What is a retrieval model?

- Model is an idealization or abstraction of an actual process (retrieval)
- Mathematical models are used to study the properties of the process, draw conclusions, make predictions
- Conclusions derived from a model depend on whether the model is a good approximation to the actual situation
- Statistical models represent repetitive processes , make predictions about frequencies of interesting events, use probability as the fundamental tool
- Retrieval models can describe the computational process
  - e.g. how documents are ranked
  - note that inverted file is an implementation not a model
- Retrieval models can attempt to describe the human process
  - e.g. the information need, ASK
- Retrieval variables: queries, documents, terms, relevance judgements, users, information needs
- Retrieval models have an explicit or implicit definition of relevance

Vector Space Retrieval Model: Introduction

Overview:

- **Variations:**
  - Vector space retrieval model
  - Extended Boolean model
  - Latent Semantic Indexing
- **Key idea: Everything (documents, queries, terms) is a vector in a high-dimensional space**
- **Example system: SMART, developed by Salton and students at Cornell starting in the 60's.**

Vector Space Retrieval Model: Introduction

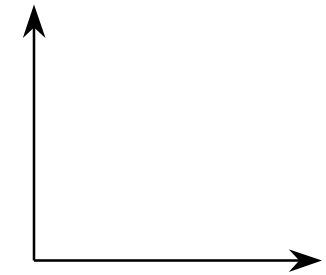
- **How are documents represented in a binary model?**

TermID:	**T_1**	**T_2**	**T_3**	**T_4**	**…**	**T_n**
D1:	**1**	**0**	**0**	**1**	**…**	**1**
D2:	**0**	**1**	**1**	**0**	**…**	**0**
D3:	**1**	**0**	**1**	**0**	**…**	**0**
:	**:**	**:**	**:**	**:**	**:**	**:**

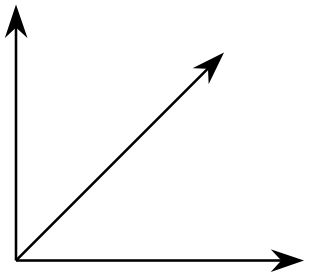
  A document is represented as a vector of terms.
- **A query can also be represented as a vector of terms.**
  - For now, assume no AND, OR, NOT, etc, query operators.
- **Linear algebra can be used to discover which document vectors are the most similar to the query vector.**

Vector Space Representation

- **Formally, a *vector space* is defined by a set of *linearly independent* basis vectors.**
- **Basis vectors:**
  - correspond to the *dimensions* or *directions* in the vector space;
  - determine what can be described in the vector space; and
  - must be *orthogonal*, or *linearly independent*, i.e. a value along one dimension implies nothing about a value along another.



Basis vectors
for 2 dimensions



Basis vectors
for 3 dimensions

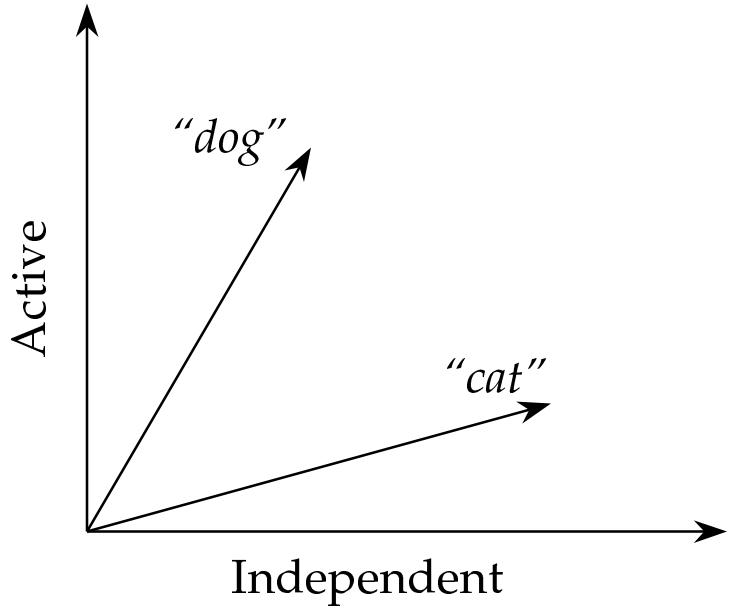
Vector Space Representation

How do basis vectors relate to terms?

