# Latent Semantic Analysis (LSA)

Monica Gavrila
gavrila@nats.informatik.uni-hamburg.de

13. June 2006

# Contents

- Latent Semantic Analysis Algorithm Description
- PROLIV Presentation
- Discussions

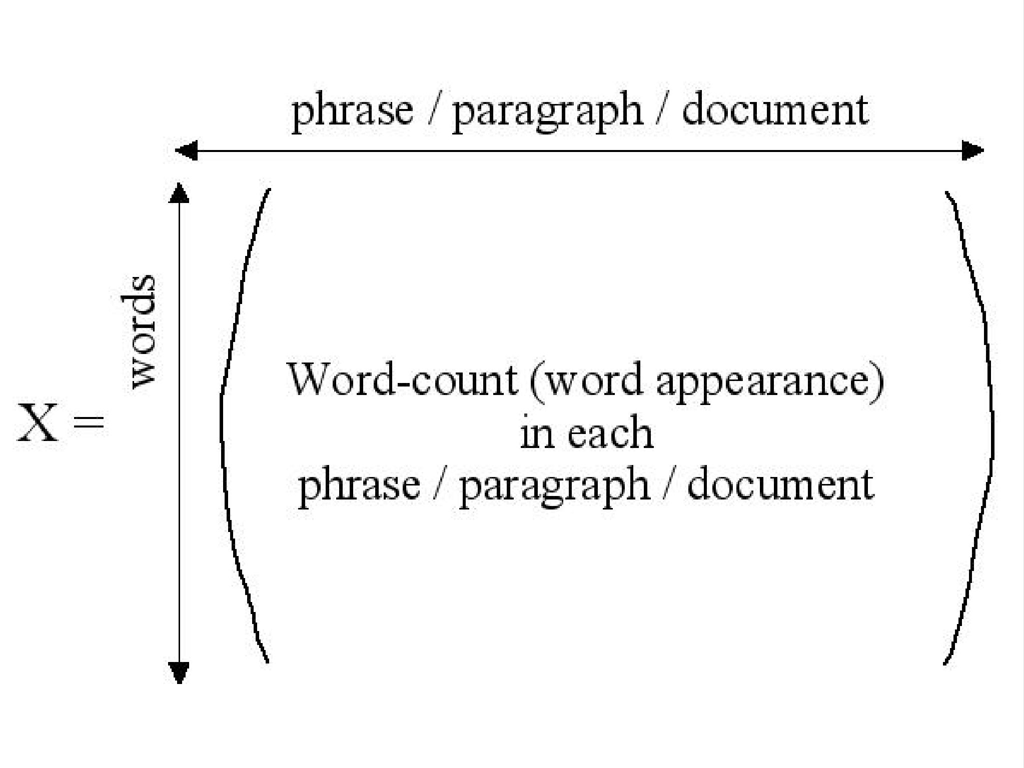
# What is LSA?

- LSA is a fully automatic statistics-algebraic technique for extracting and inferring relations of expected contextual usage of words in documents
- It uses no humanly constructed dictionaries, knowledge bases, semantic networks, parsers, morphology, grammars
- Motivation: finding similarity between words, texts

# Method:
# Co-occurrence Matrix

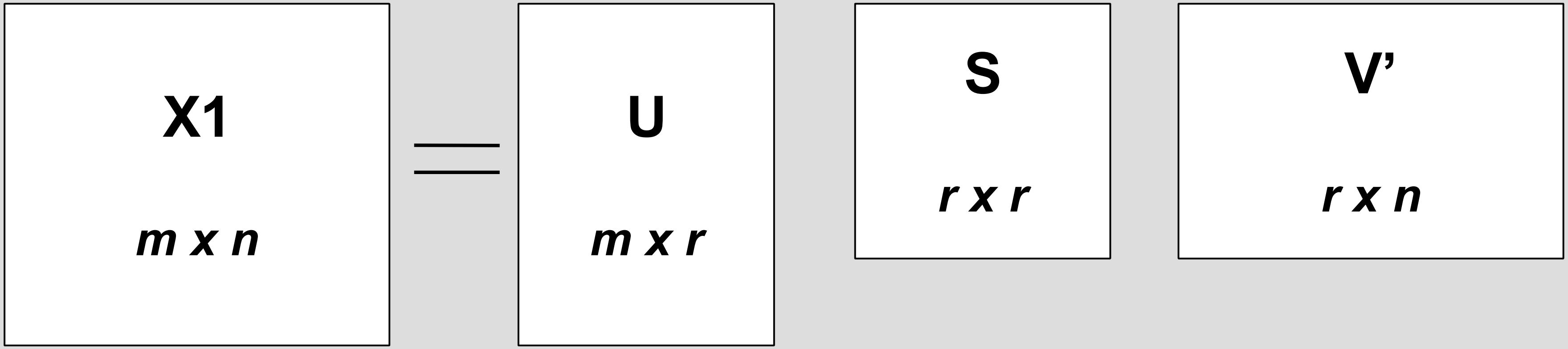
- Takes as input row text
  - text segmented in words
  - text segmented in passages
- The text „is introduced“

