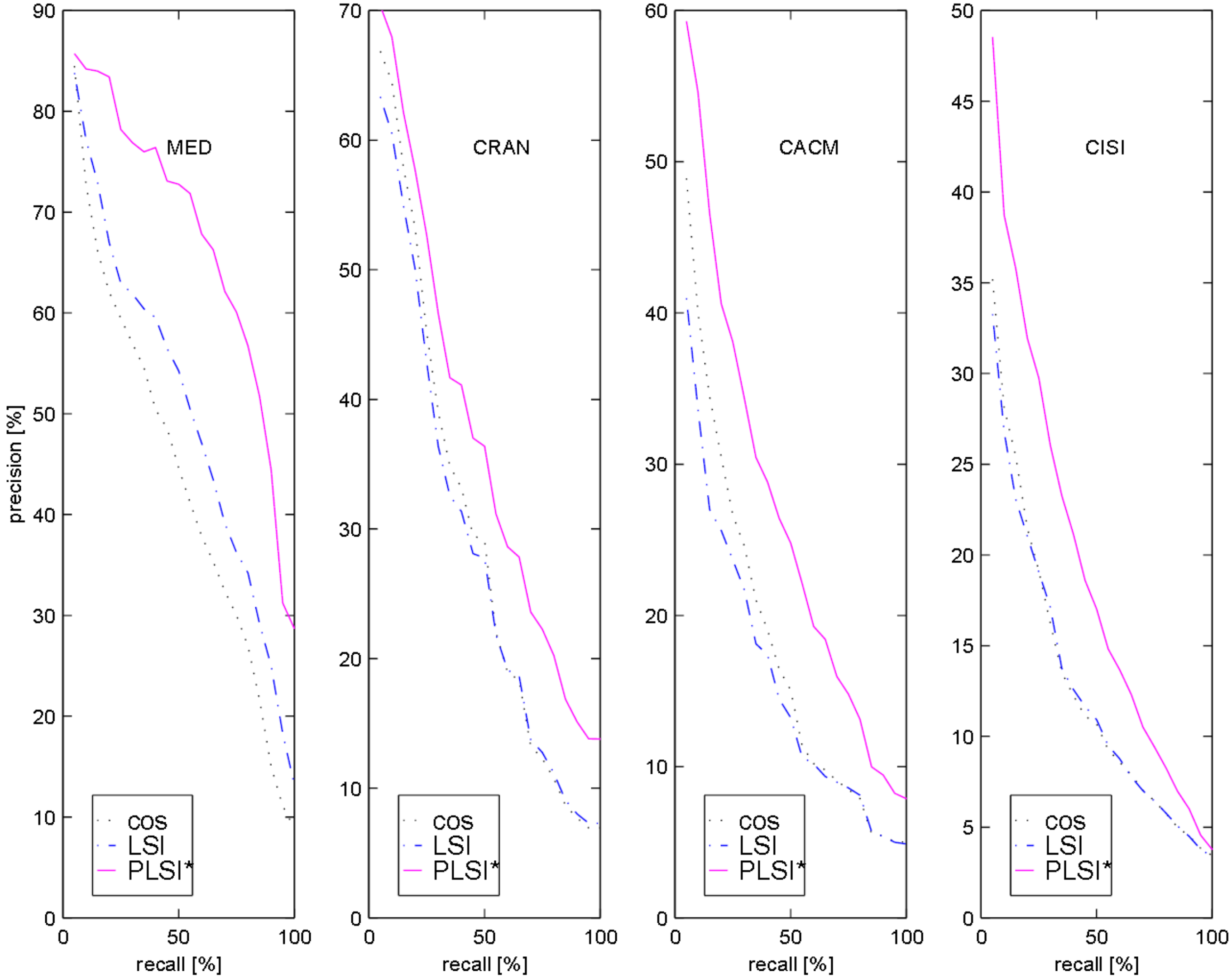
Comparisons – PLSA & LSA

- 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 KL-divergence measure

- 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

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 ([illegible]

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

Revisiting the Objective Function

$$KL(T||E) = \sum\nolimits_{x \in \mathbf{X}} T(x) log \frac{T(x)}{E(x)}$$

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

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

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

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

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

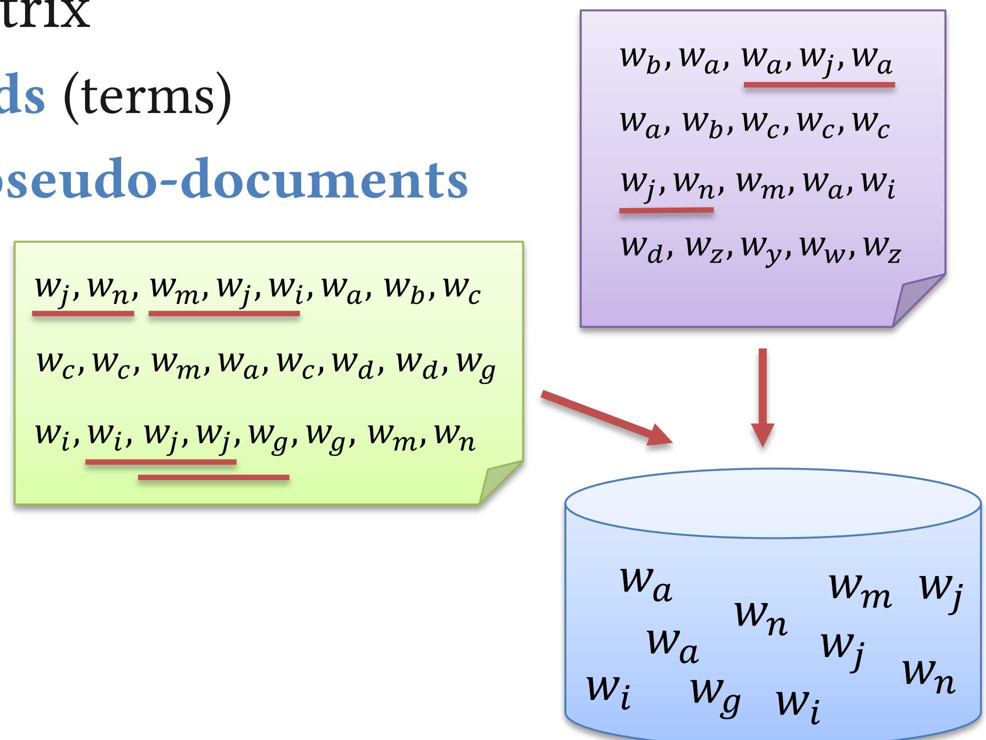
$$= \sum_{d_j \in \mathbf{D}} |d_j| \sum_{w_i \in V} \Big(\underbrace{P(w_i|d_j) log P(d_j)}_{\text{Constant}} + \underbrace{P(w_i|d_j) log P_{PLSA}(w_i|d_j)}_{\text{KL-Divergence}} \Big)$$

Word Topic Modeling – 1

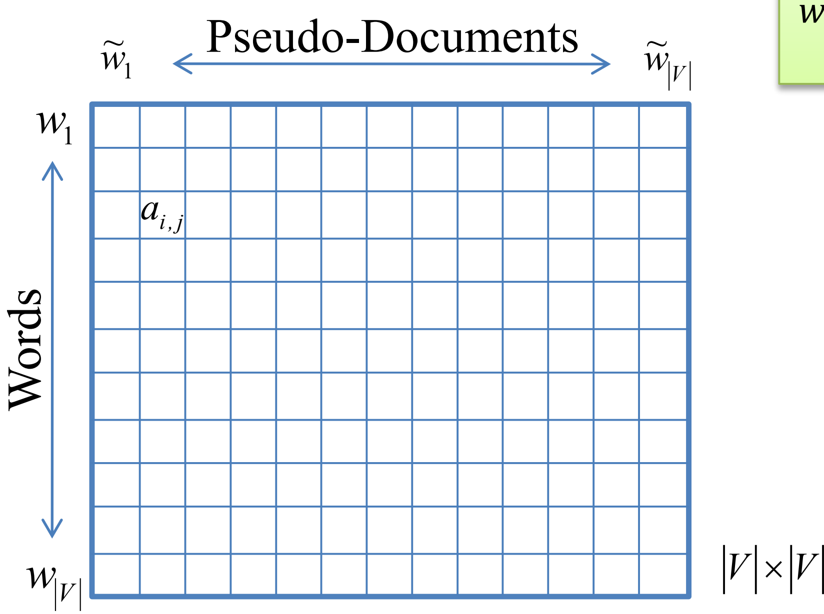
- WTM is a novel extension of the PLSA model
  - The model parameters of PLSA grow linearly with the size of the corpus
    - For WTM, the number of parameters is fixed
  - PLSA explores the co-occurrence relationship between words and documents
    - WTM focuses on the word-word co-occurrence relationship
  - For a new document, fold-in strategy is time consuming
    - For WTM, a linear combination technique can be applied

