Latent Semantic Analysis (LSA)

Berlin Chen
Department of Computer Science & Information Engineering
National Taiwan Normal University

References:
Taxonomy of Classic IR Models

User Task

Retrieval: Adhoc Filtering

Structured Models

Non-Overlapping Lists
Proximal Nodes

Boolean Vector Probabilistic

Classic Models

Set Theoretic

Fuzzy Extended Boolean

Algebraic

Generalized Vector
Latent Semantic Analysis (LSA)
Neural Networks

Probabilistic

Inference Network
Belief Network

Browsing

Flat Structure Guided Hypertext

Browsing
Classification of IR Models Along Two Axes

• **Matching Strategy**
  – Literal term matching
    • E.g., Vector Space Model (VSM), Hidden Markov Model (HMM), Language Model (LM)
  – Concept matching
    • E.g., Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Topical Mixture Model (TMM)

• **Learning Capability**
  – Term weighting, query expansion, document expansion, etc
    • E.g., Vector Space Model, Latent Semantic Indexing
    • Most models are based on linear algebra operations
  – Solid statistical foundations (optimization algorithms)
    • E.g., Hidden Markov Model (HMM), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA)
    • Most models belong to the language modeling approach
Two Perspectives for IR Models (cont.)

- Literal Term Matching vs. Concept Matching

- There are usually many ways to express a given concept, so literal terms in a user's query may not match those of a relevant document.

  香港星島日報篇報導引述軍事觀察家的話表示，到二零零五年台灣將完全喪失空中優勢，原因是中國大陸戰機不論數量或是性能都將超越台灣，報導指出中國在大量引進俄羅斯先進武器的同時也加快研發自製武器系統，目前西安飛機製造廠任職的改進型飛豹戰機即將部署尚未與蘇愷三十通道地對地攻擊住宅飛機，以督促遇到挫折的監控其戰機目前也已經取得了重大階段性的認知成果。根據日本媒體報導在台海戰爭隨時可能爆發情況之下北京方面的基本方針，使用高科技答應局部戰爭。因此，解放軍打算在二零零四年前又有包括蘇愷三十二期在內的兩百架蘇霍伊戰鬥機。
Latent Semantic Analysis (LSA)

- Also called Latent Semantic Indexing (LSI), Latent Semantic Mapping (LSM), or Two-Mode Factor Analysis
  - 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

Latent Semantic Analysis: Schematic

- Dimension Reduction and Feature Extraction
  - **PCA**
    \[ y_i = \phi_i^T X \]
    \[ \sum_{i=1}^{k} y_i \phi_i \]
  - **SVD (in LSA)**
    \[ r \leq \min(m,n) \]
    \[ \min \| A' - A \|_F^2 \text{ for a given } k \]
LSA: An Example

- Singular Value Decomposition (SVD) used for the word-document matrix
  - A least-squares method for dimension reduction

<table>
<thead>
<tr>
<th>Query</th>
<th>Term 1</th>
<th>Term 2</th>
<th>Term 3</th>
<th>Term 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document 1</td>
<td>user</td>
<td>interface</td>
<td>HCI</td>
<td>interaction</td>
</tr>
<tr>
<td>Document 2</td>
<td>user</td>
<td>interface</td>
<td>HCI</td>
<td>interaction</td>
</tr>
</tbody>
</table>

Projection of a Vector \( \mathbf{x} \):

\[
y_1 = \|\mathbf{x}\| \cos \theta_1 = \|\mathbf{x}\| \frac{\mathbf{\phi}_1^T \mathbf{x}}{\|\mathbf{\phi}_1\|} = \mathbf{\phi}_1^T \mathbf{x}
\]

, where \(\|\mathbf{\phi}_1\| = 1\)
LSA: Latent Structure Space

• Two alternative frameworks to circumvent vocabulary mismatch
# LSA: Another Example (1/2)