in a matrix


phrase / paragraph / document
X =
words
Word-count (word appearance)
in each
phrase / paragraph / document

# Method: Singular Value Decomposition

- The matrix is normalized (weighted) – not always
- Matrix decomposed (Singular Value Decomposition)



X1

*m x n*

=

U

*m x r*

S

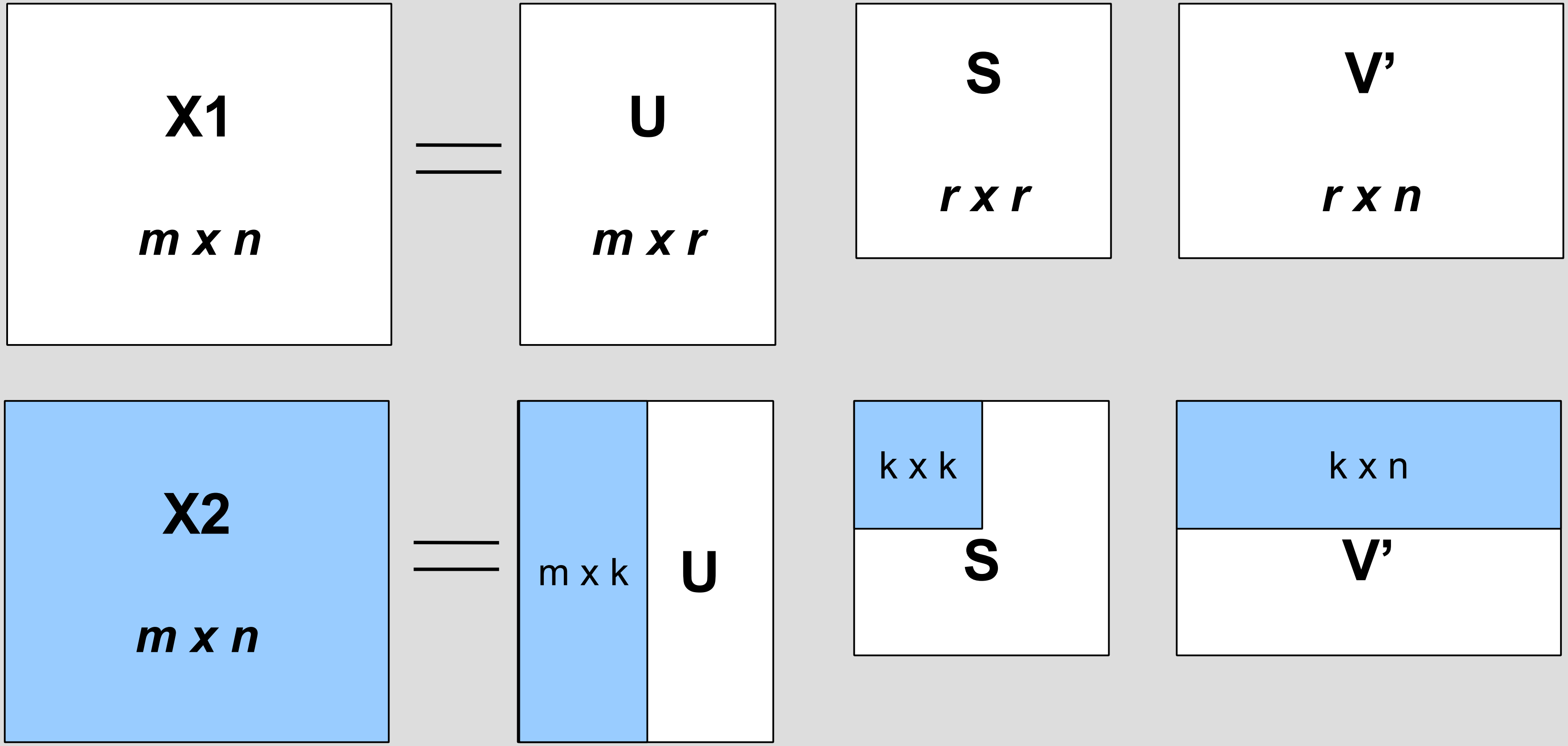
*r x r*

V’