Word Topic Modeling – 2

- A given document collection can be represented as a word-by-pseudo-document matrix
  - Row: composed of **words** (terms)
  - Column: composed of **pseudo-documents**



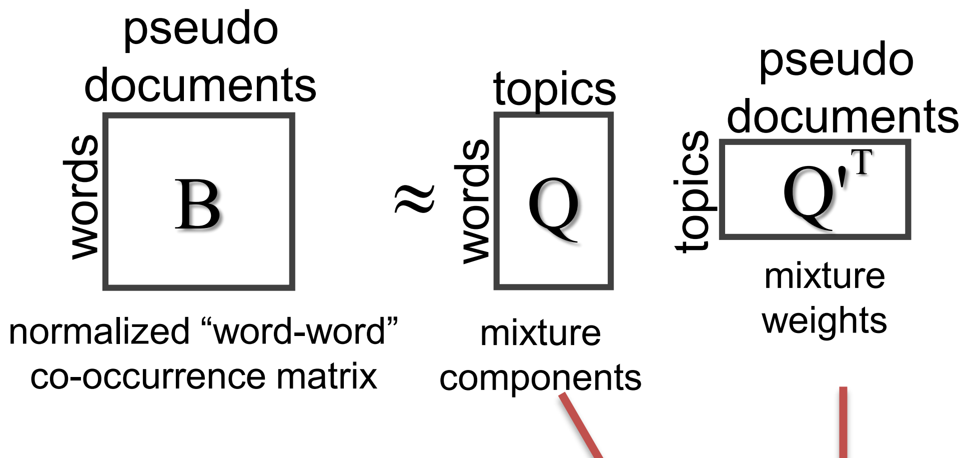
pseudo-document $\tilde{w}_j$



- The window size is set to 3
- $a_{i,j} = 2 \quad a_{m,j} = 1$

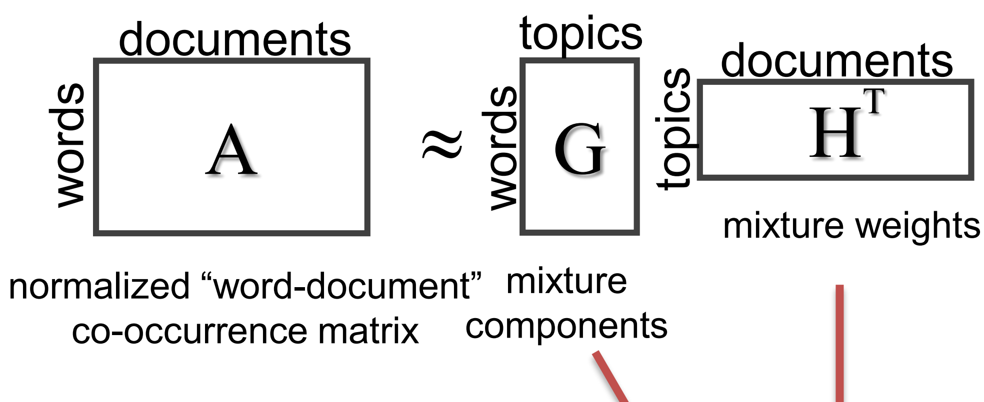
Word Topic Modeling – 3

- WTM



$$P_{\mathrm{WTM}}\left(w_i \middle| \widetilde{w}_j\right) = \sum_{k=1}^{K} P(w_i|T_k)P\left(T_k \middle| \widetilde{w}_j\right)$$

- PLSA



$$P_{\mathrm{PLSA}}\left(w_i \middle| d_j\right) = \sum_{k=1}^{K} P(w_i|T_k)P\left(T_k \middle| d_j\right)$$

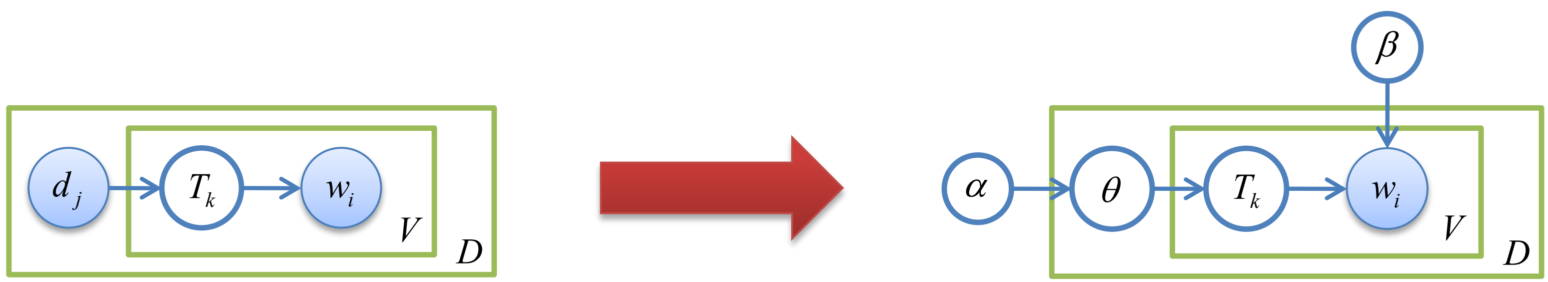
Word Topic Modeling – 4

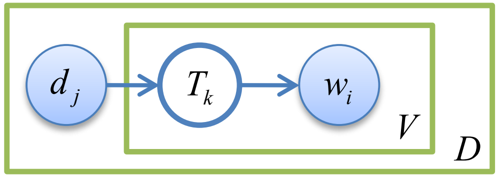
- For a new document, we can linearly combine the associated WTM models of the words occurring in the document to form a composite WTM model

$$
\begin{aligned}
P_{\mathrm{WTM}}\left(w_i \middle| d_j\right) &= \frac{1}{|d_j|} \sum_{j'=1}^{|d_j|} \sum_{k=1}^{K} P(w_i|T_k) P\left(T_k \middle| \widetilde{w}_{j'}\right) \\
&= \sum_{k=1}^{K} P(w_i|T_k) \sum_{j'=1}^{|d_j|} \frac{P\left(T_k \middle| \widetilde{w}_{j'}\right)}{|d_j|} \\
&= \sum_{k=1}^{K} P(w_i|T_k) \hat{P}\left(T_k \middle| d_j\right)
\end{aligned}
$$

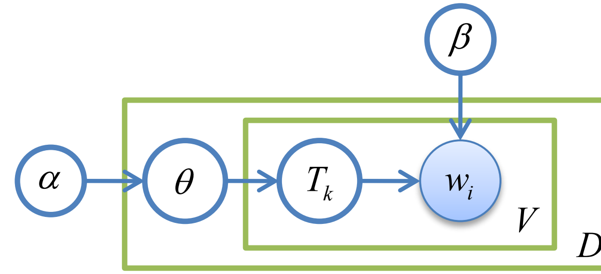
From PLSA to Latent Dirichelet Allocation