**Titles**
- **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*

<table>
<thead>
<tr>
<th>Terms</th>
<th>c1</th>
<th>c2</th>
<th>c3</th>
<th>c4</th>
<th>c5</th>
<th>m1</th>
<th>m2</th>
<th>m3</th>
<th>m4</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>interface</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>computer</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>user</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>system</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>response</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>time</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EPS</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>survey</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>trees</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>graph</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>minors</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
LSA: Another Example (2/2)

2-D Plot of Terms and Docs from Example

Query: “human computer interaction”

An OOV word

FIG. 1. A two-dimensional plot of 12 Terms and 9 Documents from the same TM set. Terms are represented by filled circles. Documents are shown as open squares, and component terms are indicated parenthetically. The query ("human computer interaction") is represented as a pseudo-document at point q. Axes are scaled for Document-Document or Term-Term comparisons. The dotted cone represents the region whose points are within a cosine of .9 from the query q. All documents about human-computer (c1–c5) are “near” the query (i.e., within this cone), but none of the graph theory documents (m1–m4) are nearby. In this reduced space, even documents c3 and c5 which share no terms with the query are near it.
LSA: Theoretical Foundation

- **Singular Value Decomposition (SVD)**

\[
A = U\Sigma V^T
\]

\[
A' = U'\Sigma' V'^T
\]

Docs and queries are represented in a \( k \)-dimensional space. The quantities of the axes can be properly weighted according to the associated diagonal values of \( \Sigma_k \).

Row \( A \in \mathbb{R}^n \)

Col \( A \in \mathbb{R}^m \)

Both \( U \) and \( V \) has orthonormal column vectors

\[
U^TU = I_{r \times r}
\]

\[
V^TV = I_{r \times r}
\]

\[
||A||_F^2 \geq ||A'||_F^2
\]

\[
||A||_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2
\]

IR – Berlin Chen 11
LSA: Theoretical Foundation

• “term-document” matrix \( A \) has to do with the co-occurrences between terms (or units) and documents (or compositions)
  – Contextual information for words in documents is discarded
  • “bag-of-words” modeling

• Feature extraction for the entities \( a_{i,j} \) of matrix \( A \)
  1. Conventional \( tf-idf \) statistics
  2. Or, \( a_{i,j} \) : occurrence frequency weighted by negative entropy

\[
a_{i,j} = \frac{f_{i,j}}{|d_j|} \times (1 - \varepsilon_i), \quad |d_j| = \sum_{i=1}^{m} f_{i,j}
\]

\[
\varepsilon_i = -\frac{1}{\log n} \sum_{j=1}^{n} \left( \frac{f_{i,j}}{\tau_i} \log \frac{f_{i,j}}{\tau_i} \right), \quad \tau_i = \sum_{j=1}^{n} f_{i,j}
\]

\( 0 \leq \varepsilon_i \leq 1 \)
LSA: Theoretical Foundation

- **Singular Value Decomposition (SVD)**
  - $A^T A$ is symmetric $n \times n$ matrix
  - All eigenvalues $\lambda_j$ are nonnegative real numbers
    \[
    \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0 \quad \Sigma^2 = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)
    \]
  - All eigenvectors $v_j$ are orthonormal ($\in \mathbb{R}^n$)
    \[
    V = \begin{bmatrix} v_1 & v_2 & \ldots & v_n \end{bmatrix} \quad \forall j \quad v_j^T v_j = 1 \quad (V^T V = I_{n \times n})
    \]

- Define **singular values**: $\sigma_j = \sqrt{\lambda_j}, \quad j = 1, \ldots, n$
  - As the square roots of the eigenvalues of $A^T A$
  - As the lengths of the vectors $Av_1, Av_2, \ldots, Av_n$

\[
\text{For } \lambda_i \neq 0, \quad i=1,\ldots,r, \\
\{Av_1, Av_2, \ldots, Av_r\} \text{ is an orthogonal basis of Col } A
\]

\[
\sigma_1 = \|Av_1\| \\
\sigma_2 = \|Av_2\| \\
\ldots
\]

\[
\|Av_i\|^2 = v_i^T A^T A v_i = v_i^T \lambda_i v_i = \lambda_i \\
\Rightarrow \|Av_i\| = \sigma_i
\]
LSA: Theoretical Foundation