*r x n*

# Method: Dimension Reduction

- Dimension reduction
  - X2 is an approximation of X1



# Method: Calculating Similarity

- Calculating similarity measures
  - Cosine
  - ..........
- Obtaining similarity results
  - Word - word
  - Word – passage
  - Passage - passage

# Example - Corpus

- c1: Human machine interface for 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, ordered trees
- m2: The intersection graph paths in trees
- m3: Graph minors IV: Widths of trees and well-quasi-ordering
- m4: Graph minor: A survey

Taken from Landauer et al., 1998

# Example - Terms Considered

- c1: **Human** machine **interface** for 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, ordered **trees**
- m2: The intersection **graph** paths in **trees**
- m3: **Graph minors** IV: Widths of **trees** and well-quasi-ordering
- m4: **Graph minors**: A **survey**

**Words (appear 2 times):** human, interface, computer, user system, response, time, EPS, survey, trees, graph, minors.

Taken from Landauer et al., 1998

# Example - Passages Considered

- **c1**: Human machine interface for 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, ordered trees
- **m2:** The intersection graph paths in trees
- **m3:** Graph minors IV: Widths of trees and well-quasi-ordering
- **m4:** Graph minors: A survey

**Text passages:** c1, c2, c3, c4, c5, m1, m2, m3, m4.

Taken from Landauer et al., 1998

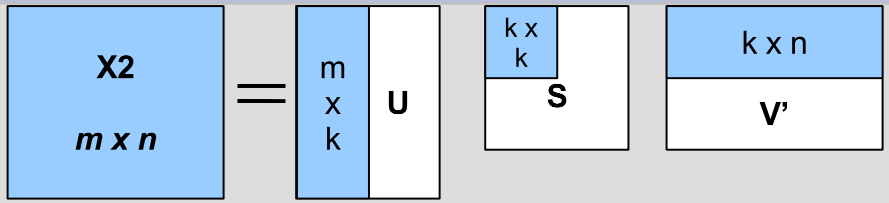
# Example - Co-occurrence Matrix

$$X=\begin{array}{lccccccccc}
 & \mathbf{c1} & \mathbf{c2} & \mathbf{c3} & \mathbf{c4} & \mathbf{c5} & \mathbf{m1} & \mathbf{m2} & \mathbf{m3} & \mathbf{m4} \\
\textbf{human} & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
\textbf{interface} & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
\textbf{computer} & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\textbf{user} & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
\textbf{system} & 0 & 1 & 1 & 2 & 0 & 0 & 0 & 0 & 0 \\
\textbf{response} & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
\textbf{time} & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
\textbf{EPS} & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
\textbf{survey} & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
\textbf{trees} & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\
\textbf{graph} & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
\textbf{minors} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1
\end{array}$$

Taken from Landauer et al., 1998

# Example - Reduced Matrix

After SVD and dimension reduction:



X2=

| | c1 | c2 | c3 | c4 | c5 | m1 | m2 | m3 | m4 |
|---|---|---|---|---|---|---|---|---|---|
| **human** | 0.16 | 0.4 | 0.38 | 0.47 | 0.18 | -0.05 | -0.12 | -0.16 | -0.09 |
| **interface** | 0.14 | 0.37 | 0.33 | 0.4 | 0.16 | -0.03 | -0.07 | -0.1 | -0.04 |
| **computer** | 0.15 | 0.51 | 0.36 | 0.41 | 0.24 | 0.02 | 0.06 | 0.09 | 0.12 |
| **user** | 0.26 | 0.84 | 0.61 | 0.7 | 0.39 | 0.03 | 0.08 | 0.12 | 0.19 |
| **system** | 0.45 | 1.23 | 1.05 | 1.27 | 0.56 | -0.07 | -0.15 | -0.21 | -0.05 |
| **response** | 0.16 | 0.58 | 0.38 | 0.42 | 0.28 | 0.06 | 0.13 | 0.19 | 0.22 |
| **time** | 0.16 | 0.58 | 0.38 | 0.42 | 0.28 | 0.06 | 0.13 | 0.19 | 0.22 |
| **EPS** | 0.22 | 0.55 | 0.51 | 0.63 | 0.24 | -0.07 | -0.14 | -0.2 | -0.11 |
| **survey** | 0.1 | 0.53 | 0.23 | 0.21 | 0.27 | 0.14 | 0.31 | 0.44 | 0.42 |
| **trees** | -0.06 | 0.23 | -0.14 | -0.27 | 0.14 | 0.24 | 0.55 | 0.77 | 0.66 |
| **graph** | -0.06 | 0.34 | -0.15 | -0.3 | 0.2 | 0.31 | 0.69 | 0.98 | 0.85 |
| **minors** | -0.04 | 0.25 | -0.1 | -0.21 | 0.15 | 0.22 | 0.5 | 0.71 | 0.62 |