- 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

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

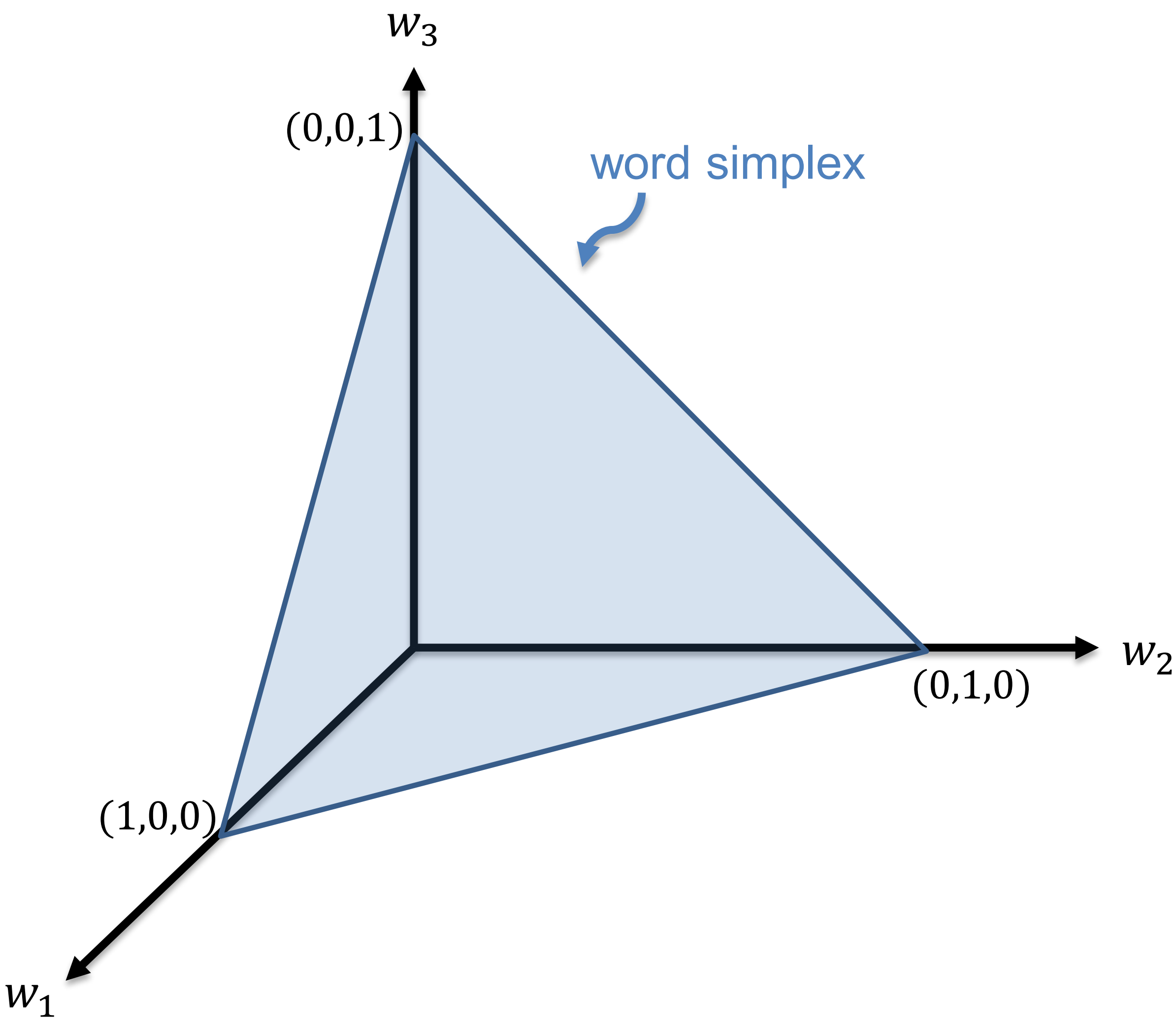
$$= \prod_{d_j \in \mathbf{D}} \prod_{i=1}^{|d_j|} \left(P(d_j) \sum_{k=1}^{K} P(w_i|T_k) P(T_k|d_j) \right)$$

