NLP

Text Similarity

Dimensionality Reduction

Problems with the Simple Vector Approaches to Similarity

- Polysemy (sim < cos)
  - bar, bank, jaguar, hot
- Synonymy (sim > cos)
  - building/edifice, large/big, spicy/hot
- Relatedness (people are really good at figuring this)
  - doctor/patient/nurse/treatment
- Sparse matrix
- Needed
  - dimensionality reduction

TOEFL Synonyms and SAT Analogies

- Word similarity vs. analogies

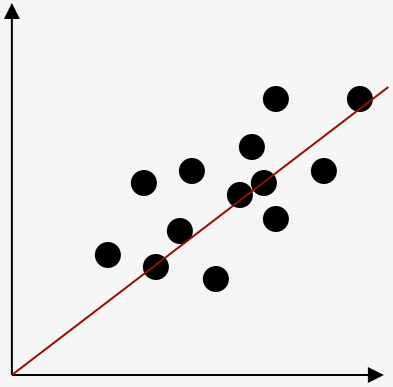
Stem:		levied
Choices:	(a)	imposed
	(b)	believed
	(c)	requested
	(d)	correlated
Solution:	(a)	imposed

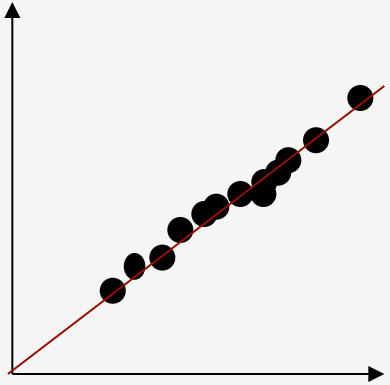
Stem:	mason:stone
Choices:	(a) teacher:chalk
	(b) carpenter:wood
	(c) soldier:gun
	(d) photograph:camera
	(e) book:word
Solution:	(b) carpenter:wood

Example from Peter Turney

Dimensionality Reduction

- Looking for hidden similarities in data
- Based on matrix decomposition
- Height/weight example





Vectors and Matrices

- A matrix is an *m* x *n* table of objects (in our case, numbers)
- Each row (or column) is a vector.
- Matrices of compatible dimensions can be multiplied together.
- What is the result of the multiplication below?

$$\begin{bmatrix} 1 & 2 & 4 \\ 2 & 5 & 7 \\ 4 & 9 & 14 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} \quad \\ \quad \\ \quad \end{bmatrix}$$

Answer to the Quiz

$$\begin{bmatrix} 1 & 2 & 4 \\ 2 & 5 & 7 \\ 4 & 9 & 14 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \times 2 + 2 \times 1 + 4 \times (-1) \\ 2 \times 2 + 5 \times 1 + 7 \times (-1) \\ 4 \times 2 + 9 \times 1 + 14 \times (-1) \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ 3 \end{bmatrix}$$

Eigenvectors and Eigenvalues

- An eigenvector is an implicit "direction" for a matrix $A\vec{v} = \lambda\vec{v}$
- v (the eigenvector) is non-zero, though λ (the eigenvalue) can be any complex number in principle.
- Computing eigenvalues: $\det(A - \lambda I) = 0$

Eigenvectors and Eigenvalues

- Example:

$$A = \begin{pmatrix} -1 & 3 \\ 2 & 0 \end{pmatrix} \qquad A - \lambda I = \begin{pmatrix} -1-\lambda & 3 \\ 2 & -\lambda \end{pmatrix}$$

- $\det(A-\lambda I) = (-1-\lambda)*(-\lambda)-3*2=0$
- Then: $\lambda+\lambda^2-6=0$; $\lambda_1=2$; $\lambda_2=-3$
- For $\lambda_1=2$:

$$\begin{pmatrix} -3 & 3 \\ 2 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0$$

