Word Meanings

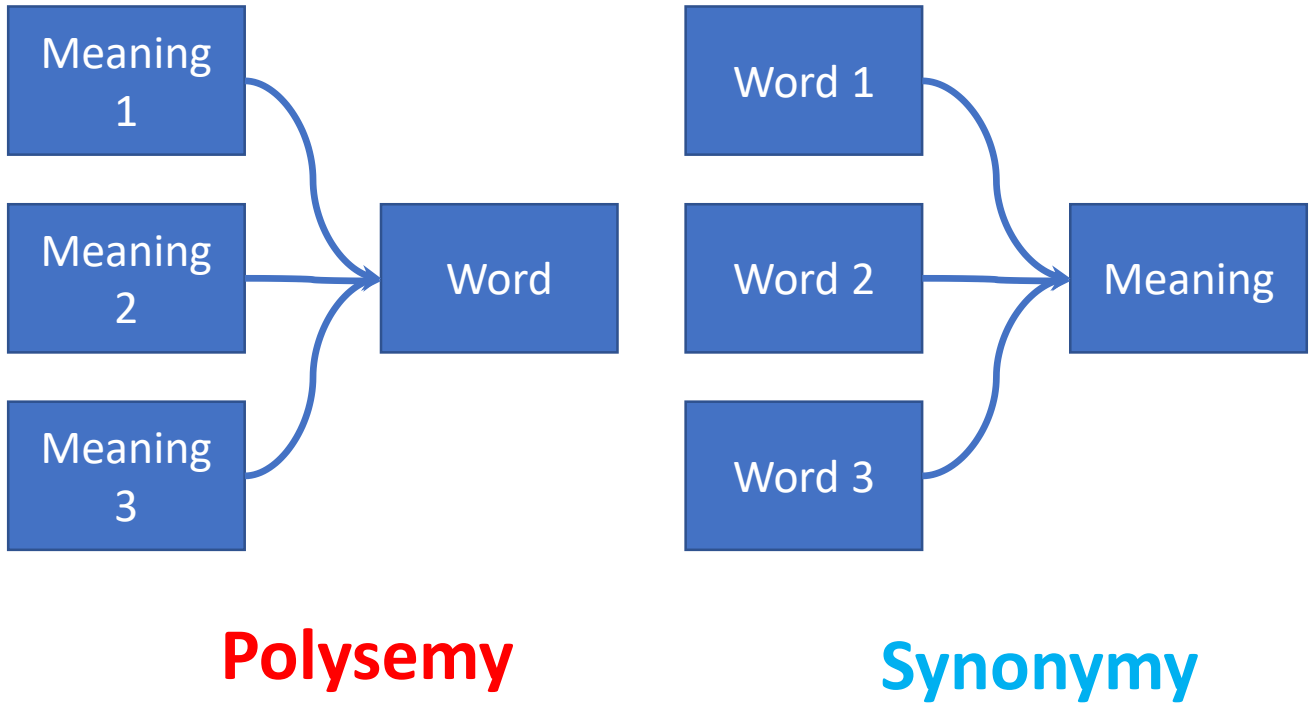
Latent Semantic Analytics

Dr. Uzair Ahmad

Program

- Word meanings and context
- The distributional hypothesis
- Design dimensions of word representations
- Latent semantic analysis
- Evaluation of representations

Word meanings: NLP Challenge



The distributional hypothesis

Acquire meaningful representations from unlabeled data

A bottle of ____ is on the table.

Everybody likes ____.

Don't have ____ before you drive.

We make ____ out of corn.

Lin, D. (1998). Automatic retrieval and clustering of similar words. See col (1998), pp.768–774.

The distributional hypothesis

(14.1) A bottle of ____ is on the table.

(14.2) Everybody likes _____.