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

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

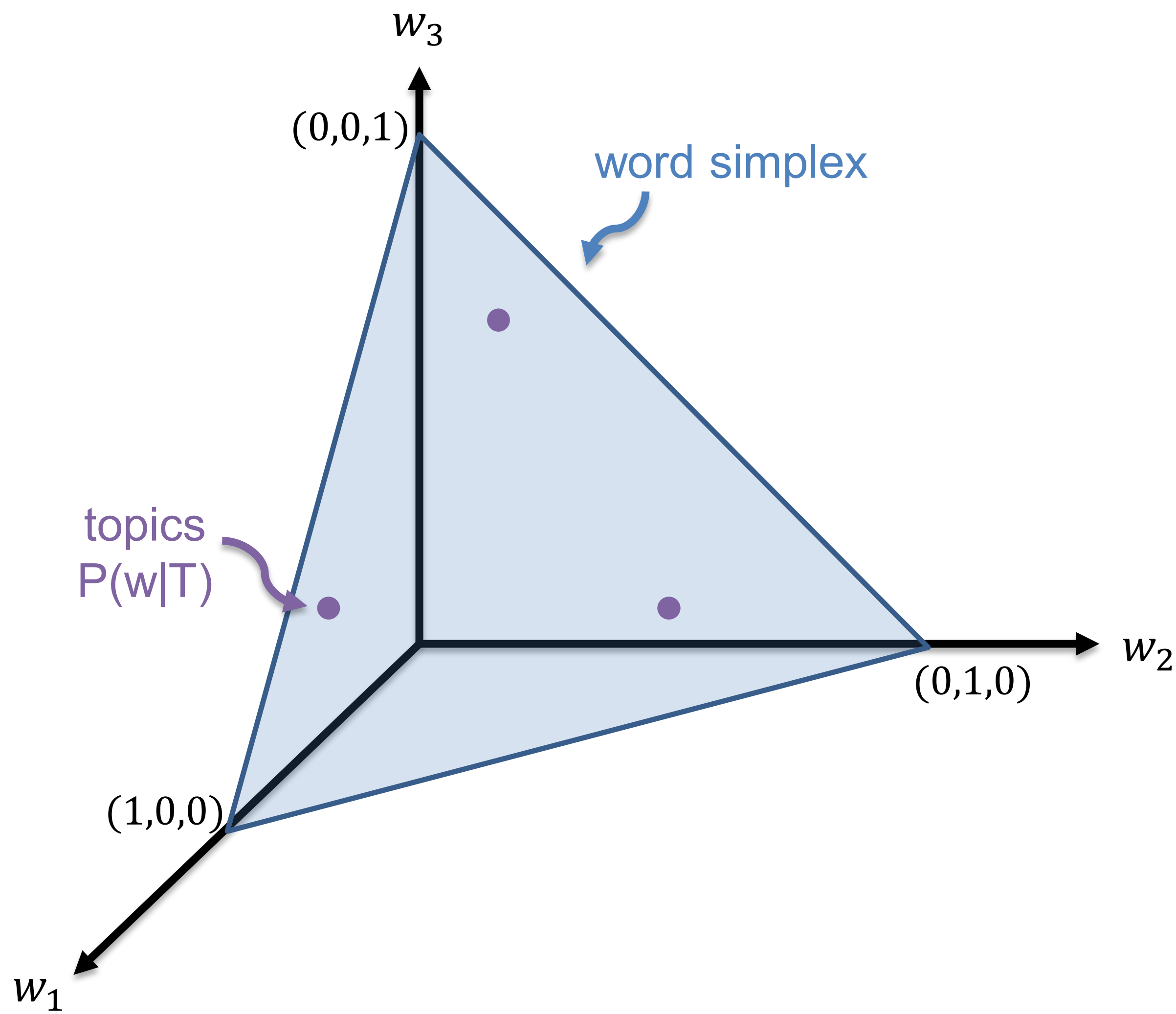
PLSA & LDA – 2

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



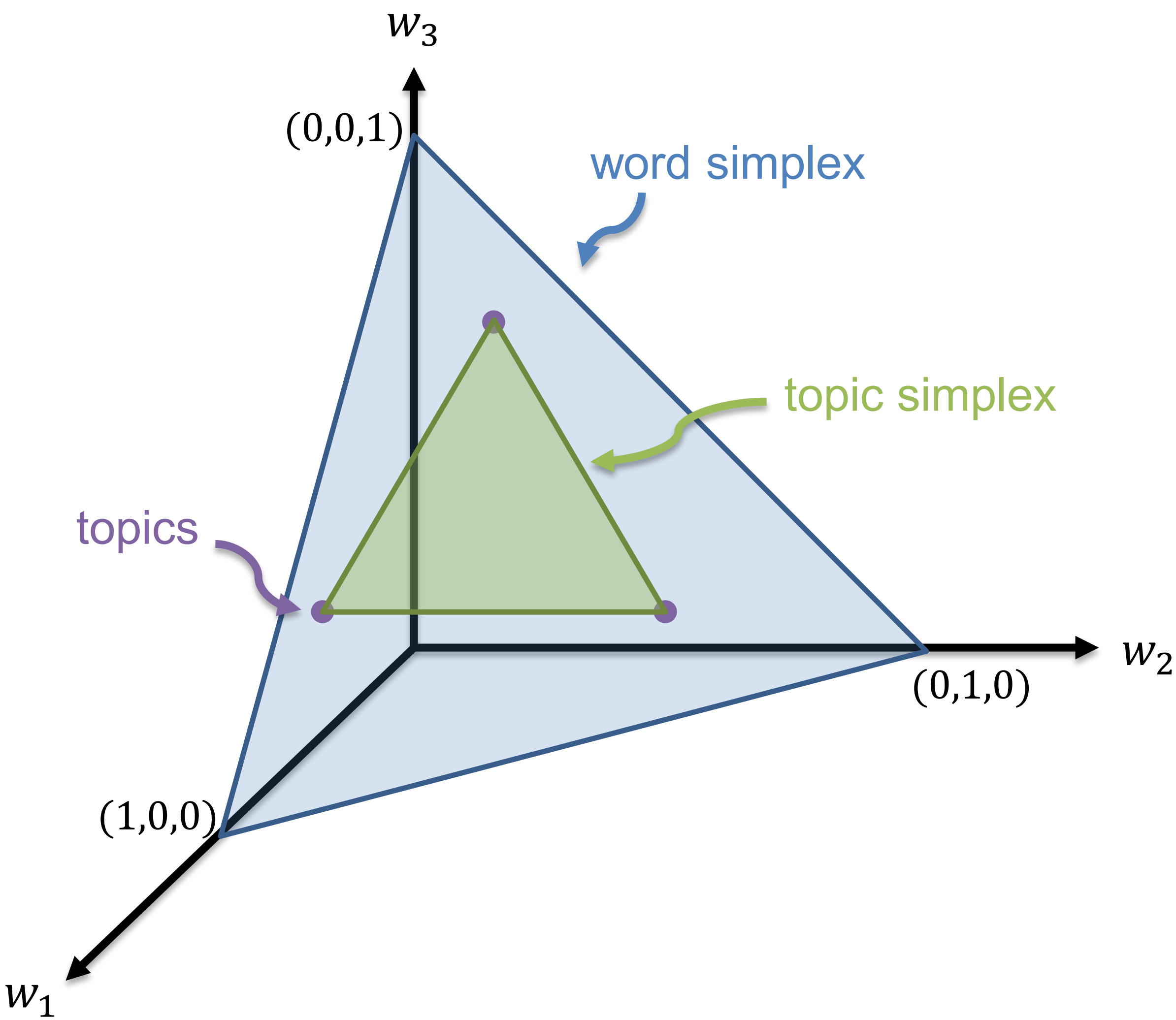
PLSA & LDA – 3

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



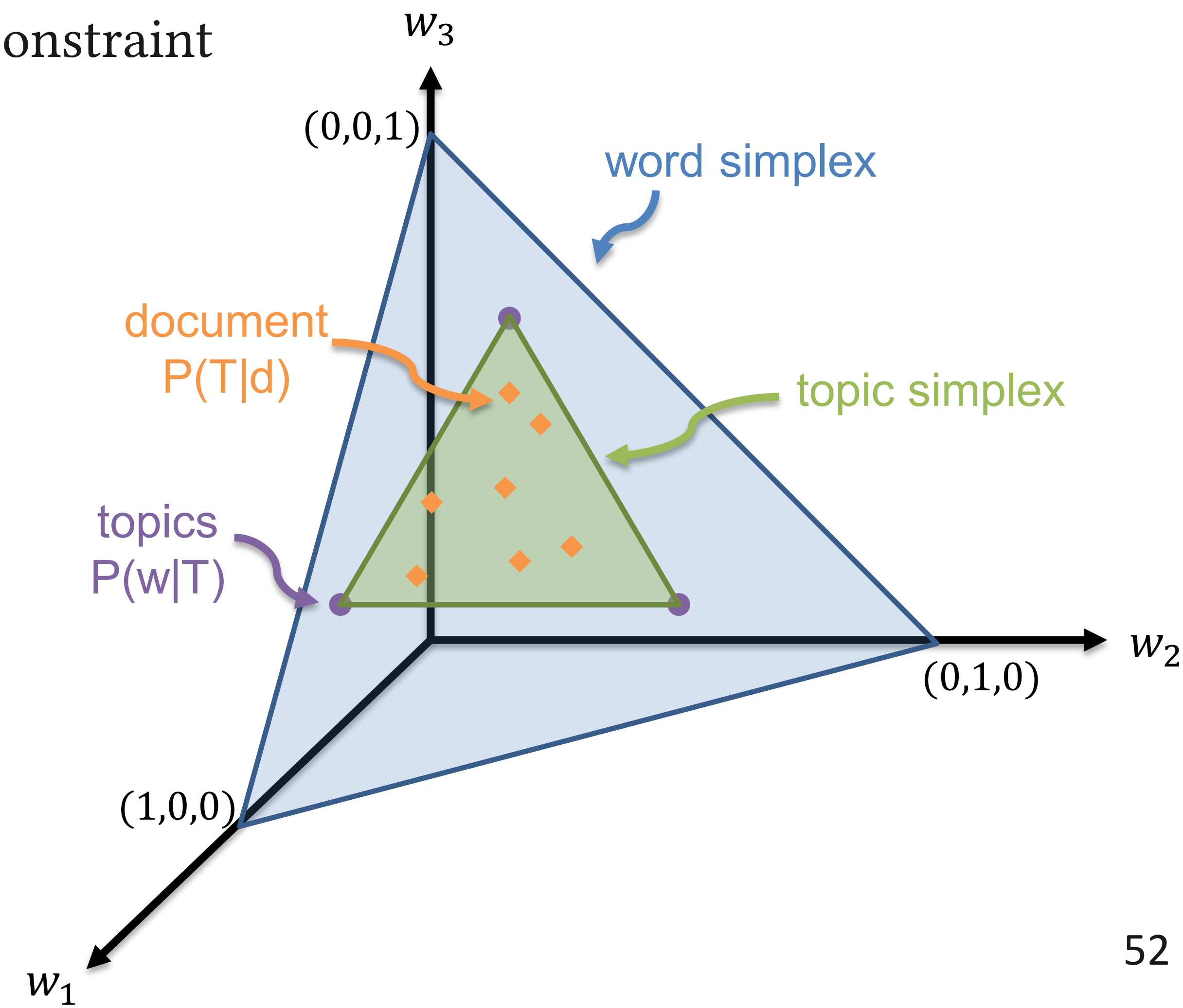
PLSA & LDA – 4

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



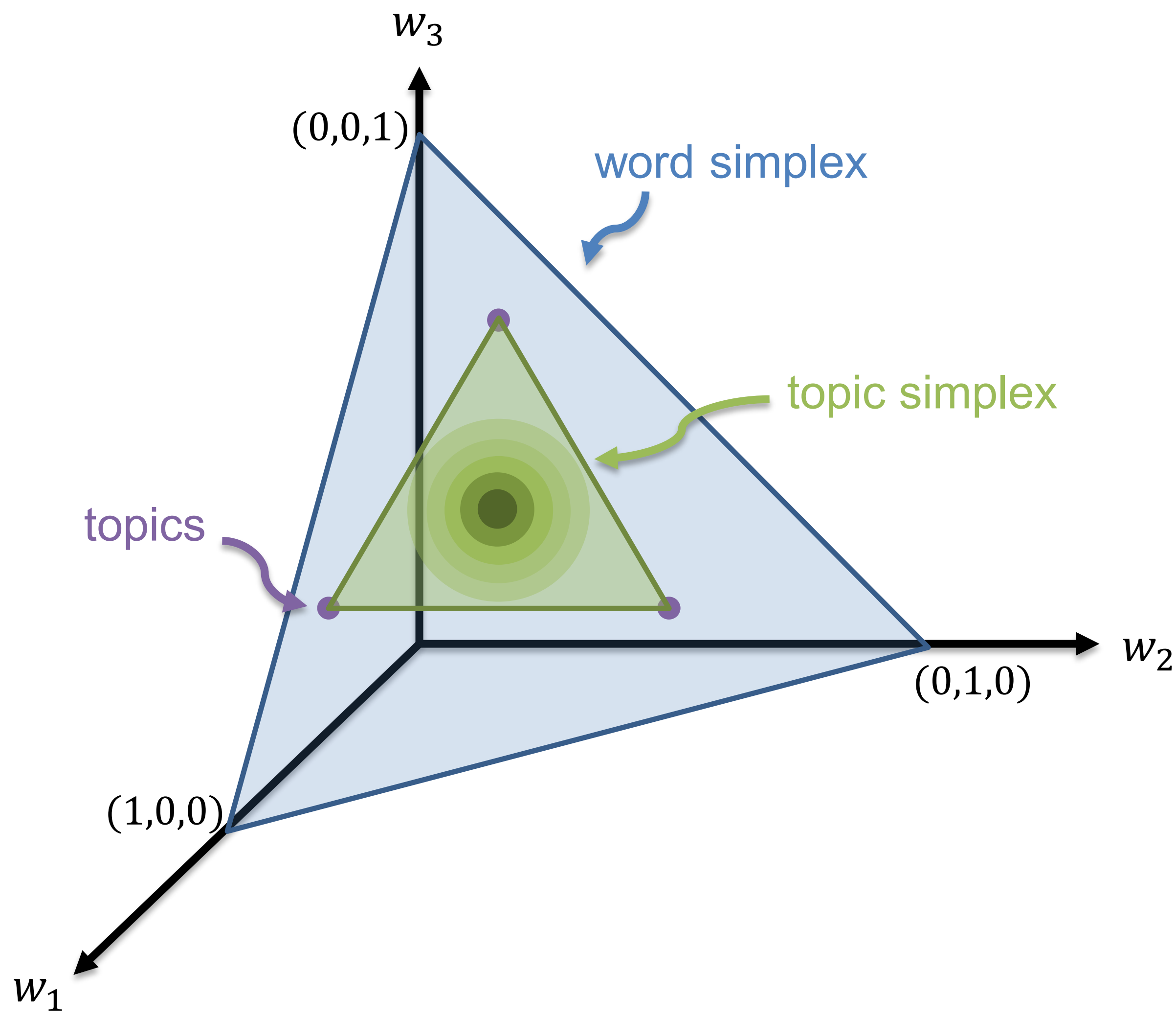
PLSA & LDA – 5

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



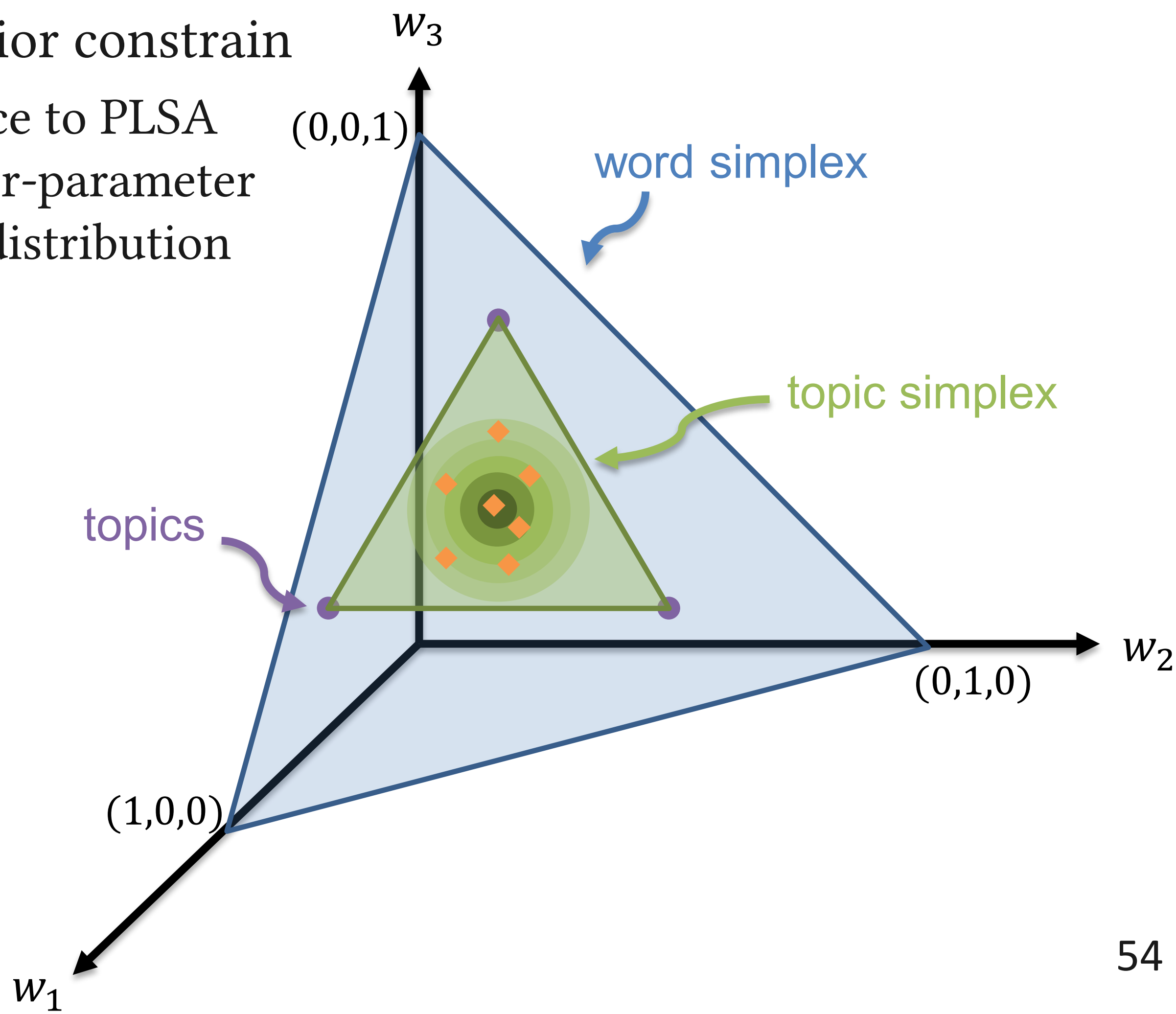
PLSA & LDA – 6

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

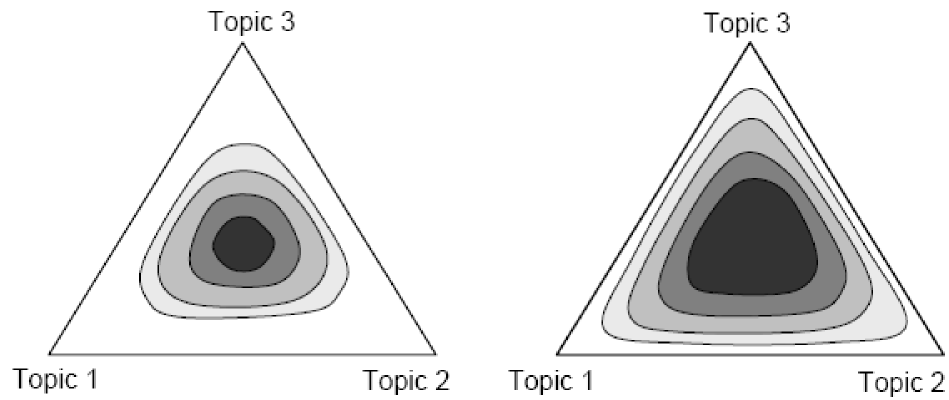


PLSA & LDA – 7