• \{Av_1, Av_2, \ldots, Av_r\} is an orthogonal basis of Col A

\[ Av_i \cdot Av_j = (Av_i)^T Av_j = v_i^T A^T Av_j = \lambda_j v_i^T v_j = 0 \]

– Suppose that A (or \(A^T A\)) has rank \(r \leq n\)

\[ \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0, \quad \lambda_{r+1} = \lambda_{r+2} = \ldots = \lambda_n = 0 \]

– Define an orthonormal basis \{u_1, u_2, \ldots, u_r\} for Col A

\[ u_i = \frac{1}{\|Av_i\|} Av_i = \frac{1}{\sigma_i} Av_i \implies \sigma_i u_i = Av_i \]

\[ \Rightarrow [u_1 u_2 \ldots u_r] \Sigma_r = \begin{bmatrix} v_1 & v_2 & \cdots & v_r \end{bmatrix} \]

\[ \Sigma_r = \begin{bmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_r \end{bmatrix} \]

\(V: an \text{ orthonormal matrix} \)

\(U: \text{an orthonormal matrix} (m \times r)\)

– Extend to an orthonormal basis \{u_1, u_2, \ldots, u_m\} of \(\mathbb{R}^m\)

\[ \Rightarrow [u_1 u_2 \ldots u_r u_{r+1} \ldots u_m] \Sigma = A[v_1 v_2 \ldots v_r v_{r+1} \ldots v_n] \]

\[ \Rightarrow U \Sigma = AV \implies U \Sigma V^T = AVV^T \]

\[ \Rightarrow A = U \Sigma V^T \]

\[ \sum_{m \times n} = \begin{bmatrix} \Sigma_r & 0_{(m-r) \times r} \\ 0_{(m-r) \times m} & 0_{(m-r) \times (m-r)} \end{bmatrix} I_{n \times n} \]

\[ \|A\|_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2 \]

\[ \|A\|_F^2 = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_r^2 \]
LSA: Theoretical Foundation

$v_i$ spans the row space of $A$

$u_i$ spans the row space of $A^T$

$U \Sigma V^T = (U_1 \quad U_2) \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix}$

$= U_1 \Sigma_1 V_1^T$

$= AV_1 V_1^T$

$U \Sigma = AV$

FIGURE 4 The four fundamental subspaces and the action of $A$. 

$AX = 0$
LSA: Theoretical Foundation

• Additional Explanations
  – Each row of $U$ is related to the projection of a corresponding row of $A$ onto the basis formed by columns of $V$
    
    $$A = U\Sigma V^T$$

    $$\Rightarrow AV = U\Sigma V^TV = U\Sigma \quad \Rightarrow U\Sigma = AV$$

  – the $i$-th entry of a row of $U$ is related to the projection of a corresponding row of $A$ onto the $i$-th column of $V$

  – Each row of $V$ is related to the projection of a corresponding row of $A^T$ onto the basis formed by $U$
    
    $$A = U\Sigma V^T$$

    $$\Rightarrow A^TU = (U\Sigma V^T)^TU = V\Sigma U^TU = V\Sigma$$

    $$\Rightarrow V\Sigma = A^TU$$

  – the $i$-th entry of a row of $V$ is related to the projection of a corresponding row of $A^T$ onto the $i$-th column of $U$
LSA: Theoretical Foundation

- Fundamental comparisons based on SVD
  - The original word-document matrix (A)
    - compare two terms → dot product of two rows of A
      - or an entry in $AA^\top$
    - compare two docs → dot product of two columns of A
      - or an entry in $A^\top A$
    - compare a term and a doc → each individual entry of A