- Solutions: $x_1=x_2$

Matrix Decomposition

- If Σ is a square matrix, it can be decomposed into $U\Lambda U^{-1}$, where

  U = matrix of eigenvectors

  Λ = diagonal matrix of eigenvalues

$$\Sigma U = U\Lambda$$

$$U^{-1}\Sigma U = \Lambda$$

$$\Sigma = U\Lambda U^{-1}$$

Example

$$S = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \lambda_1 = 1, \lambda_2 = 3$$

$$U = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$

$$U^{-1} = \begin{pmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{pmatrix}$$

$$S = U\Lambda U^{-1} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}\begin{pmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{pmatrix}$$

SVD: Singular Value Decomposition

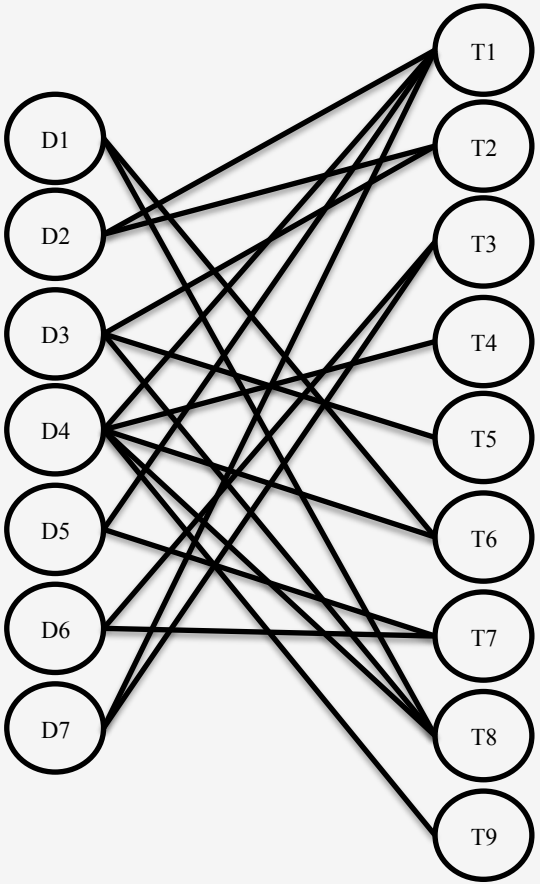
- $A=U\Sigma V^T$
  - U is the matrix of orthogonal eigenvectors of AA^T
  - V is the matrix of orthogonal eigenvectors of A^TA
  - The components of Σ are the eigenvalues of A^TA
- This decomposition exists for all matrices, dense or sparse
- If A has 5 columns and 3 rows, then U will be 5x5 and V will be 3x3
- In Matlab, use *[U,S,V] = svd (A)*

Example

D1: T6, T9
D2: T1, T2
D3: T2, T5, T8
D4: T1, T4, T6, T8, T9
D5: T1, T7
D6: T3, T7
D7: T1, T3

Example

D1: T6, T9
D2: T1, T2
D3: T2, T5, T8
D4: T1, T4, T6, T8, T9
D5: T1, T7
D6: T3, T7
D7: T1, T3



Document-Term Matrix

raw

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

normalized

$$A^{(n)} = \begin{bmatrix} 0 & 0.58 & 0 & 0.45 & 0.71 & 0 & 0.71 \\ 0 & 0.58 & 0.58 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.71 & 0.71 \\ 0 & 0 & 0 & 0.45 & 0 & 0 & 0 \\ 0 & 0.58 & 0.58 & 0 & 0 & 0 & 0 \\ 0.71 & 0 & 0 & 0.45 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.71 & 0.71 & 0 \\ 0 & 0 & 0.58 & 0.45 & 0 & 0 & 0 \\ 0.71 & 0 & 0 & 0.45 & 0 & 0 & 0 \end{bmatrix}$$