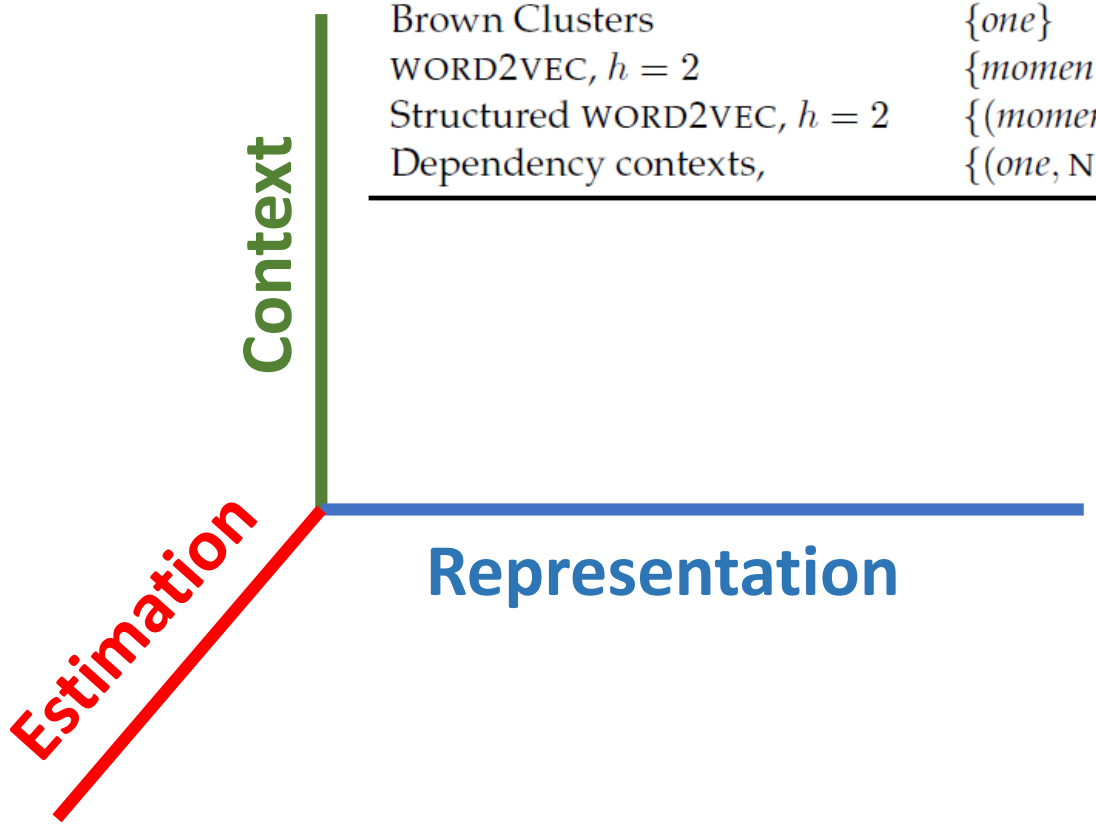
(14.3) Don't have ____ before you drive.

(14.4) We make _____ out of corn.

	contextual properties				
	(14.1)	(14.2)	(14.3)	(14.4)	...
tezgüino	1	1	1	1	
loud	0	0	0	0	
motor oil	1	0	0	1	
tortillas	0	1	0	1	
choices	0	1	0	0	
wine	1	1	1	0	

YOU SHALL KNOW A WORD BY THE COMPANY IT KEEPS.
(FIRTH 1957)

Word representations



Brown Clusters	$\{one\}$
WORD2VEC, $h = 2$	$\{moment, one, English, complications\}$
Structured WORD2VEC, $h = 2$	$\{(moment, -2), (one, -1), (English, +1), (complications, +2)\}$
Dependency contexts,	$\{(one, \text{NSUBJ}), (English, \text{DOBJ}), (moment, \text{ACL}^{-1})\}$

