Applied Natural Language Processing

Info 256

Lecture 9: Lexical semantics (Feb 19, 2019)

David Bamman, UC Berkeley

Lexical semantics

"You shall know a word by the company it keeps"

[Firth 1957]

DISTRIBUTIONAL STRUCTURE

Zellig S. Harris

(b) The fact that, for example, not every adjective occurs with every noun can be used as a measure of meaning difference. For it is not merely that different members of the one class have different selections of members of the other class with which they are actually found. More than that: if we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution.

Harris 1954

The distribution of an element will be understood as the sum of all its environments. An environment of an element A is an existing array of its co-occurrents, i.e. the other elements, each in a particular position, with which A occurs to yield an utterance. A's co-occurrents in a particular position are called its selection for that position.

Harris 1954

everyone likes	________________	
a bottle of	________________	is on the table
you can drink	________________	if you're over 21
a cocktail with	________________	and seltzer

Context

"You shall know a word by the company it keeps"

[Firth 1957]

- A few different ways we can encode the notion of "company" (or context).

context

everyone likes	______________	
a bottle of	______________	is on the table
you can drink	______________	if you’re over 21
a cocktail with	______________	and seltzer

Distributed representation

- Vector representation that encodes information about the distribution of contexts a word appears in

- Words that appear in similar contexts have similar representations (and similar meanings, by the distributional hypothesis).

Term-document matrix

	Hamlet	Macbeth	Romeo & Juliet	Richard III	Julius Caesar	Tempest	Othello	King Lear
knife	1	1	4	2		2		2
dog	2		6	6		2		12
sword	17	2	7	12		2		17
love	64		135	63		12		48
like	75	38	34	36	34	41	27	44

Context = appearing in the same document.

Vector

Vector representation of the document; vector size = V

Hamlet	King Lear
1	2
2	12
17	17
64	48
75	44

Vectors

knife	1	1	4	2		2		2
sword	17	2	7	12		2		17

Vector representation of the term; vector size = number of documents

Weighting dimensions

- Not all dimensions are equally informative

TF-IDF

- Term frequency-inverse document frequency
- A scaling to represent a feature as function of how frequently it appears in a data point but accounting for its frequency in the overall collection
- IDF for a given term = the number of documents in collection / number of documents that contain term

TF-IDF

- Term frequency ($tf_{t,d}$) = the number of times term t occurs in document d; several variants (e.g., passing through log function).

- Inverse document frequency = inverse fraction of number of documents containing (D_t) among total number of documents N

$$tfidf(t,d) = tf_{t,d} \times \log \frac{N}{D_t}$$

IDF

	Hamlet	Macbeth	Romeo & Juliet	Richard III	Julius Caesar	Tempest	Othello	King Lear
knife	1	1	4	2		2		2
dog	2		6	6		2		12
sword	17	2	7	12		2		17
love	64		135	63		12		48
like	75	38	34	36	34	41	27	44

IDF
0.12
0.20
0.12
0.20
0

IDF for the informativeness of the terms when comparing documents

PMI

- Mutual information provides a measure of how independent two variables (X and Y) are.
- Pointwise mutual information measures the independence of two outcomes (x and y)

PMI

$$\log_2 \frac{P(x,y)}{P(x)P(y)}$$

w = word, c = context

$$\log_2 \frac{P(w,c)}{P(w)P(c)}$$

What's this value for w and c that never occur together?

$$PPMI = \max\left(\log_2 \frac{P(w,c)}{P(w)P(c)}, 0\right)$$

	Hamlet	Macbeth	Romeo & Juliet	Richard III	Julius Caesar	Tempest	Othello	King Lear	total
knife	1	1	4	2		2		2	12
dog	2		6	6		2		12	28
sword	17	2	7	12		2		17	57
love	64		135	63		12		48	322
like	75	38	34	36	34	41	27	44	329
total	159	41	186	119	34	59	27	123	748

$$PMI(\text{love}, \text{R\&J}) = \frac{\frac{135}{748}}{\frac{186}{748} \times \frac{322}{748}}$$

Term-context matrix

- Rows and columns are both words; cell counts = the number of times word w_i and w_j show up in the same document.
- More common to define document = some smaller context (e.g., a window of 2 tokens)

Dataset

- the big dog ate dinner
- the small cat ate dinner
- the white dog ran down the street
- the yellow cat ran inside

DOG terms (window = 2)

the big ate dinner the white ran down

CAT terms (window = 2)

the small ate dinner the yellow ran inside

Term-context matrix

contexts

term

	the	big	ate	dinner	…
dog	2	1	1	1	…
cat	2	0	1	1	…

- Each cell enumerates the number of time a context word appeared in a window of 2 words around the term.