Decomposition

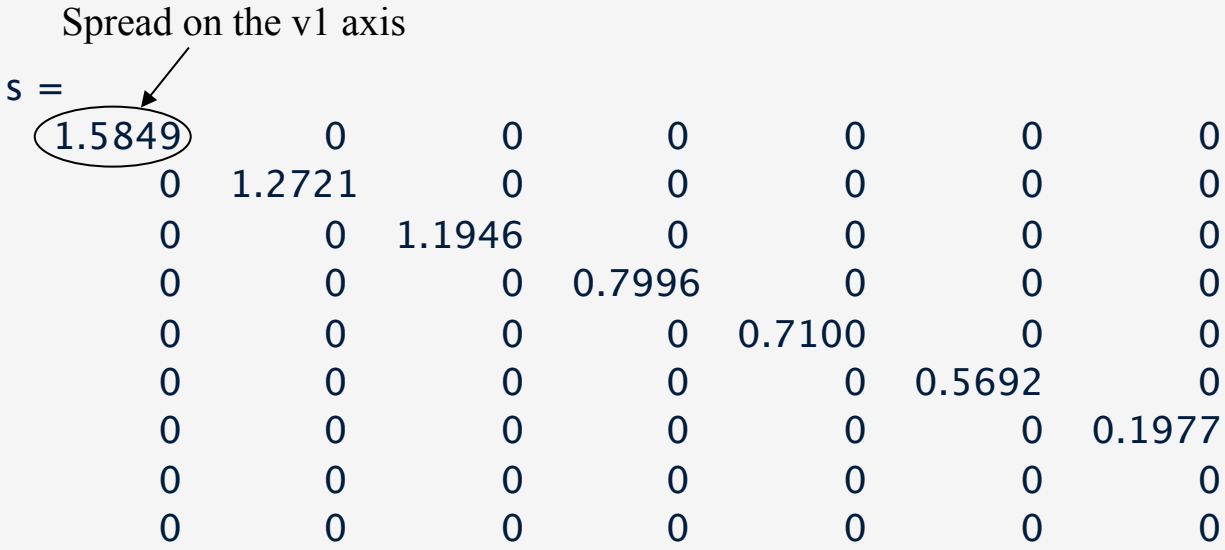
u =

-0.6976	-0.0945	0.0174	-0.6950	0.0000	0.0153	0.1442	-0.0000	0
-0.2622	0.2946	0.4693	0.1968	-0.0000	-0.2467	-0.1571	-0.6356	0.3098
-0.3519	-0.4495	-0.1026	0.4014	0.7071	-0.0065	-0.0493	-0.0000	0.0000
-0.1127	0.1416	-0.1478	-0.0734	0.0000	0.4842	-0.8400	0.0000	-0.0000
-0.2622	0.2946	0.4693	0.1968	0.0000	-0.2467	-0.1571	0.6356	-0.3098
-0.1883	0.3756	-0.5035	0.1273	-0.0000	-0.2293	0.0339	-0.3098	-0.6356
-0.3519	-0.4495	-0.1026	0.4014	-0.7071	-0.0065	-0.0493	0.0000	-0.0000
-0.2112	0.3334	0.0962	0.2819	-0.0000	0.7338	0.4659	-0.0000	0.0000
-0.1883	0.3756	-0.5035	0.1273	-0.0000	-0.2293	0.0339	0.3098	0.6356

v =

-0.1687	0.4192	-0.5986	0.2261	0	-0.5720	0.2433
-0.4472	0.2255	0.4641	-0.2187	0.0000	-0.4871	-0.4987
-0.2692	0.4206	0.5024	0.4900	-0.0000	0.2450	0.4451
-0.3970	0.4003	-0.3923	-0.1305	0	0.6124	-0.3690
-0.4702	-0.3037	-0.0507	-0.2607	-0.7071	0.0110	0.3407
-0.3153	-0.5018	-0.1220	0.7128	-0.0000	-0.0162	-0.3544
-0.4702	-0.3037	-0.0507	-0.2607	0.7071	0.0110	0.3407

Decomposition



Related to Dimensionality Reduction

- Singular value decomposition:
  - $A=U\Sigma V^T$
- Dimensionality reduction
  - $A^*=U\Sigma^*V^T$
  - Where Σ^* keeps only the largest eigenvalues