- **Each term is represented as a linear combination of basis vectors.**

Dictionary

Term	Active	Independent
cat	0.25	0.75
dog	0.75	0.25
gravity	?	?



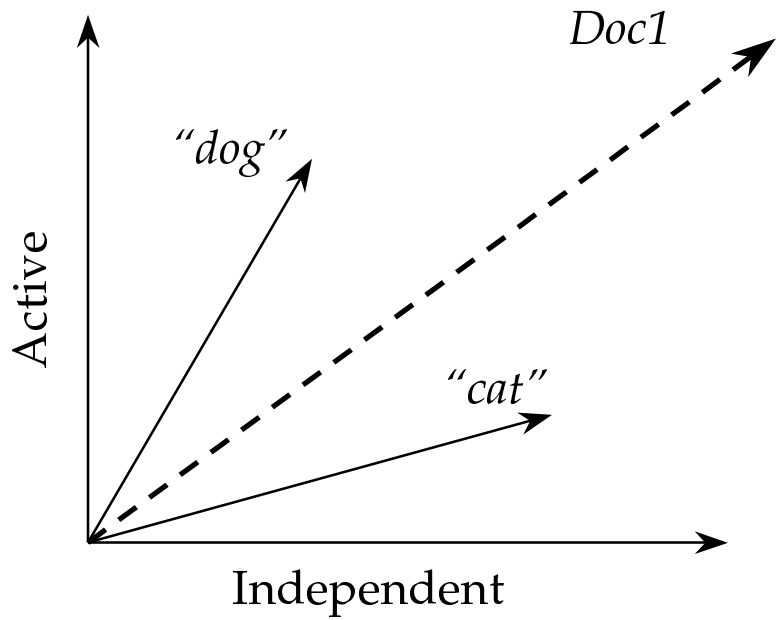
Vector Space Representation

How are documents represented?

- **A document is represented as the sum of its term vectors.**

Doc1

dog
cat



Vector Space Representation

What should be the basis vectors for information retrieval?

- **"Core" concepts:**
  - difficult to determine (Philosophy? Cognitive science?)
  - orthogonal (by definition)
  - a relatively static vector space ("there are no new ideas")
- **Terms:**
  - easy to determine
  - not *really* orthogonal (orthogonal enough?)
  - a constantly growing vector space (new vocabulary)