K=2

Taken from Landauer et al., 1998

# Example - Interesting results

m1: The generation of random, binary, ordered **trees**
m2: The intersection **graph** paths in **trees**
m3: **Graph minors** IV: Widths of **trees** and well-quasi-ordering
m4: **Graph minors**: A **survey**

| | c1 | c2 | c3 | c4 | c5 | m1 | m2 | m3 | m4 |
|---|---|---|---|---|---|---|---|---|---|
| **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 |

| | c1 | c2 | c3 | c4 | c5 | m1 | m2 | m3 | m4 |
|---|---|---|---|---|---|---|---|---|---|
| **survey** | 0.1 | 0.53 | 0.23 | 0.21 | 0.27 | 0.14 | 0.31 | 0.44 | 0.42 |
| **trees** | -0.06 | 0.23 | -0.14 | -0.27 | 0.14 | 0.24 | 0.55 | 0.77 | 0.66 |
| **graph** | -0.06 | 0.34 | -0.15 | -0.3 | 0.2 | 0.31 | 0.69 | 0.98 | 0.85 |
| **minors** | -0.04 | 0.25 | -0.1 | -0.21 | 0.15 | 0.22 | 0.5 | 0.71 | 0.62 |

# Example - Similarity Measures – Unreduced Case

| | 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 |

R=-0.38

R=-0.29

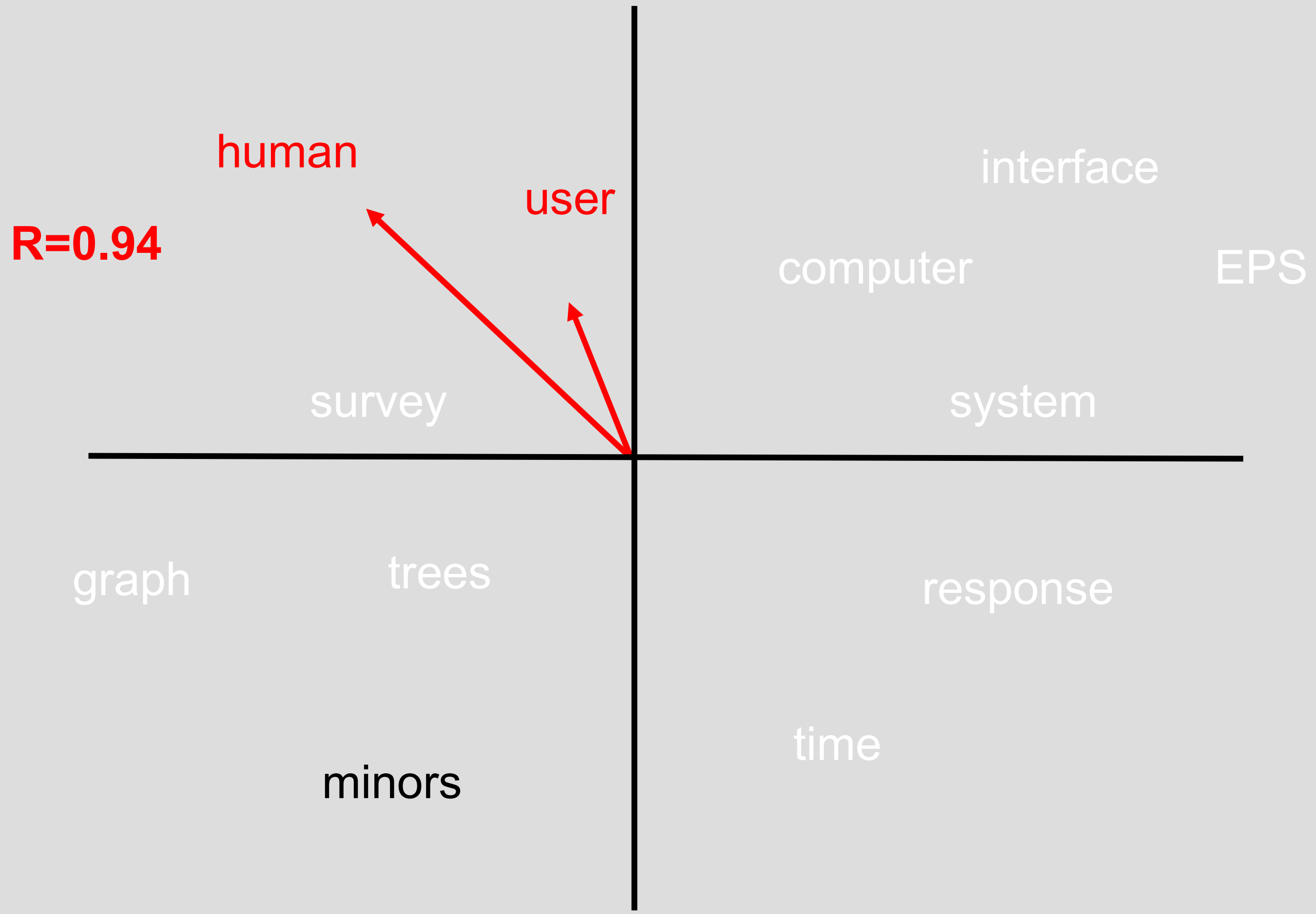
# Example- Similarity Measures – Reduced Case