Rank-4 Approximation

s4 =

1.5849	0	0	0	0	0	0
0	1.2721	0	0	0	0	0
0	0	1.1946	0	0	0	0
0	0	0	0.7996	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

Rank-4 Approximation

```
u*s4*v'
   -0.0019    0.5985   -0.0148    0.4552    0.7002    0.0102    0.7002
   -0.0728    0.4961    0.6282    0.0745    0.0121   -0.0133    0.0121
    0.0003   -0.0067    0.0052   -0.0013    0.3584    0.7065    0.3584
    0.1980    0.0514    0.0064    0.2199    0.0535   -0.0544    0.0535
   -0.0728    0.4961    0.6282    0.0745    0.0121   -0.0133    0.0121
    0.6337   -0.0602    0.0290    0.5324   -0.0008    0.0003   -0.0008
    0.0003   -0.0067    0.0052   -0.0013    0.3584    0.7065    0.3584
    0.2165    0.2494    0.4367    0.2282   -0.0360    0.0394   -0.0360
    0.6337   -0.0602    0.0290    0.5324   -0.0008    0.0003   -0.0008
```

Rank-4 Approximation

```
u*s4

   -1.1056   -0.1203    0.0207   -0.5558         0         0         0
   -0.4155    0.3748    0.5606    0.1573         0         0         0
   -0.5576   -0.5719   -0.1226    0.3210         0         0         0
   -0.1786    0.1801   -0.1765   -0.0587         0         0         0
   -0.4155    0.3748    0.5606    0.1573         0         0         0
   -0.2984    0.4778   -0.6015    0.1018         0         0         0
   -0.5576   -0.5719   -0.1226    0.3210         0         0         0
   -0.3348    0.4241    0.1149    0.2255         0         0         0
   -0.2984    0.4778   -0.6015    0.1018         0         0         0
```

Rank-4 Approximation

```
s4*v'

   -0.2674   -0.7087   -0.4266   -0.6292   -0.7451   -0.4996   -0.7451
    0.5333    0.2869    0.5351    0.5092   -0.3863   -0.6384   -0.3863
   -0.7150    0.5544    0.6001   -0.4686   -0.0605   -0.1457   -0.0605
    0.1808   -0.1749    0.3918   -0.1043   -0.2085    0.5700   -0.2085
         0         0         0         0         0         0         0
         0         0         0         0         0         0         0
         0         0         0         0         0         0         0
         0         0         0         0         0         0         0
         0         0         0         0         0         0         0
```

Rank-2 Approximaton

```
s2 =
   1.5849         0         0         0         0         0         0
        0    1.2721         0         0         0         0         0
        0         0         0         0         0         0         0
        0         0         0         0         0         0         0
        0         0         0         0         0         0         0
        0         0         0         0         0         0         0
        0         0         0         0         0         0         0
        0         0         0         0         0         0         0
        0         0         0         0         0         0         0
```

Rank-2 Approximation

```
u*s2*v'

    0.1361    0.4673    0.2470    0.3908    0.5563    0.4089    0.5563
    0.2272    0.2703    0.2695    0.3150    0.0815   -0.0571    0.0815
   -0.1457    0.1204   -0.0904   -0.0075    0.4358    0.4628    0.4358
    0.1057    0.1205    0.1239    0.1430    0.0293   -0.0341    0.0293
    0.2272    0.2703    0.2695    0.3150    0.0815   -0.0571    0.0815
    0.2507    0.2412    0.2813    0.3097   -0.0048   -0.1457   -0.0048
   -0.1457    0.1204   -0.0904   -0.0075    0.4358    0.4628    0.4358
    0.2343    0.2454    0.2685    0.3027    0.0286   -0.1073    0.0286
    0.2507    0.2412    0.2813    0.3097   -0.0048   -0.1457   -0.0048
```