- 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



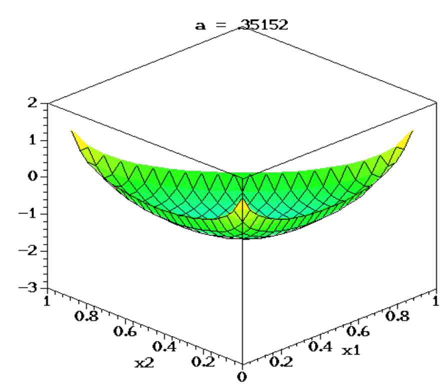
PLSA & LDA – 8



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

- 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

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



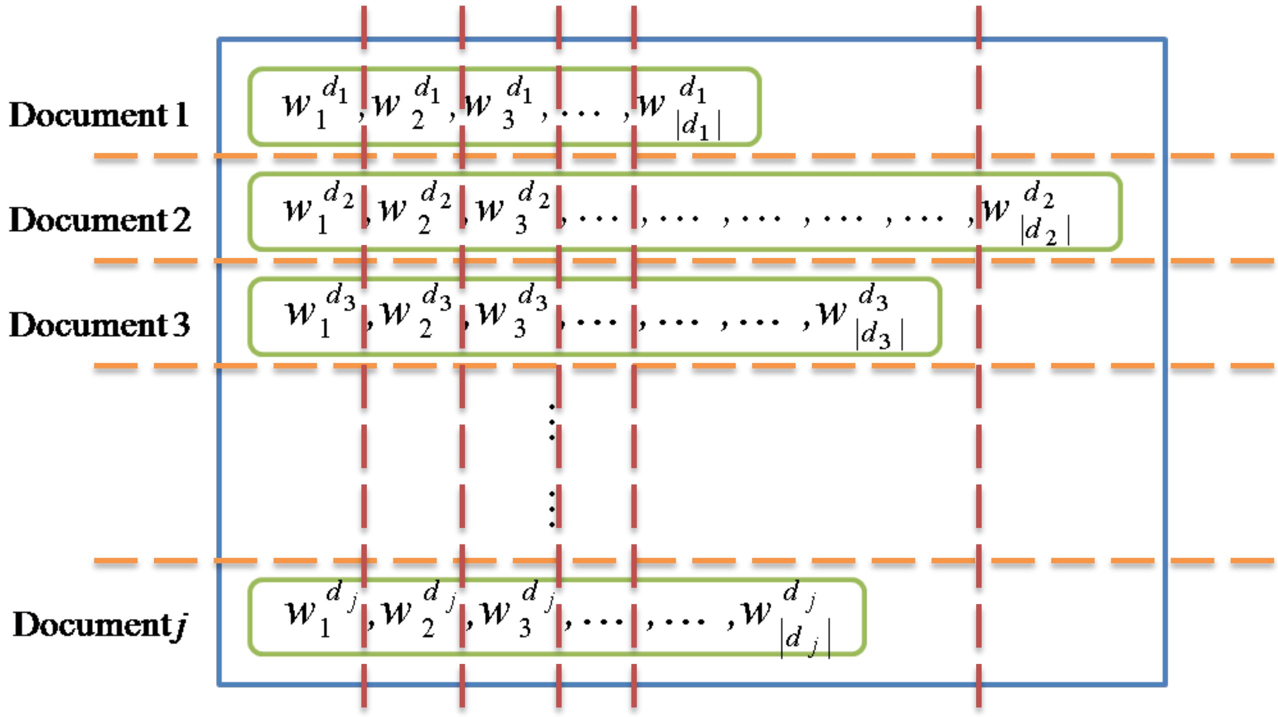
LDA – Experiments

- 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*

PLSA, LDA and WTM – 1

- PLSA, LDA and WTM can be analyzed from several perspectives
  - Explore the latent information from different points of view
    - “between words and documents” vs. “between words”



PLSA, LDA and WTM – 2

- Use different methods when calculating the topic mixture weights
  - "time-consuming" vs. "simple combination"