Latent Semantic Analysis

Contexts |C|

V Words	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

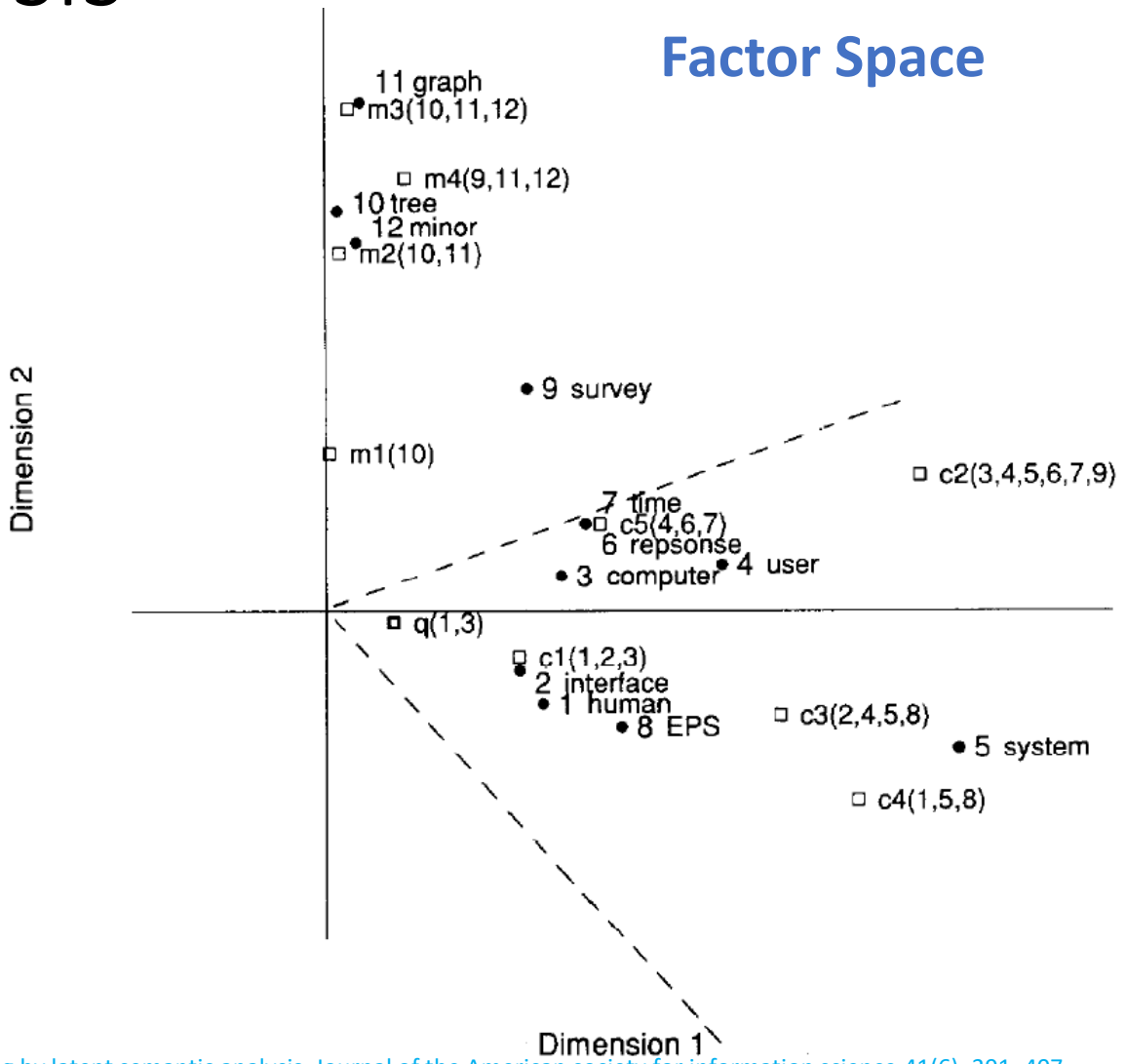
Term-Document Matrix

Deerwester, S. C., S. T. Dumais, T. K. Landauer, G.W. Furnas, and R. A. Harshman (1990). Indexing by latent semantic analysis. Journal of the American society for information science 41(6), 391–407.