Rank-2 Approximation

u*s2 – word vector representation in concept space

-1.1056	-0.1203	0	0	0	0	0
-0.4155	0.3748	0	0	0	0	0
-0.5576	-0.5719	0	0	0	0	0
-0.1786	0.1801	0	0	0	0	0
-0.4155	0.3748	0	0	0	0	0
-0.2984	0.4778	0	0	0	0	0
-0.5576	-0.5719	0	0	0	0	0
-0.3348	0.4241	0	0	0	0	0
-0.2984	0.4778	0	0	0	0	0

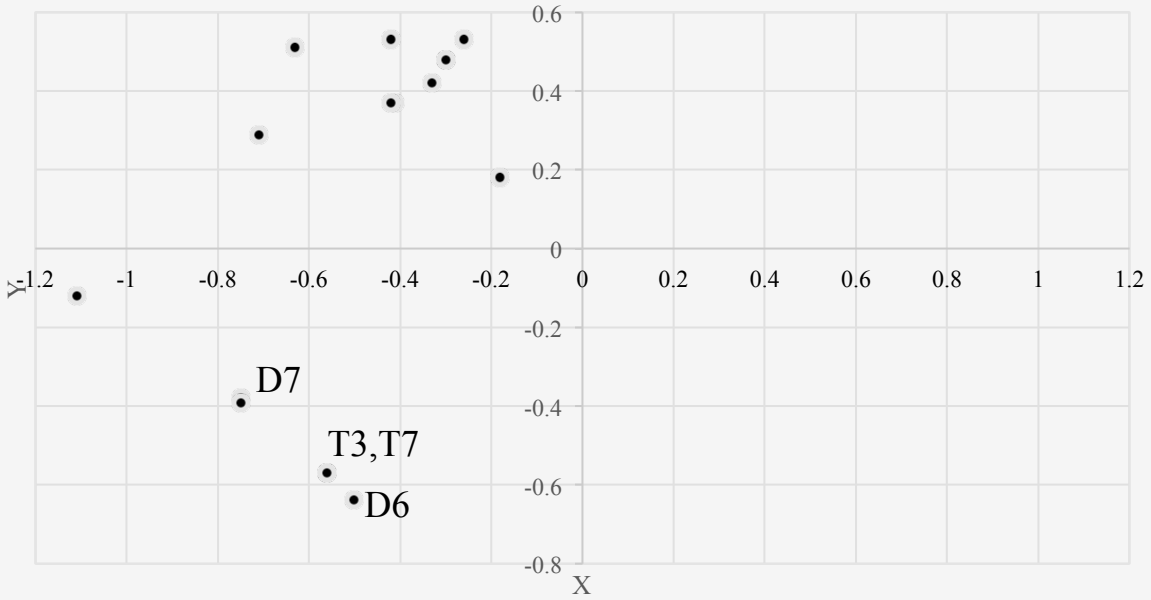
Rank-2 Approximation

s2*v' – new concept representation of the documents

-0.2674	-0.7087	-0.4266	-0.6292	-0.7451	-0.4996	-0.7451
0.5333	0.2869	0.5351	0.5092	-0.3863	-0.6384	-0.3863
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

