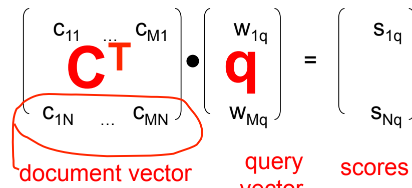
Latent Semantic Indexing: Introduction

- Analysis of term-document interaction for corpus of text documents
- Standard vector model:
  - document vector of term weights
- Goals:
  - reduce dimension of document vectors
  - uncover latent factors:
    - document as vector of factor weights
- uses of theory of linear algebra

1

Matrix formulation

M - number of terms in lexicon
N - number of documents in collection
C the M×N (term×doc.) matrix of weights ≥ 0 (our old w_{ij})



$s_{xq} = \Sigma^t_{i=1}(c_{ix} * w_{iq})$

2

Set-up

C the M×N (term×doc.) matrix of non-negative weights
- of rank r (r ≤ min(M,N))
- documents are *columns* of C

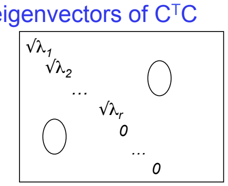
consider CC^T and C^TC:
- symmetric,
- share the same eigenvalues $\lambda_1, \lambda_2, \ldots$
  - $\lambda_1, \lambda_2, \ldots$ are indexed in decreasing order

- $C^TC(i,j)$ measures similarity documents i and j
- $CC^T(i,j)$ measures strength co-occurrence terms i and j

3

Use Singular Value Decomposition (SVD)

Theorem:
M×N matrix C of rank r has a
singular value decomposition $C = U\Sigma V^T$
Where:
U M×M matrix
with columns = orthogonal eigenvectors of CC^T
V N×N matrix
with columns = orthogonal eigenvectors of C^TC
Σ M×N diagonal matrix:
$\Sigma(i,i) = \sqrt{\lambda_i}$ for $1 \leq i \leq r$
$\Sigma(i,j) = 0$ otherwise
$\sqrt{\lambda_i}$ *called singular values*

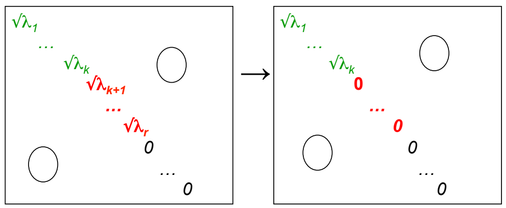


4

Reduce Rank

- Reduce rank of Σ from r to **k**
  keep only k largest singular values

Σ_K is M×N diagonal matrix: $\Sigma(i,i) = \sqrt{\lambda_i}$ for $1 \le i \le$ **k**
$\Sigma(i,j) = 0$ otherwise



5

Reduced Rank Approximation of C

- Approximation:

$$C_k = U\Sigma_k V^T$$

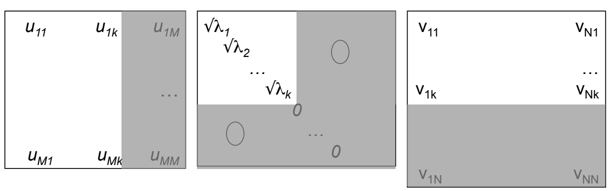
[M×N] [M×M] [M×N] [N×N]

- Theorem:
  C_k is the best rank-k approximation to C under the least square fit (Frobenius) norm

$$= \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{N} (C(i,j) - C_k(i,j))^2}$$

6

Reduced dimension matrices



C_k =	**U'_k**	**Σ'_k**	**V'^T_k**
M×N	M×k	k×k	k×N

7

Semantic Interpretation

- remaining k dimensions: k factors
- View V'^T_k as a representation of documents in the k-dimensional space
- View U'^T_k as a representation of terms in the k-dimensional space
- Σ_k scales between them
- find some semantic relationship?
  - "concept space"?
  - correlating terms to find structure
    - synonomy
    - polysomy

"people choose same main terms <20% time"

8

Using the Approximation

- V'^T_k as a representation of documents in a k-dimensional space
- $C_k^T C_k = (U'_k \Sigma'_k V'^T_k)^T (U'_k \Sigma'_k V'^T_k)$
  $= (V'_k \Sigma'^T_k U'^T_k)(U'_k \Sigma'_k V'^T_k)$
  $= V'_k (\Sigma'_k)^2 (V'_k)^T$ compares documents
- Transform query vector **q** into that space:

  $U'_k \Sigma'_k V'^T_k = C_k \Rightarrow V'^T_k = (\Sigma'_k)^{-1} (U'_k)^T C_k$

  Then $(\Sigma'_k)^{-1} (U'_k)^T \boldsymbol{q} = \boldsymbol{q}_k$

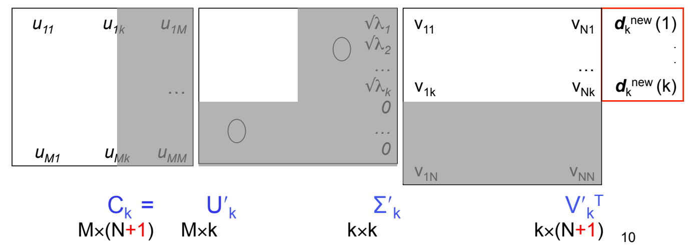
recalling $(V'^T_k)(V'_k) = (U'^T_k)(U'_k) = I$

Adding a new document

add new document $\boldsymbol{d}^{new}$ to C_k => add column $\boldsymbol{d}_k^{new}$ to V'^T_k

Transform $\mathbf{d}^{new}$ into the k-dimensional space version $\boldsymbol{d}_k^{new}$

$V'^T_k = (\Sigma'_k)^{-1} (U'_k)^T C_k \Rightarrow (\Sigma'_k)^{-1} (U'_k)^T \boldsymbol{d}^{new} = \boldsymbol{d}_k^{new}$



Original LSI paper:

Deerwester, Dumais, et. al.
Indexing by Latent Semantic Analysis
Journal of the Society for Information Science, 41(6), 1990, 391-407.

Example from that paper follows

Deerwester, Dumais et. al. Table:

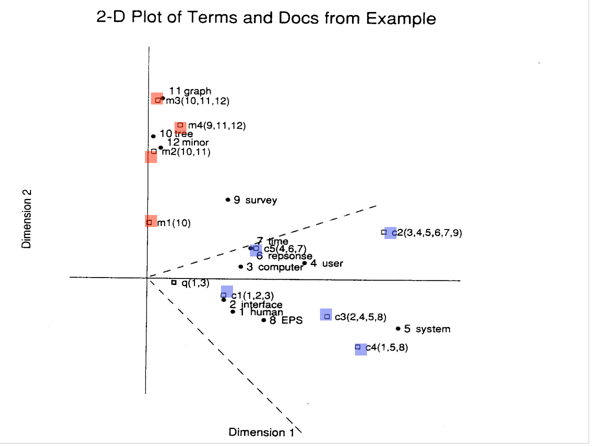
Terms				**Documents**					
	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
System	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

Deerwester, Dumais et. al. example, cont.:

Matrix V'^T_k for k=2

0.20	0.61	0.46	0.54	0.28	0.00	0.02	0.02	0.08
-0.06	0.17	-0.13	-0.23	0.11	0.19	0.44	0.62	0.53

Deerwester, Dumais, et al Figure 1



Summary

- LSI uses SVD to get a reduced-rank and reduced-size approximation to C
- LSI can be viewed as a preprocessor for
  - query evaluation
  - clustering
- SVD computation can be costly
  - do once (or rarely)