Vector Space Similarity: Weighted Features Example

$D_1 = (3T_1 + 1T_2 + 4T_3)$
$D_2 = (8T_2 + 2T_2 + 6T_3)$

$Q = (0T_1 + 2T_2 + 0T_3)$

Correlated Terms

Term	Cat	Dog	Lion
cat	1.00	-0.20	0.50
dog	-0.20	1.00	-0.40
lion	0.50	-0.40	1.00

Orthogonal Terms

Term	Cat	Dog	Lion
cat	1.00	0.00	0.00
dog	0.00	1.00	0.00
lion	0.00	0.00	1.00

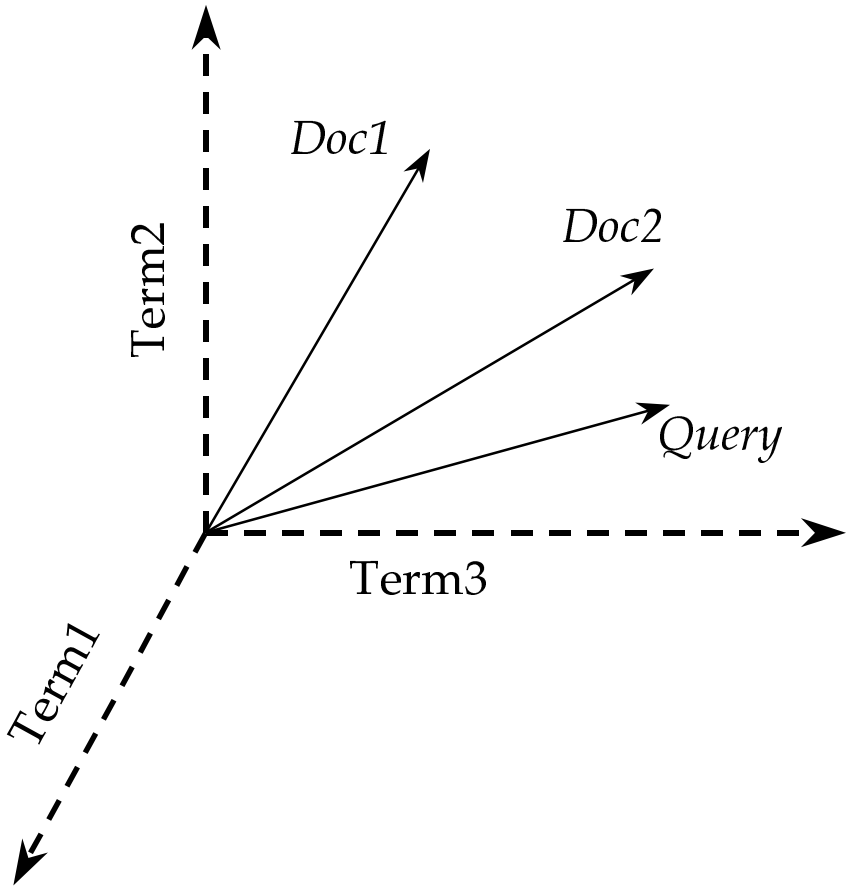
$$\begin{aligned}\text{Sim}(D_1,Q) &= (3T_1 + 1T_2 + 4T_3) \bullet (2T_2) \\ &= 6T_1 \bullet T_2 + 2T_2 \bullet T_2 + 8T_3 \bullet T_2 \\ &= -6 \bullet 0.2 + 2 \bullet 1 - 8 \bullet 0.4 \\ &= -1.2 + 2 - 3.2 \\ &= -2.4\end{aligned}$$

$$\begin{aligned}\text{Sim}(D_1,Q) &= 3 \bullet 0 + 1 \bullet 2 + 4 \bullet 0 \\ &= 2\end{aligned}$$

Vector Space Representation: Vector Coefficients

- **The coefficients (vector elements, term weights) represent term presence, importance, or "aboutness"**
- **The model gives no guidance on how to set term weights**
- **Some common choices:**
  - Binary: 1 = term is present, 0 = term not present in document
  - *tf*: The frequency of the term in the document
  - *tf* • *idf*: *idf* indicates the discriminatory power of the term

Vector Space Similarity



Similarity is
inversely related
to the angle
between the vectors.

Doc2 is the
most similar
to the *Query*.

Rank the documents
by their similarity
to the *Query*.

Vector Space Similarity: Common Measures

Sim(X,Y)	Binary Term Vectors	Weighted Term Vectors
Inner product	$\lvert X \cap Y \rvert$	$\sum x_i . y_i$
Dice coefficient	$\frac{2 \lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert}$	$\frac{2 \sum x_i . y_i}{\sum x_i^2 + \sum y_i^2}$
Cosine coefficient	$\frac{\lvert X \cap Y \rvert}{\sqrt{\lvert X \rvert} \sqrt{\lvert Y \rvert}}$	$\frac{\sum x_i . y_i}{\sqrt{\sum x_i^2 \cdot \sum y_i^2}}$
Jaccard coefficient	$\frac{\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert - \lvert X \cap Y \rvert}$	$\frac{\sum x_i . y_i}{\sum x_i^2 + \sum y_i^2 - \sum x_i . y_i}$

Vector Space Similarity:
Cosine Coefficient (Correlation) Example

$D_1 = (0.5T_1 + 0.8T_2 + 0.3T_3)$ $\quad$ $Q = (1.5T_1 + 1T_2 + 0T_3)$

$$\text{Sim}(D_1, Q) = \frac{(0.5 \times 1.5) + (0.8 \times 1)}{\sqrt{(0.5^2 + 0.8^2 + 0.3^2)(1.5^2 + 1^2)}}$$