- The new word-document matrix (A')
  - compare two terms \[ A'A^\top = (U' \Sigma' V'^\top) (U' \Sigma' V'^\top)^\top = U' \Sigma' V'^\top V'^\top \Sigma' U'^\top = (U' \Sigma')(U' \Sigma')^\top \]
  - compare two docs \[ A'^\top A' = (U' \Sigma' V'^\top)^\top (U' \Sigma' V'^\top) = V'^\top U'^\top \Sigma' U'^\top V'^\top = (V' \Sigma')(V' \Sigma')^\top \]
  - compare a query word and a doc → each individual entry of $A'$
LSA: Theoretical Foundation

• **Fold-in**: find representations for pseudo-docs \( q \)
  - For objects (new queries or docs) that did not appear in the original analysis
• Fold-in a new \( m \times 1 \) query (or doc) vector

\[
\hat{q}_{1 \times k} = \left( q^T \right)_{1 \times m} U_{m \times k} \Sigma^{-1} k \times k
\]

- **Row vectors**
- Query represented by the weighted sum of its constituent term vectors
- The separate dimensions are differentially weighted
- Represented as the weighted sum of its component word (or term) vectors
- Cosine measure between the query and doc vectors in the latent semantic space

\[
sim(\hat{q}, \hat{d}) = \text{coine} \left( \hat{q} \Sigma, \hat{d} \Sigma \right) = \frac{\hat{q} \Sigma \hat{d}^T}{\|\hat{q} \Sigma\| \|\hat{d} \Sigma\|}
\]

See Figure A in next page
LSA: Theoretical Foundation

• Fold-in a new $1 \times n$ term vector

$$\hat{t}_{1 \times k} = t_{1 \times n} V_{n \times k} \Sigma_{k \times k}^{-1}$$

See Figure B below
LSA: A Simple Evaluation

- Experimental results
  - HMM is consistently better than VSM at all recall levels
  - LSA is better than VSM at higher recall levels

Recall-Precision curve at 11 standard recall levels evaluated on TDT-3 SD collection. (Using word-level indexing terms)
LSA: Pro and Con (1/2)

• Pro (Advantages)
  – A clean formal framework and a clearly defined optimization criterion (least-squares)
    • Conceptual simplicity and clarity
  – Handle synonymy problems (“heterogeneous vocabulary”)
    • Replace individual terms as the descriptors of documents by independent “artificial concepts” that can specified by any one of several terms (or documents) or combinations
  – Good results for high-recall search
    • Take term co-occurrence into account
LSA: Pro and Con (2/2)

- Disadvantages
  - High computational complexity (e.g., SVD decomposition)
  - Exhaustive comparison of a query against all stored documents is needed (cannot make use of inverted files)
  - LSA offers only a partial solution to polysemy (e.g., bank, bass, ...)
    - Every term is represented as just one point in the latent space (represented as weighted average of different meanings of a term)
LSA: Junk E-mail Filtering

- One vector represents the centroid of all e-mails that are of interest to the user, while the other the centroid of all e-mails that are not of interest.
LSA: Dynamic Language Model Adaptation (1/4)

- Let \( w_q \) denote the word about to be predicted, and \( H_{q-1} \) the admissible LSA history (context) for this particular word
  - The vector representation of \( H_{q-1} \) is expressed by \( \tilde{d}_{q-1} \)
  - Which can be then projected into the latent semantic space

\[
\tilde{v}_{q-1} = \tilde{v}_{q-1}S = \tilde{d}_{q-1}^T U \quad \text{[change of notation : } S = \Sigma]\]

- Iteratively update \( \tilde{d}_{q-1} \) and \( \tilde{v}_{q-1} \) as the decoding evolves

\[
\tilde{d}_q = \frac{n_q - 1}{n_q} \tilde{d}_{q-1} + \frac{1 - \varepsilon_i}{n_q} [0...1...0]^T \]

\[
\tilde{v}_q = \tilde{v}_q S = \tilde{d}_{q-1}^T U = \frac{1}{n_q} \left[ (n_q - 1) \tilde{v}_{q-1} + (1 - \varepsilon_i) u_i \right]
\]

or

\[
= \frac{1}{n_q} \left[ \lambda \cdot (n_q - 1) \tilde{v}_{q-1} + (1 - \varepsilon_i) u_i \right] \quad \text{with exponential decay}
\]
LSA: Dynamic Language Model Adaptation (2/4)