- Model parameters

Model	PLSA	LDA	WTM
Relationships	doc-word	doc-word	word-word
No. of parameters	$\lvert V \rvert \times K + \lvert \mathbf{D} \rvert \times K$	$K + \lvert V \rvert \times K$	$2 \times \lvert V \rvert \times K$

Homework 3 – Description

- In this project, you will have
  - 16 Short Queries
  - 2265 Documents
  - A Background Language Model (BGLM.txt)
  - A Set of Documents for Topic Model Training (Collection.txt)

- Our goal is to implement the PLSA model, and incorporate the PLSA and query likelihood measure for retrieval
  - Thus, the ultimate goal is to enhance the estimation of each document language model

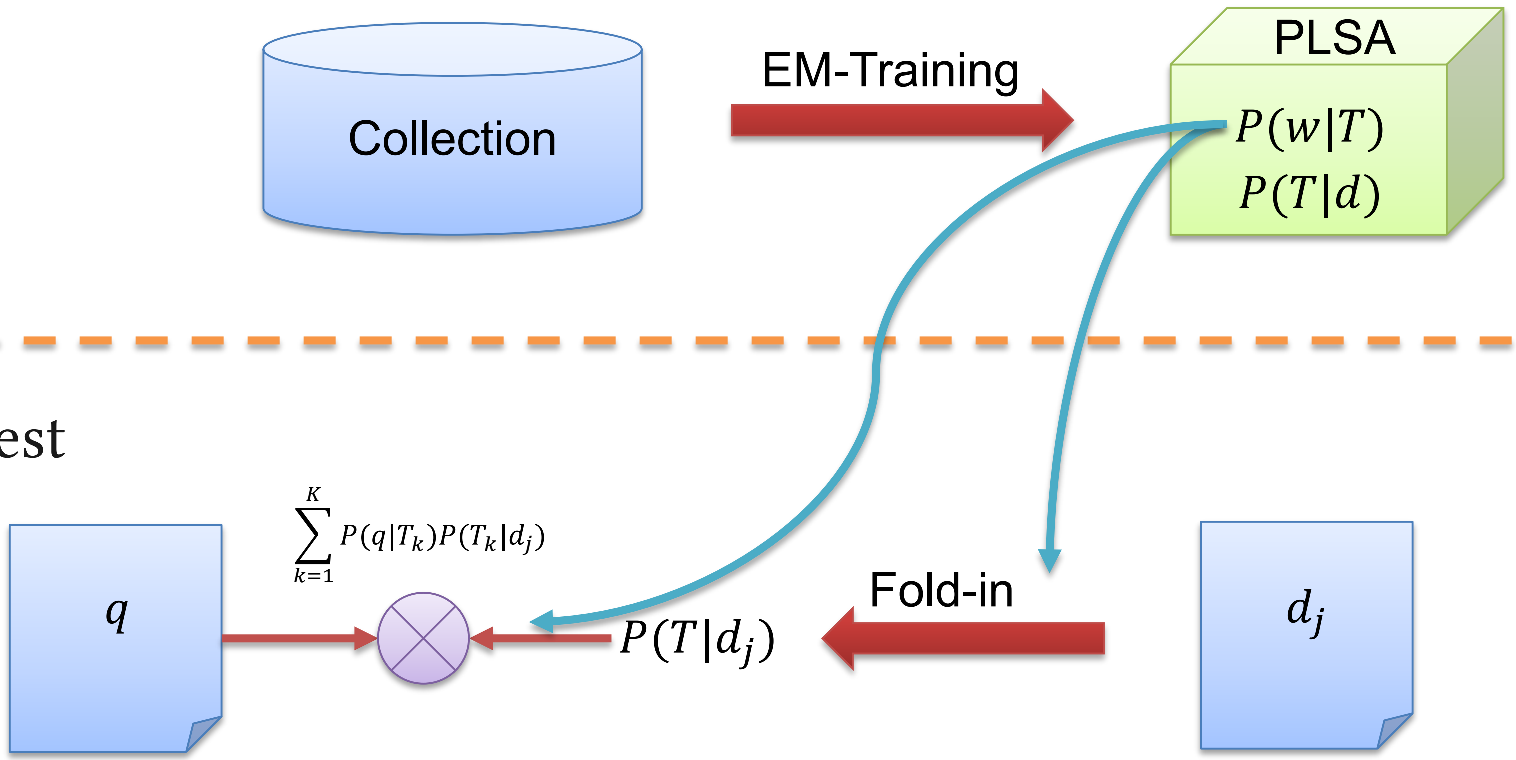
$$P(q|d_j) \approx \prod_{i=1}^{|q|} P'(w_i|d_j)$$