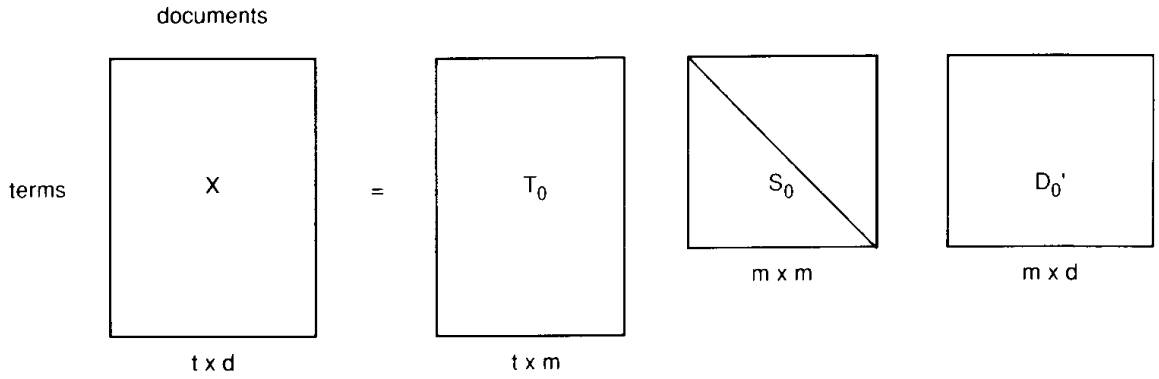
$$= \frac{1.55}{\sqrt{.98 \times 3.25}}$$

$$= .868$$

Vector Space Representation:
Latent Semantic Indexing (LSI)

- **Use Singular Value Decomposition (a dimensionality reduction technique) to identify uncorrelated, significant basis vectors or factors**
- **Replace original words with a subset of the new factors (say 100) in both documents and queries**
- **Compute similarities in this new space**
- **Computationally expensive, uncertain effectiveness**

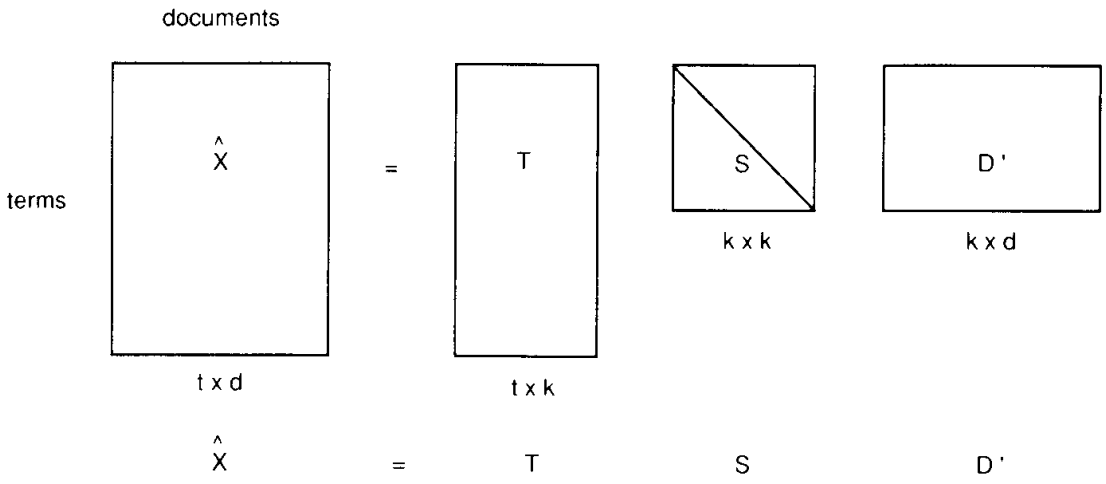
LSI



Singular value decomposition of the term x document matrix, X. Where:

T_0 has orthogonal, unit-length columns (T_0' T_0 = I)
D_0 has orthogonal, unit-length columns (D_0' D_0 = I)
S_0 is the diagonal matrix of singular values
t is the number of rows of X
d is the number of columns of X
m is the rank of X ($\leq$ min(t,d))

LSI



Reduced singular value decomposition of the term x document matrix, X. Where:

T has orthogonal, unit-length columns (T' T = I)
D has orthogonal, unit-length columns (D' D = I)
S is the diagonal matrix of singular values

t is the number of rows of X
d is the number of columns of X
m is the rank of X ($\leq$ min(t,d))
k is the chosen number of dimensions in the reduced model (k $\leq$ m)