- Integration of LSA with N-grams

\[
\Pr(w_q \mid H_{q-1}^{(n+l)}) = \Pr(w_q \mid H_{q-1}^{(n)}, H_{q-1}^{(l)})
\]

where \( H_{q-1} \) denotes some suitable history for word \( w_q \), and the superscripts \((n)\) and \((l)\) refer to the \( n \)-gram component \((w_{q-1}w_{q-2}...w_{q-n+1}, \text{ with } n > 1)\), the LSA component \((\tilde{d}_{q-1})\):

This expression can be rewritten as:

\[
\Pr(w_q \mid H_{q-1}^{(n+l)}) = \frac{\Pr(w_q, H_{q-1}^{(l)} \mid H_{q-1}^{(n)})}{\sum_{w_i \in V} \Pr(w_i, H_{q-1}^{(l)} \mid H_{q-1}^{(n)})}
\]
LSA: Dynamic Language Model Adaptation (3/4)

• Integration of LSA with N-grams (cont.)

\[
\Pr(w_q, H^{(l)}_{q-1} | H^{(n)}_{q-1}) = \\
\Pr(w_q | H^{(n)}_{q-1}) \cdot \Pr(H^{(l)}_{q-1} | w_q, H^{(n)}_{q-1}) \\
= \Pr(w_q | w_{q-1}w_{q-2} \cdots w_{q-n+1}) \cdot \Pr(\tilde{d}_{q-1} | w_qw_{q-1}w_{q-2} \cdots w_{q-n+1}) \\
= \Pr(w_q | w_{q-1}w_{q-2} \cdots w_{q-n+1}) \cdot \Pr(\tilde{d}_{q-1} | w_q) \\
= \Pr(w_q | w_{q-1}w_{q-2} \cdots w_{q-n+1}) \cdot \frac{\Pr(w_q | \tilde{d}_{q-1}) \Pr(\tilde{d}_{q-1})}{\Pr(w_q)} \\
= \frac{\Pr(w_q | w_{q-1}w_{q-2} \cdots w_{q-n+1}) \cdot \Pr(w_q | \tilde{d}_{q-1}) \Pr(\tilde{d}_{q-1})}{\sum_{w_i \in V} \Pr(w_i | w_{q-1}w_{q-2} \cdots w_{q-n+1}) \cdot \Pr(w_i | \tilde{d}_{q-1}) \Pr(\tilde{d}_{q-1})}
\]

Assume the probability of the document history given the current word is not affected by the immediate context preceding it.
LSA: Dynamic Language Model Adaptation

Intuitively, $\Pr(w_q \mid \tilde{d}_{q-1})$ reflects the "relevance" of word $w_q$ to the admissible history, as observed through $\tilde{d}_{q-1}$:

$$\Pr(w_q \mid \tilde{d}_{q-1})$$

$$\approx K(w_q, \tilde{d}_{q-1})$$

$$= \cos( u_q S^{1/2} , \tilde{v}_{q-1} S^{1/2} ) = \frac{u_q S \tilde{v}_{q-1}^T}{\| u_q S^{1/2} \| \| \tilde{v}_{q-1} S^{1/2} \|}$$

As such, it will be highest for words whose meaning aligns most closely with the semantic fabric of $\tilde{d}_{q-1}$ (i.e., relevant "content" words), and lowest for words which do not convey any particular information about this fabric (e.g., "function" works like "the").
LSA: Cross-lingual Language Model Adaptation (1/2)

- Assume that a document-aligned (instead of sentence-aligned) Chinese-English bilingual corpus is provided

\[ W = U \times S \times V^T \]

SVD of a word-document matrix for CL-LSA.

\[ \bar{W} = \bar{U} \times \bar{S} \times \bar{V}^T \]

Folding-in a monolingual corpus into LSA.