Obtained by using fold-in strategy

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

Homework 3 – Flowchart

- Training



- Test

Homework 3 – BGLM

- The vocabulary size is 51253
  - The word id starts from 0 to 51252

BGLM.txt - 記事本

檔案(F) 編輯(E) 格式(O) 檢視(V) 說明(H)

word id	$lnP(w_i)$
51215	-22.42474748728176700000
51216	-22.42474748728176700000
51217	-22.42474748728176700000
51218	-22.42474748728176700000
51219	-22.42474748728176700000
51220	-22.42474748728176700000
51221	-22.42474748728176700000
51222	-22.42474748728176700000
51223	-15.53506750127963300000
51224	-22.42474748728176700000
51225	-22.42474748728176700000
51226	-22.42474748728176700000
51227	-22.42474748728176700000
51228	-22.42474748728176700000
51229	-22.42474748728176700000
51230	-10.73104836924566800000
51231	-22.42474748728176700000
51232	-13.13717380611637800000
51233	-22.42474748728176700000
51234	-22.42474748728176700000
51235	-22.42474748728176700000
51236	-22.42474748728176700000
51237	-22.42474748728176700000
51238	-22.42474748728176700000
51239	-22.42474748728176700000
51240	-22.42474748728176700000
51241	-22.42474748728176700000
51242	-22.42474748728176700000
51243	-22.42474748728176700000
51244	-22.42474748728176700000
51245	-22.42474748728176700000
51246	-22.42474748728176700000
51247	-22.42474748728176700000
51248	-14.84192031073563500000
51249	-22.42474748728176700000
51250	-22.42474748728176700000
51251	-15.53506750127963300000
51252	-22.42474748728176700000

Homework 3 – Collection

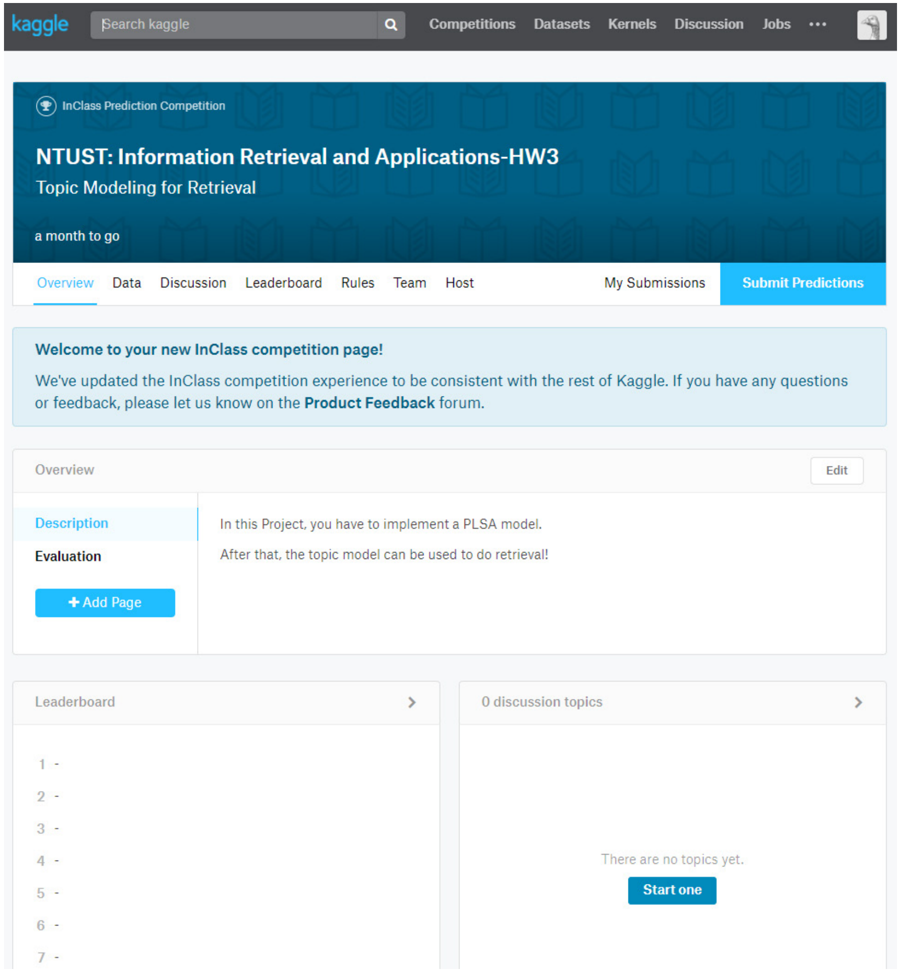
- In the collection file, each line refers to a document
  - Thus, we have about 18 thousand documents

Collection.txt

```
2717 3401 2279 1407 30048 465 2584 465 2584 906 596 27799 1027 1985 906 596 27799 1027 1985 38527 612 1407
40085 29686 20709 3020 3439 38527 3020 3439 20709 612 1407 2986 25607 1948 739 2746 8594 42072 42472 2376 14468 40085 1407 23088 2
591 3367 4165 1521 7680 41224 3383 18094 38527 40747 612 1407 2986 25607 1405 1250 2188 591 3367 4165 7680 1548 1929 3460 1407 1332
1152 2091 3596 31424 49951 7680 38527 45624 612 1407 2986 25607 820 6211 1744 764 1263 1943 49951 24554 24022 7680 12037 8824 3015
2851 44049 7764 44482 9024 19444 38527 23743 612 1407 2986 25607 3483 2343 44049 27646 22572 40526 40645 1407 596 3107 12085 2851
39180 26406 23785 710 43385 38527 21818 612 1407 2986 25607 1860 3755 10736 11185 8107 7546 596 39180 26704 9743 1330 8024 8813 2
18046 3835 26763 25460 29626 24310 3015 26273 2572 29978 38527 4165 3015 3981 4165 3388 3460 612 1407 2986 25607 3204 3293 1974 2962
11164 18329 644 35313 28940 27960 16906 17541 32882 38527 43039 612 1407 2986 25607 1405 1188 11164 14468 2619 2091 1547 1407 2913
2280 3068 457 26763 27631 14645 732 4332 2690 2923 20646 38527 27352 612 1407 2986 25607 432 2148 3140 27414 31424 7680 2280 3068 45
1484 2515 1164 1135 8594 596 40889 3015 30241 14452 38527 40889 612 1407 2986 25607 444 950 1298 8594 33266 40889 31622 1484 2515 11
2837 1072 7570 8221 28727 19721 39250 2572 1583 12620 38527 2619 2091 2054 612 1407 2986 25607 432 1263 5558 29626 1072 732 7570 822
20258 27646 23728 23771 30861 16678 38527 49648 612 1407 2986 25607 1860 11963 50171 3100 1715 612 1407 1542 18651 28342 33744 29
27647 7871 42763 1374 2763 22001 34382 38527 34616 612 1407 2986 27647 7871 42763 1374 2763 22001 34382 38527 25607 1657 31884 42763
27647 7871 1715 612 27647 31743 26254 38527 612 1407 2986 1715 612 27647 31743 26254 38527 25607 1164 2262 2343 8221 9966 24702 29
0 3015 30457 1072 732 2572 31011 35244 25164 22586 12335 3015 27646 15284 2572 7574 1072 732 3015 27564 8994 14881 24453 9283 21108 25
18316 855 20529 16729 18663 855 24485 38527 30628 612 1407 2986 25607 2055 3941 4171 18316 855 20529 33697 7546 30066 3015 30802 346
7712 41841 3130 2620 30580 3015 1715 38527 20364 612 1407 2986 7712 41841 3130 2620 30580 3015 1715 38527 25607 530 2001 2515 4184
20709 26594 26596 18790 457 1741 25167 38527 9898 40359 612 1407 2986 25607 1948 739 2746 21915 20709 3015 26594 26596 1407 596 924
1894 30064 28152 14433 14787 24546 8986 38527 50810 612 1407 2986 25607 3204 2273 47495 8994 612 1407 30064 27957 28152 14433 9561
43960 732 49029 38527 43960 3198 612 1407 2986 43960 732 2387 12710 21130 29618 1407 596 31085 1759 3015 2999 8891 34650 17045 35898
11726 13182 18402 9567 1408 23141 29978 8594 2397 38527 20364 612 1407 2986 25607 2965 3403 456 21413 18402 38361 1407 30066 34698
10736 18526 20258 1898 474 8221 2625 28421 20826 38527 49648 612 1407 2986 25607 1860 11963 2220 50171 38925 2572 10736 33266 20258
9879 7689 1135 21073 28421 8812 29752 38527 27067 612 1407 2986 25607 1960 3020 1444 9879 8994 7689 1657 13957 24554 596 9245 1200
38208 19194 39037 23873 9715 18126 1551 22194 38527 43965 612 1407 2986 25607 2668 491 1943 38208 38988 42680 1407 30066 17257 191
27647 353 2016 11408 16761 18964 37448 11708 38527 34616 612 1407 2986 1715 16761 18964 7609 41399 3015 44939 38506 44962 27924 596 6
35125 12026 2453 39607 38120 33229 1375 27920 38527 50612 612 1407 2986 25607 1076 3531 1507 41134 457 33229 21032 1152 3631 1109 33
16761 18964 34099 41399 21413 41611 2914 3383 15438 32565 38527 48455 612 1407 2986 25607 488 2495 1474 16761 18964 34099 41399 245
18047 3981 3383 36362 21385 7547 7609 37941 38527 1826 3017 1003 3015 4165 612 1407 2986 25607 2530 2442 2170 8670 8202 7609 37941 1
21413 1416 2386 39936 23728 43347 29602 38527 34616 612 1407 2986 21413 1416 3981 2786 1376 1109 25394 39936 651 612 1407 596 5016
25653 24022 10054 36776 33229 25197 15285 26088 38527 25653 612 1407 2986 25607 626 983 456 24554 25653 1257 18395 713 33624 36776
15912 17639 16665 24022 10054 18586 25197 33229 26088 38527 17639 612 1407 2986 25607 2626 22715 24554 18896 1278 820 662 1788 17
```