LSI: example

Technical Memo Example

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*

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

LSI: example

$T_0 =$

0.22	−0.11	0.29	−0.41	−0.11	−0.34	0.52	−0.06	−0.41
0.20	−0.07	0.14	−0.55	0.28	0.50	−0.07	−0.01	−0.11
0.24	0.04	−0.16	−0.59	−0.11	−0.25	−0.30	0.06	0.49
0.40	0.06	−0.34	0.10	0.33	0.38	0.00	0.00	0.01
0.64	−0.17	0.36	0.33	−0.16	−0.21	−0.17	0.03	0.27
0.27	0.11	−0.43	0.07	0.08	−0.17	0.28	−0.02	−0.05
0.27	0.11	−0.43	0.07	0.08	−0.17	0.28	−0.02	−0.05
0.30	−0.14	0.33	0.19	0.11	0.27	0.03	−0.02	−0.17
0.21	0.27	−0.18	−0.03	−0.54	0.08	−0.47	−0.04	−0.58
0.01	0.49	0.23	0.03	0.59	−0.39	−0.29	0.25	−0.23
0.04	0.62	0.22	0.00	−0.07	0.11	0.16	−0.68	0.23
0.03	0.45	0.14	−0.01	−0.30	0.28	0.34	0.68	0.18

$S_0 =$

3.34								
	2.54							
		2.35						
			1.64					
				1.50				
					1.31			
						0.85		
							0.56	
								0.36

$D_0 =$

0.20	−0.06	0.11	−0.95	0.05	−0.08	0.18	−0.01	−0.06
0.61	0.17	−0.50	−0.03	−0.21	−0.26	−0.43	0.05	0.24
0.46	−0.03	0.21	0.04	0.38	0.72	−0.24	0.01	0.02
0.54	−0.23	0.57	0.27	−0.21	−0.37	0.26	−0.02	−0.08
0.28	0.11	−0.51	0.15	0.33	0.03	0.67	−0.06	−0.26
0.00	0.19	0.10	0.02	0.39	−0.30	−0.34	0.45	−0.62
0.01	0.44	0.19	0.02	0.35	−0.21	−0.15	−0.76	0.02
0.02	0.62	0.25	0.01	0.15	0.00	0.25	0.45	0.52
0.08	0.53	0.08	−0.03	−0.60	0.36	−0.04	−0.07	−0.45