Lexical triggers and latent semantic analysis for cross-lingual language model adaptation, TALIP 2004, 3(2)
LSA: Cross-lingual Language Model Adaptation (2/2)

• CL-LSA adapted Language Model

\[
P_{\text{Adapt}} \left( c_k \| c_{k-1}, c_{k-2}, d^E_i \right) \\
\approx \lambda \cdot P_{\text{CL-LSA-Unigram}} \left( c_k \| d^E_i \right) + (1 - \lambda) \cdot P_{BG} \left( c_k \| c_{k-1}, c_{k-2} \right)
\]

\[
P_{\text{CL-LSA-Unigram}} \left( c_k \| d^E_i \right) = \sum_e P_T \left( c \| e \right) P \left( e \| d^E_i \right)
\]

\[
P_T \left( c \| e \right) \approx \frac{\text{sim} \left( \tilde{c}, \tilde{e} \right)^\gamma}{\sum_{c'} \text{sim} \left( \tilde{c}', \tilde{e} \right)^\gamma} \quad (\gamma \gg 1)
\]

\(d^E_i\) is a relevant English doc of the Mandarin doc being transcribed, obtained by CL-IR.
LSA: SVDLIBC

- Doug Rohde's SVD C Library version 1.3 is based on the SVDPACKC library

- Download it at http://tedlab.mit.edu/~dr/
LSA: Exercise (1/4)

• Given a sparse term-document matrix
  – E.g., 4 terms and 3 docs
  – Each entry can be weighted by $TF \times IDF$ score

• Perform SVD to obtain term and document vectors represented in the latent semantic space

• Evaluate the information retrieval capability of the LSA approach by using varying sizes (e.g., 100, 200, ..., 600 etc.) of LSA dimensionality
LSA: Exercise (2/4)

• Example: term-document matrix

<table>
<thead>
<tr>
<th>Indexing</th>
<th>Doc no.</th>
<th>Nonzero entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term no.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>51253</td>
<td>2265</td>
<td>218852</td>
</tr>
<tr>
<td>77</td>
<td></td>
<td></td>
</tr>
<tr>
<td>508</td>
<td>7.725771</td>
<td></td>
</tr>
<tr>
<td>596</td>
<td>16.213399</td>
<td></td>
</tr>
<tr>
<td>612</td>
<td>13.080868</td>
<td></td>
</tr>
<tr>
<td>709</td>
<td>7.725771</td>
<td></td>
</tr>
<tr>
<td>713</td>
<td>7.725771</td>
<td></td>
</tr>
<tr>
<td>744</td>
<td>7.725771</td>
<td></td>
</tr>
<tr>
<td>1190</td>
<td>7.725771</td>
<td></td>
</tr>
<tr>
<td>1200</td>
<td>16.213399</td>
<td></td>
</tr>
<tr>
<td>1259</td>
<td>7.725771</td>
<td></td>
</tr>
</tbody>
</table>

......

• SVD command (IR_svd.bat)

```
svd -r st -o LSA100 -d 100 Term-Doc-Matrix
```

output

LSA100-Ut
LSA100-S
LSA100-Vt
LSA: Exercise (3/4)

- **LSA100-Ut**
  - Word vector \( (u^T) \): 1x100
  - 51253 words

- **LSA100-Vt**
  - Doc vector \( (v^T) \): 1x100
  - 2265 docs

- **LSA100-S**
  - 100 eigenvalues
  - 100
LSA: Exercise (4/4)

• Fold-in a new $m \times 1$ query vector

$$\hat{q}_{1 \times k} = \left( q^T \right)_{1 \times m} U_{m \times k} \Sigma^{-1}_{k \times k}$$

The separate dimensions are differentially weighted

Just like a row of $V$

Query represented by the weighted sum of its constituent term vectors

• Cosine measure between the query and doc vectors in the latent semantic space

$$\text{sim} \ (\hat{q}, \hat{d}) = \text{coine} \ (\hat{q} \Sigma, \hat{d} \Sigma) = \frac{\hat{q} \Sigma^2 \hat{d}^T}{|\hat{q} \Sigma| |\hat{d} \Sigma|}$$