| | c1 | c2 | c3 | c4 | c5 | m1 | m2 | m3 | m4 |
|---|---|---|---|---|---|---|---|---|---|
| **human** | 0.16 | 0.4 | 0.38 | 0.47 | 0.18 | -0.05 | -0.12 | -0.16 | -0.09 |
| **interface** | 0.14 | 0.37 | 0.33 | 0.4 | 0.16 | -0.03 | -0.07 | -0.1 | -0.04 |
| **computer** | 0.15 | 0.51 | 0.36 | 0.41 | 0.24 | 0.02 | 0.06 | 0.09 | 0.12 |
| **user** | 0.26 | 0.84 | 0.61 | 0.7 | 0.39 | 0.03 | 0.08 | 0.12 | 0.19 |
| **system** | 0.45 | 1.23 | 1.05 | 1.27 | 0.56 | -0.07 | -0.15 | -0.21 | -0.05 |
| **response** | 0.16 | 0.58 | 0.38 | 0.42 | 0.28 | 0.06 | 0.13 | 0.19 | 0.22 |
| **time** | 0.16 | 0.58 | 0.38 | 0.42 | 0.28 | 0.06 | 0.13 | 0.19 | 0.22 |
| **EPS** | 0.22 | 0.55 | 0.51 | 0.63 | 0.24 | -0.07 | -0.14 | -0.2 | -0.11 |
| **survey** | 0.1 | 0.53 | 0.23 | 0.21 | 0.27 | 0.14 | 0.31 | 0.44 | 0.42 |
| **trees** | -0.06 | 0.23 | -0.14 | -0.27 | 0.14 | 0.24 | 0.55 | 0.77 | 0.66 |
| **graph** | -0.06 | 0.34 | -0.15 | -0.3 | 0.2 | 0.31 | 0.69 | 0.98 | 0.85 |
| **minors** | -0.04 | 0.25 | -0.1 | -0.21 | 0.15 | 0.22 | 0.5 | 0.71 | 0.62 |