Latent Semantic Analysis

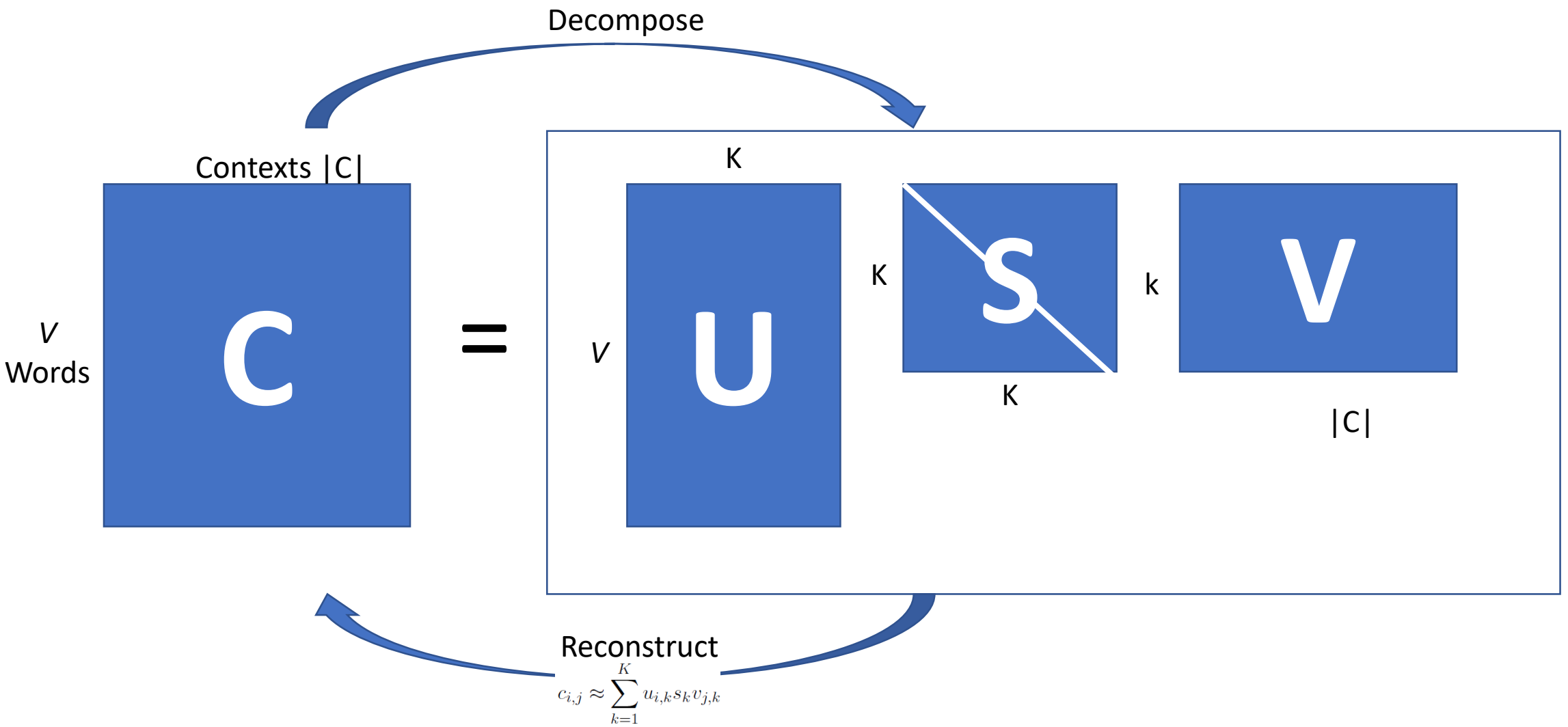
		c1	c2	c3	c4	c5	m1	m2	m3	m4
1	*human*	1	0	0	1	0	0	0	0	0
2	*interface*	1	0	1	0	0	0	0	0	0
3	*computer*	1	1	0	0	0	0	0	0	0
4	*user*	0	1	1	0	1	0	0	0	0
5	*system*	0	1	1	2	0	0	0	0	0
6	*response*	0	1	0	0	1	0	0	0	0
7	*time*	0	1	0	0	1	0	0	0	0
8	*EPS*	0	0	1	1	0	0	0	0	0
9	*survey*	0	1	0	0	0	0	0	0	1
10	*trees*	0	0	0	0	0	1	1	1	0
11	*graph*	0	0	0	0	0	0	1	1	1
12	*minors*	0	0	0	0	0	0	0	1	1