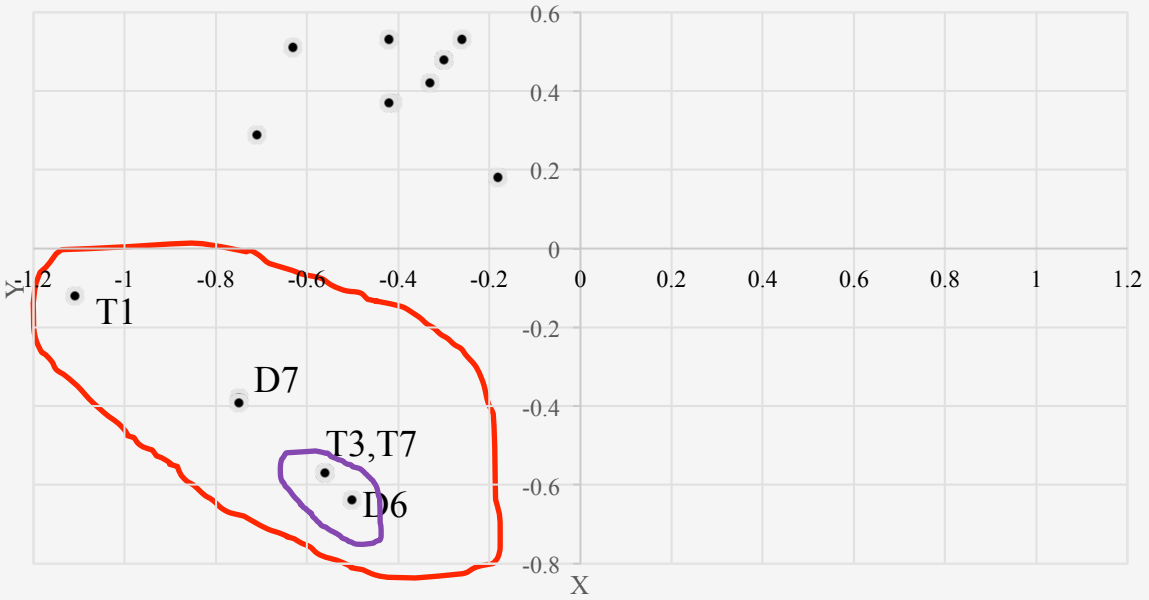
D1	D2	D3	D4	D5	D6	D7
-0.26	-0.71	-0.42	-0.63	-0.75	-0.5	-0.75
0.53	0.29	0.53	0.51	-0.38	-0.64	-0.39

T1	T2	T3	T4	T5	T6	T7	T8	T9
-1.11	-0.41	-0.56	-0.18	-0.42	-0.3	-0.56	-0.33	-0.3
-0.12	0.37	-0.57	0.18	0.37	0.48	-0.57	0.42	0.48



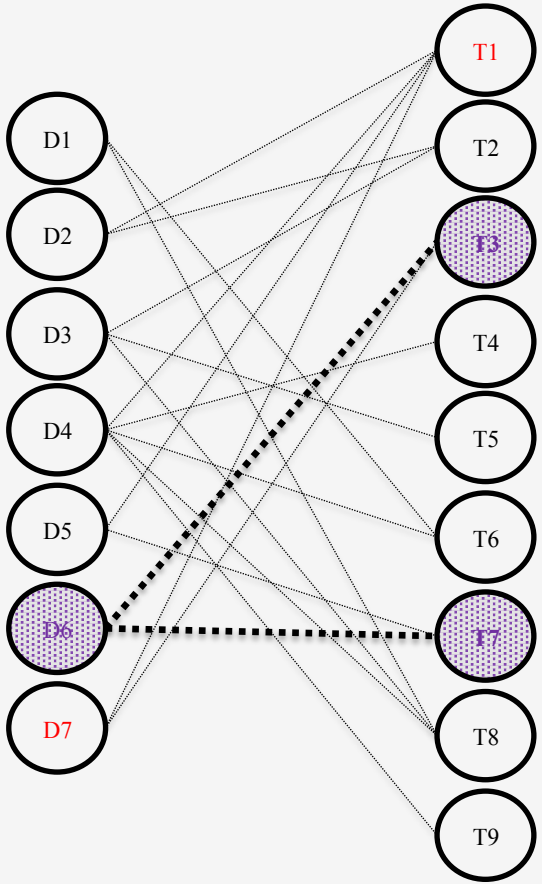
D1	D2	D3	D4	D5	D6	D7
-0.26	-0.71	-0.42	-0.63	-0.75	-0.5	-0.75
0.53	0.29	0.53	0.51	-0.38	-0.64	-0.39