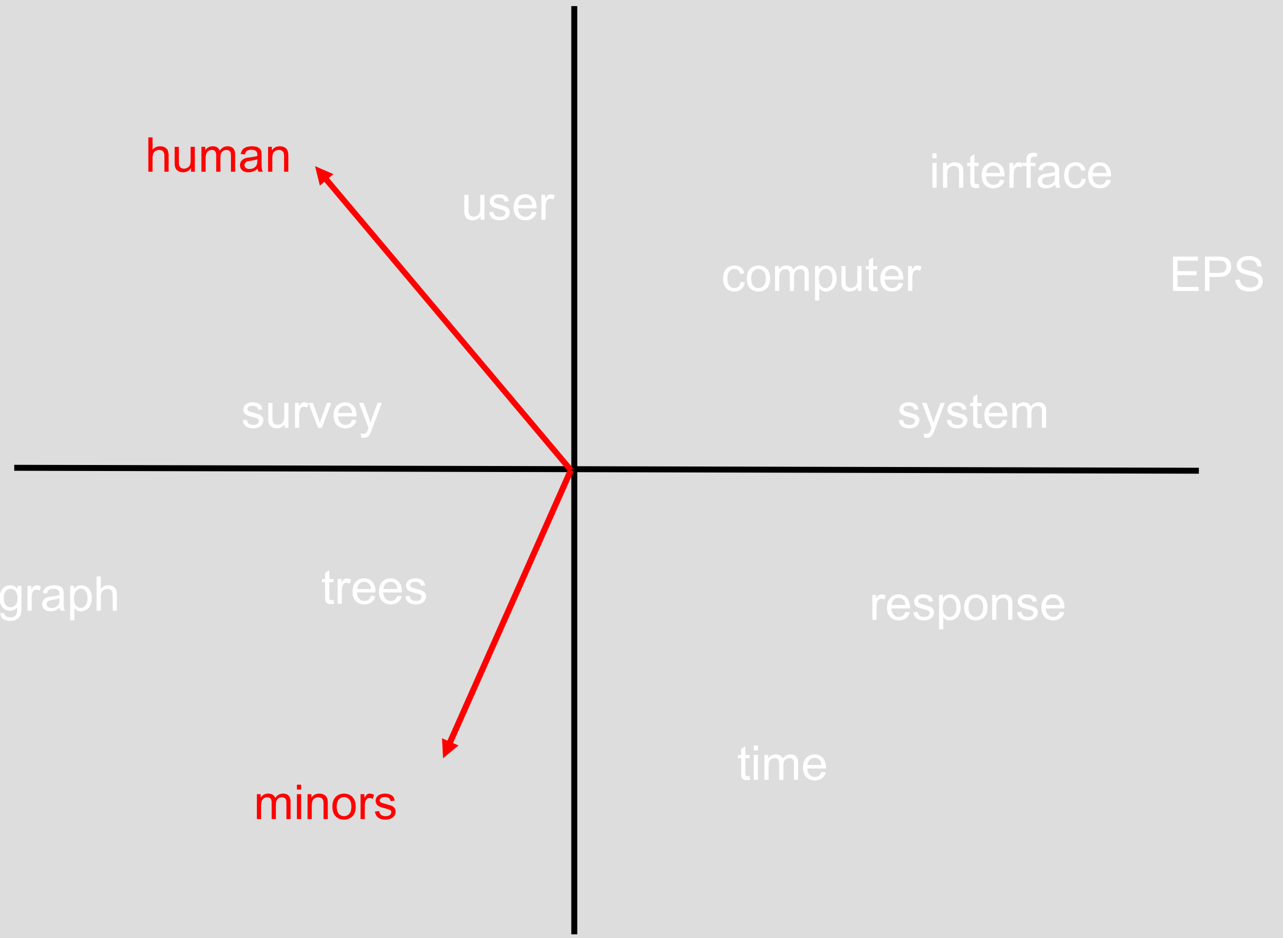
R=0.94

R=-0.83

# Example - Graphic Representation (human, user)



# Example - Graphic Representation (human, minors)



R=-0.83

# LSA‘s Ability to Model Human Conceptual Knowledge

- Predictor of query-document topic similarity judgments
- Simulation of agreed upon word-word relations and of human vocabulary test synonym judgments
- Simulation of human choices on subject-matter multiple-choice tests
- Predictor of text coherence and resulting comprehension
- Simulation of word-word, passage-word relations found in lexical priming experiments
- Predictor of subjective ratings of text properties
- Predictor of appropriate matches of instructional text to learners
- Used to simulate synonym, antonym, singular-plural and compound-compound word relations.

# What is LSA used for?

- Ability to model human conceptual knowledge
- Searching, information retrieval (queries and documents are in different language, or the same language), indexing (Latent Semantic Indexing - LSI)
- Semantic representation (text comparison – Foltz et al. 1996)
- Vocabulary acquisition (Landauer & Dumais, 1997)
- Text comprehension (Lemaire et al.)
- Free text assessment (Haley et al. 2005)

# LSA and PROLIV

http://lsa.colorado.edu

Run DEMO

Discussions

# Technology Overview

- Generated the word-passage co-occurance matrix
- Weight it
- Apply SVD
- Reduce the dimensions
- Find similarity