Term-Document Matrix



Deerwester, S. C., S. T. Dumais, T. K. Landauer, G.W. Furnas, and R. A. Harshman (1990). Indexing by latent semantic analysis. Journal of the American society for information science 41(6), 391–407.

Latent Semantic Analysis



Latent Semantic Analysis

$$\min_{\mathbf{U}\in\mathbb{R}^{V\times K},\mathbf{S}\in\mathbb{R}^{K\times K},\mathbf{V}\in\mathbb{R}^{|\mathcal{C}|\times K}} ||\mathbf{C}-\mathbf{U}\mathbf{S}\mathbf{V}^\top||_F$$

$$\text{s.t.} \quad \mathbf{U}^\top\mathbf{U}=\mathbb{I}$$

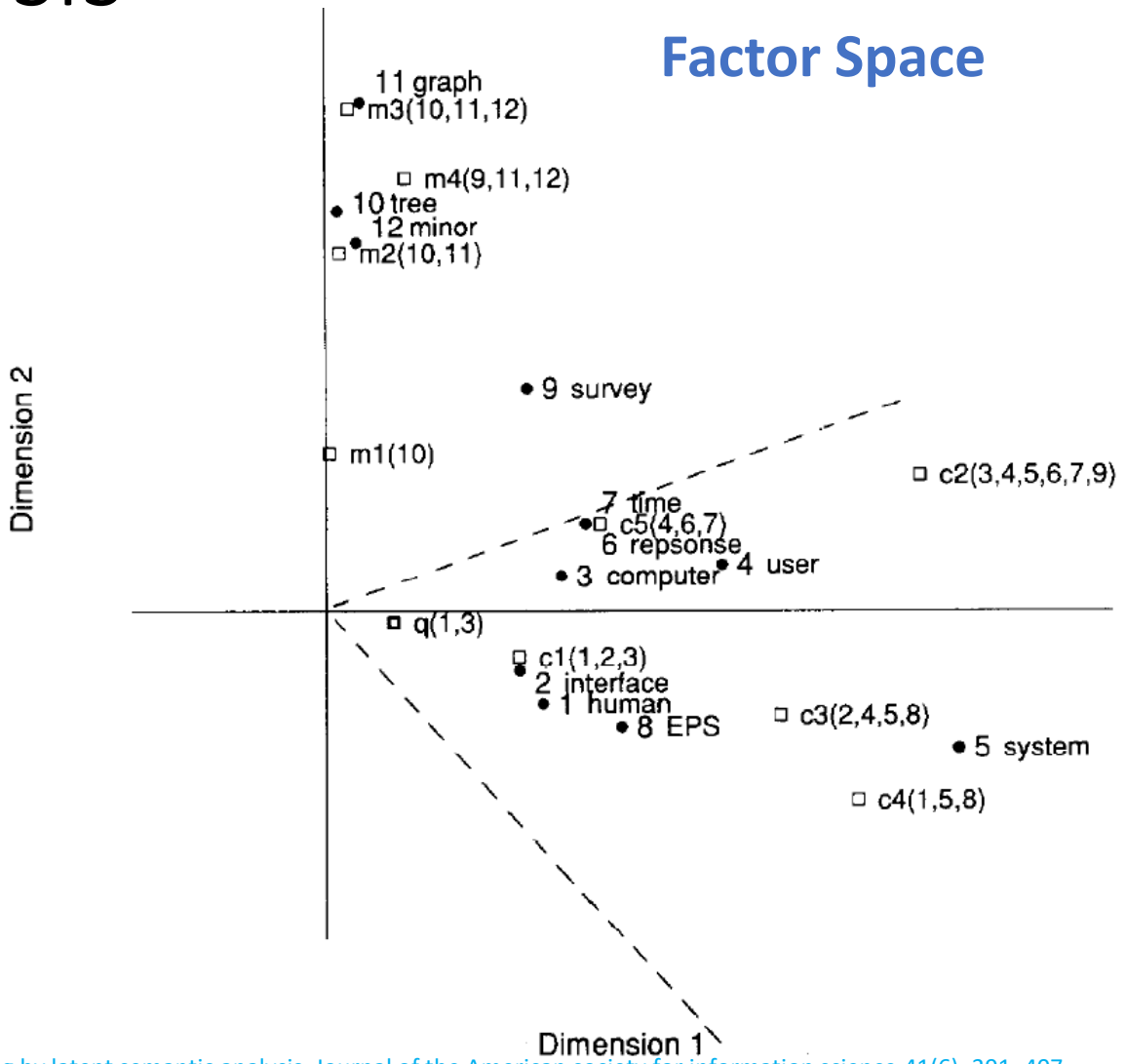
$$\mathbf{V}^\top\mathbf{V}=\mathbb{I}$$

