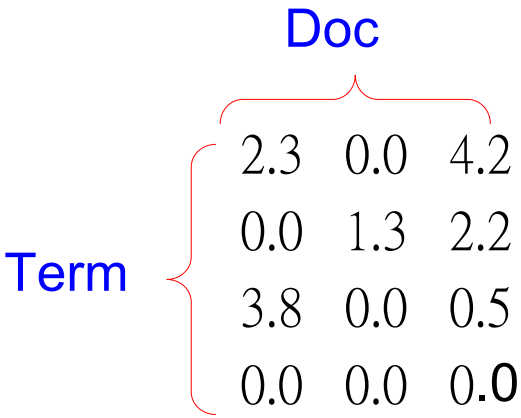
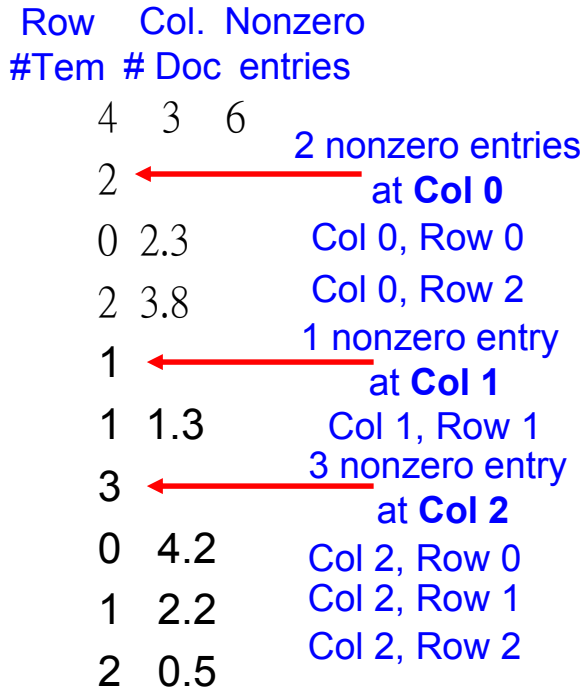
HW: Latent Semantic Analysis (LSA)

- Given a sparse term-doc matrix
  - E.g., 4 terms and 3 docs





  - Each entry is weighted by *TF*x*IDF* score

- Perform SVD to obtain corresponding term and doc 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

HW: Latent Semantic Analysis (cont.)

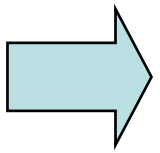
- Example: term-docmatrix

Indexing Term no. | Doc no. | Nonzero entries

```
51253 2265 218852
77
508 7.725771
596 16.213399
612 13.080868
709 7.725771
713 7.725771
744 7.725771
1190 7.725771
1200 16.213399
1259 7.725771
……
```

- SVD command (IR_svd.bat)

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



output

LSA100-Ut

LSA100-S

LSA100-Vt

sparse matrix input

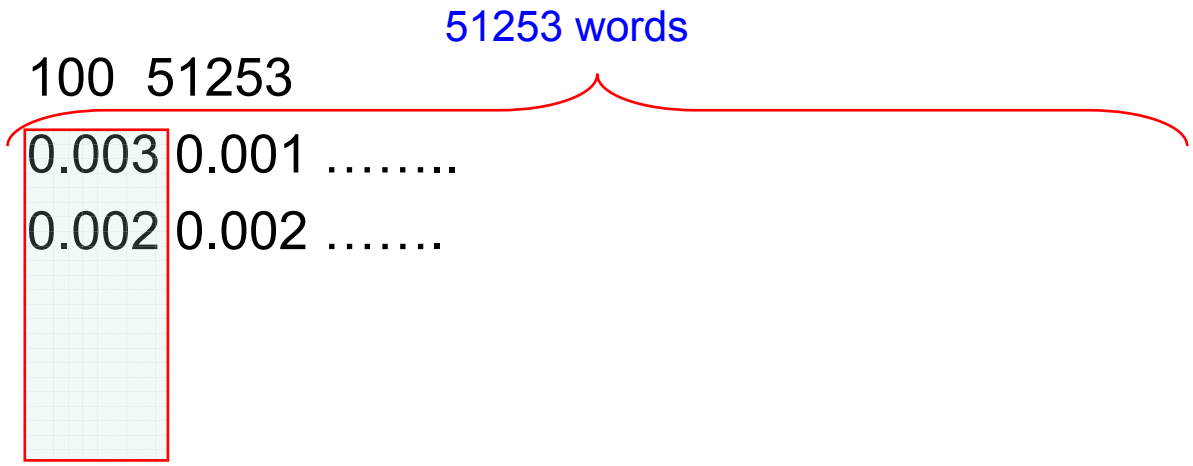
prefix of output files

No. of reserved eigenvectors

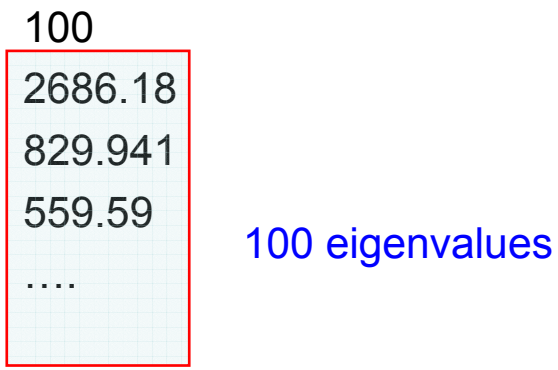
name of sparse matrix input

HW: Latent Semantic Analysis (cont.)

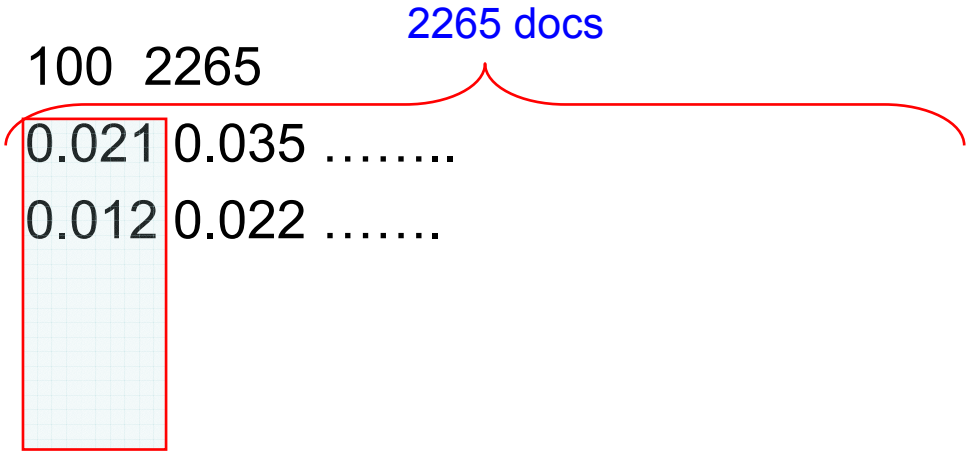
- **LSA100-Ut**



- **LSA100-S**



- **LSA100-Vt**



HW: Latent Semantic Analysis (cont.)

- Fold-in a new *m*x1 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 it constituent term vectors

TFxIDF weighted beforehand

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

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

SVDLIBC

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

- Download it at http://tedlab.mit.edu/~dr/