Homework 3 – Kaggle

- Please login our competition page at Kaggle
  - https://www.kaggle.com/t/3a0b466580294899a666e4c088be1074
- **Your team name is ID_Name**
  - M123456_陳冠宇



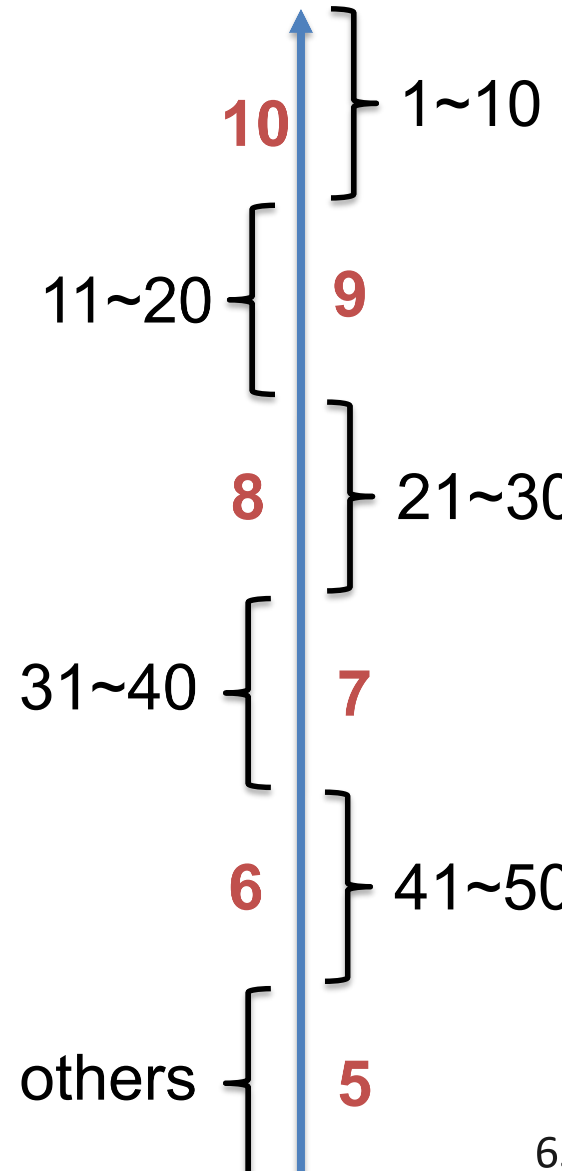
Homework 3 – Submission Format

submission.txt

```
Query,RetrievedDocuments
20001.query,VOM19980619.0700.0347 VOM19980225.0700.0510 VOM19980317.0900.0192 VOM19980317.0900.0330 VOM19980
00.0173 VOM19980302.0700.0241 VOM19980303.0700.2287 VOM19980530.0730.0166 VOM19980404.0700.2088 VOM19980616.0
216 VOM19980614.0700.0357 VOM19980626.0700.0409 VOM19980403.0700.0489 VOM19980523.0730.0220 VOM19980524.0730.
VOM19980624.0900.0077 VOM19980625.0700.0363 VOM19980605.0730.0152 VOM19980602.0730.0102 VOM19980603.0730.0280
9980522.0730.0037 VOM19980228.0700.0327 VOM19980414.0900.0260 VOM19980223.0700.0765 VOM19980505.0700.0529 VOM
503.0730.0136 VOM19980319.0900.3416 VOM19980620.0730.0034 VOM19980302.0700.0209 VOM19980302.0900.2091 VOM1998
0900.0207 VOM19980305.0900.1926 VOM19980521.0730.0029 VOM19980504.0700.0376 VOM19980314.0700.0239 VOM19980619.
.0137 VOM19980611.0700.0150 VOM19980326.0700.2112 VOM19980522.0900.0269 VOM19980503.0700.0412 VOM19980428.090
4 VOM19980422.0900.0021 VOM19980605.0700.0194 VOM19980611.0700.0046 VOM19980223.0700.2728 VOM19980614.0730.02
M19980303.0700.0696 VOM19980326.0900.0149 VOM19980505.0700.0481 VOM19980614.0730.0034 VOM19980226.0900.1964 V
80523.0730.0083 VOM19980316.0700.0356 VOM19980609.0900.0009 VOM19980314.0700.2300 VOM19980302.0700.2137 VOM19
4.0700.0458 VOM19980319.0900.2169 VOM19980305.0700.2126 VOM19980515.0700.0472 VOM19980403.0700.0129 VOM19980
30.0142 VOM19980618.0700.0234 VOM19980319.0900.0647 VOM19980527.0700.0528 VOM19980607.0730.0033 VOM19980305.0
20002.query,VOM19980530.0730.0101 VOM19980611.0900.0216 VOM19980506.0900.0089 VOM19980624.0700.0434 VOM19980
00.2021 VOM19980604.0900.0246 VOM19980606.0700.0562 VOM19980303.0900.2085 VOM19980225.0700.0999 VOM19980312.0
171 VOM19980220.0900.1979 VOM19980305.0700.0763 VOM19980627.0700.0360 VOM19980225.0700.0302 VOM19980529.0700.
VOM19980612.0730.0192 VOM19980319.0700.2737 VOM19980630.0700.0071 VOM19980526.0730.0131 VOM19980403.0700.0489
9980430.0900.0192 VOM19980502.0700.0307 VOM19980616.0700.0420 VOM19980319.0900.3468 VOM19980303.0900.1926 VOM
```

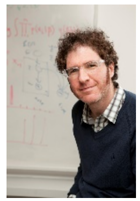
Homework 3 – Scoring

- The evaluation measure is MAP
- You can upload **ten** results each day
  - Turning the parameters
  - Please **Do Not** register several teams
- The **hard** deadline is 11/9 11:00
- You should also upload source codes and a mini report to moodle
  - TA will ask you to demo your program
  - In this HW, you can only leverage PLSA to do retrieval
  - If you use other models, you will get 0



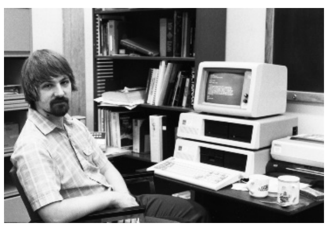
The Evolution

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

Questions?

kychen@mail.ntust.edu.tw