LSI: example

$X \approx$

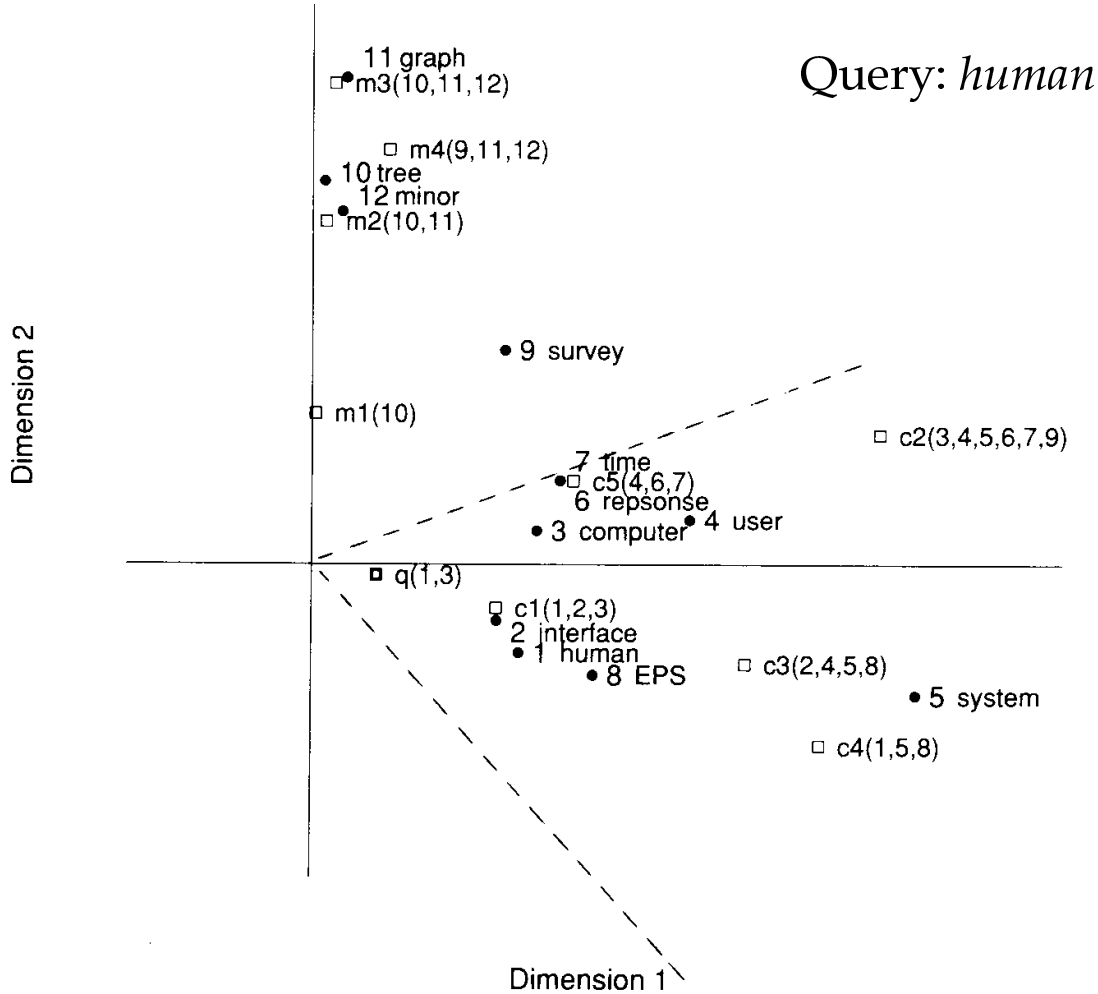
T		S	D'								
0.22	−0.11	3.34	0.20	0.61	0.46	0.54	0.28	0.00	0.02	0.02	0.08
0.20	−0.07	2.54	−0.06	0.17	−0.13	−0.23	0.11	0.19	0.44	0.62	0.53
0.24	0.04										
0.40	0.06										
0.64	−0.17										
0.27	0.11										
0.27	0.11										
0.30	−0.14										
0.21	0.27										
0.01	0.49										
0.04	0.62										
0.03	0.45										

LSI: example

$\hat{X} =$

0.16	0.40	0.38	0.47	0.18	−0.05	−0.12	−0.16	−0.09
0.14	0.37	0.33	0.40	0.16	−0.03	−0.07	−0.10	−0.04
0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
0.45	1.23	1.05	1.27	0.56	−0.07	−0.15	−0.21	−0.05
0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
0.22	0.55	0.51	0.63	0.24	−0.07	−0.14	−0.20	−0.11
0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
−0.06	0.23	−0.14	−0.27	0.14	0.24	0.55	0.77	0.66
−0.06	0.34	−0.15	−0.30	0.20	0.31	0.69	0.98	0.85
−0.04	0.25	−0.10	−0.21	0.15	0.22	0.50	0.71	0.62

LSI



Query: *human computer interaction*

Vector Space Retrieval Model: Summary

- **Standard vector space**
  - each dimension corresponds to a term in the vocabulary
  - vector elements are real-valued, reflecting term importance
  - any vector (document,query, ...) can be compared to any other
  - cosine correlation is the similarity metric used most often
- **Extended Boolean**
  - multiple nested similarity measures
- **Latent Semantic Indexing (LSI)**
  - each dimension corresponds to a “basic concept”

Vector Space Retrieval Model: Disadvantages

- **Assumed independence relationship among terms**
- **Lack of justification for some vector operations**
  - e.g. choice of similarity function
  - e.g., choice of term weights
- **Barely a retrieval model; doesn't explicitly model relevance, a person's information need, language models, etc.**
- **Assumes a query and a document can be treated the same (symmetric)**
- **Lack of a cognitive (or other) justification**

Vector Space Retrieval Model: Advantages

- **Simplicity**
- **Ability to incorporate term weights**
- **Ability to handle "distributed" term representations (e.g., LSI)**
- **Can measure similarities between almost anything:**
  - documents and queries, documents and documents, queries and queries, sentences and sentences, etc.