T1	T2	T3	T4	T5	T6	T7	T8	T9
-1.11	-0.41	-0.56	-0.18	-0.42	-0.3	-0.56	-0.33	-0.3
-0.12	0.37	-0.57	0.18	0.37	0.48	-0.57	0.42	0.48



Example

D1: T6, T9
D2: T1, T2
D3: T2, T5, T8
D4: T1, T4, T6, T8, T9
D5: T1, T7
D6: T3, T7
D7: T1, T3



Documents to Concepts and Terms to Concepts

```
>> A(:,1)'*u*s
-0.4238    0.6784   -0.8541    0.1446   -0.0000   -0.1853    0.0095

>> A(:,1)'*u*s4
-0.4238    0.6784   -0.8541    0.1446         0         0         0

>> A(:,1)'*u*s2
-0.4238    0.6784         0         0         0         0         0

>> A(:,2)'*u*s2
-1.1233    0.3650         0         0         0         0         0

>> A(:,3)'*u*s2
-0.6762    0.6807         0         0         0         0         0
```

Documents to Concepts and Terms to Concepts

```
>> A(:,4)'*u*s2
-0.9972    0.6478         0         0         0         0         0

>> A(:,5)'*u*s2
-1.1809   -0.4914         0         0         0         0         0

>> A(:,6)'*u*s2
-0.7918   -0.8121         0         0         0         0         0

>> A(:,7)'*u*s2
-1.1809   -0.4914         0         0         0         0         0
```

Cont'd

```
>> (s2*v'*A(1,:)')'
-1.7523   -0.1530         0         0         0         0         0         0         0

>> (s2*v'*A(2,:)')'
-0.6585    0.4768         0         0         0         0         0         0         0

>> (s2*v'*A(3,:)')'
-0.8838   -0.7275         0         0         0         0         0         0         0

>> (s2*v'*A(4,:)')'
-0.2831    0.2291         0         0         0         0         0         0         0

>> (s2*v'*A(5,:)')'
-0.6585    0.4768         0         0         0         0         0         0         0
```

Cont'd

```
>> (s2*v'*A(6,:)')'

   -0.4730    0.6078         0         0         0         0         0         0         0

>> (s2*v'*A(7,:)')'

   -0.8838   -0.7275         0         0         0         0         0         0         0

>> (s2*v'*A(8,:)')'

   -0.5306    0.5395         0         0         0         0         0         0         0

>> (s2*v'*A(9,:)')'

   -0.4730    0.6078         0         0         0         0         0         0         0
```

Properties

A is a document to term matrix. What is A*A'?

A*A'

1.5471	0.3364	0.5041	0.2025	0.3364	0.2025	0.5041	0.2025	0.2025
0.3364	0.6728	0	0	0.6728	0	0	0.3364	0
0.5041	0	1.0082	0	0	0	0.5041	0	0
0.2025	0	0	0.2025	0	0.2025	0	0.2025	0.2025
0.3364	0.6728	0	0	0.6728	0	0	0.3364	0
0.2025	0	0	0.2025	0	0.7066	0	0.2025	0.7066
0.5041	0	0.5041	0	0	0	1.0082	0	0
0.2025	0.3364	0	0.2025	0.3364	0.2025	0	0.5389	0.2025
0.2025	0	0	0.2025	0	0.7066	0	0.2025	0.7066

Properties

What about A'*A?

A'*A

1.0082	0	0	0.6390	0	0	0
0	1.0092	0.6728	0.2610	0.4118	0	0.4118
0	0.6728	1.0092	0.2610	0	0	0
0.6390	0.2610	0.2610	1.0125	0.3195	0	0.3195
0	0.4118	0	0.3195	1.0082	0.5041	0.5041
0	0	0	0	0.5041	1.0082	0.5041
0	0.4118	0	0.3195	0.5041	0.5041	1.0082

Latent Semantic Indexing (LSI)

- Dimensionality reduction = identification of hidden (latent) concepts
- Query matching in latent space

External Pointers

- http://lsa.colorado.edu
- http://www.cs.utk.edu/~lsi

NLP