$$\forall i\neq j, \mathbf{S}_{i,j}=0,$$

Latent Semantic Analysis

		c1	c2	c3	c4	c5	m1	m2	m3	m4
1	*human*	1	0	0	1	0	0	0	0	0
2	*interface*	1	0	1	0	0	0	0	0	0
3	*computer*	1	1	0	0	0	0	0	0	0
4	*user*	0	1	1	0	1	0	0	0	0
5	*system*	0	1	1	2	0	0	0	0	0
6	*response*	0	1	0	0	1	0	0	0	0
7	*time*	0	1	0	0	1	0	0	0	0
8	*EPS*	0	0	1	1	0	0	0	0	0
9	*survey*	0	1	0	0	0	0	0	0	1
10	*trees*	0	0	0	0	0	1	1	1	0
11	*graph*	0	0	0	0	0	0	1	1	1
12	*minors*	0	0	0	0	0	0	0	1	1

Term-Document Count Matrix



Deerwester, S. C., S. T. Dumais, T. K. Landauer, G.W. Furnas, and R. A. Harshman (1990). Indexing by latent semantic analysis. Journal of the American society for information science 41(6), 391–407.

Latent Semantic Analysis

- Transforming the count matrix **C**
  - Pointwise mutual information

$$\text{PMI}(i,j) = \log \frac{\text{p}(i,j)}{\text{p}(i)\text{p}(j)} = \log \frac{\text{p}(i \mid j)\text{p}(j)}{\text{p}(i)\text{p}(j)} = \log \frac{\text{p}(i \mid j)}{\text{p}(i)}$$

Latent Semantic Analysis

- Transforming the count matrix **C**
  - Positive Pointwise mutual information

$$\mathrm{PPMI}(i,j) = \begin{cases} \mathrm{PMI}(i,j), & \mathrm{p}(i \mid j) > \mathrm{p}(i) \\ 0, & \text{otherwise.} \end{cases}$$

Bullinaria, J. A. and J. P. Levy (2007). Extracting semantic representations from word cooccurrence statistics: A computational study. Behavior research methods 39(3), 510–526.

PMI

	computer	data	result	pie	sugar	count(w)
cherry	2	8	9	442	25	486
strawberry	0	0	1	60	19	80
digital	1670	1683	85	5	4	3447
information	3325	3982	378	5	13	7703
count(context)	4997	5673	473	512	61	11716

$$P(\text{w=information, c=data}) = \frac{3982}{11716} = .3399$$

$$P(\text{w=information}) = \frac{7703}{11716} = .6575$$

$$P(\text{c=data}) = \frac{5673}{11716} = .4842$$

$$\text{ppmi(information,data)} = \log 2(.3399/(.6575 * .4842)) = .0944$$

SLP: C6

Summary

- Distributed representations
  - Latent Semantic Analysis
  - Information retrieval
- Evaluation methods
  - Intrinsic
  - Extrinsic