Term-context matrix

	aardvark	...	computer	data	pinch	result	sugar	...
apricot	0	...	0	0	1	0	1	
pineapple	0	...	0	0	1	0	1	
digital	0	...	2	1	0	1	0	
information	0	...	1	6	0	4	0	

Figure 15.4 Co-occurrence vectors for four words, computed from the Brown corpus, showing only six of the dimensions (hand-picked for pedagogical purposes). The vector for the word *digital* is outlined in red. Note that a real vector would have vastly more dimensions and thus be much sparser.

Jurafsky and Martin 2017

Dataset

- the big dog ate dinner
- the small cat ate dinner
- the white dog ran down the street
- the yellow cat ran inside

DOG terms (window = 2)

L: the big, R: ate dinner, L: the white, R: ran down

CAT terms (window = 2)

L: the small, R: ate dinner, L: the yellow, R: ran inside

Term-context matrix

contexts

term

	L: the big	R: ate dinner	L: the small	L: the yellow	…
dog	1	1	0	0	…
cat	0	1	1	1	…

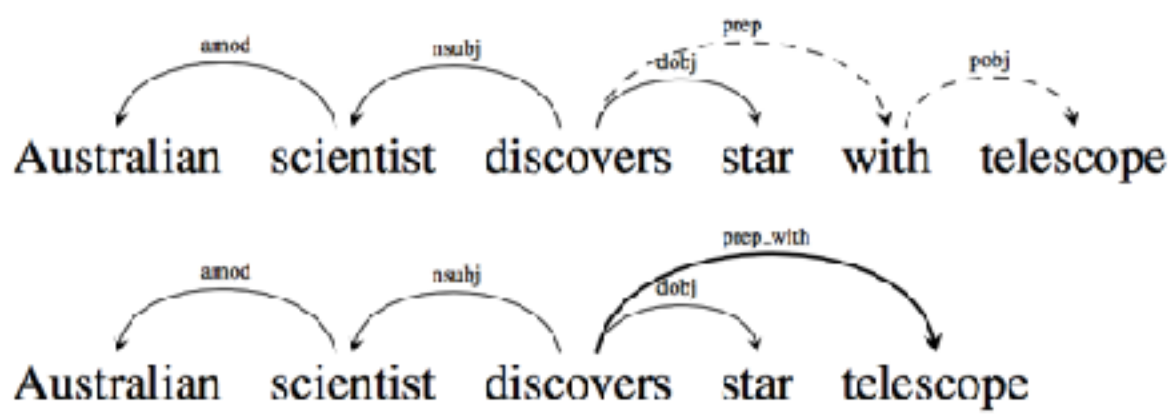
- Each cell enumerates the number of time a directional context phrase appeared in a specific position around the term.

write a book
write a poem

- First-order co-occurrence (syntagmatic association): write co-occurs with book in the same sentence.

- Second-order co-occurrence (paradigmatic association): book co-occurs with poem (since each co-occur with write)

Syntactic context



WORD	CONTEXTS
australian	scientist/amod^{-1}
scientist	australian/amod, discovers/nsubj^{-1}
discovers	scientist/nsubj, star/dobj, telescope/prep_with
star	discovers/dobj^{-1}
telescope	discovers/prep_with^{-1}

Target Word	BoW5	BoW2	DEPS
batman	nightwing aquaman catwoman superman manhunter	superman superboy aquaman catwoman batgirl	superman superboy supergirl catwoman aquaman
hogwarts	dumbledore hallows half-blood malfoy snape	evernight sunnydale garderobe blandings collinwood	sunnydale collinwood calarts greendale millfield
turing	nondeterministic non-deterministic computability deterministic finite-state	non-deterministic finite-state nondeterministic buchi primality	pauling hotelling heting lessing hamming
florida	gainesville fla jacksonville tampa lauderdale	fla alabama gainesville tallahassee texas	texas louisiana georgia california carolina
object-oriented	aspect-oriented smalltalk event-driven prolog domain-specific	aspect-oriented event-driven objective-c dataflow 4gl	event-driven domain-specific rule-based data-driven human-centered
dancing	singing dance dances dancers tap-dancing	singing dance dances breakdancing clowning	singing rapping breakdancing miming busking

Lin 1998; Levy and Goldberg 2014

Cosine Similarity

$$cos(x,y) = \frac{\sum_{i=1}^{F} x_i y_i}{\sqrt{\sum_{i=1}^{F} x_i^2}\sqrt{\sum_{i=1}^{F} y_i^2}}$$

- We can calculate the cosine similarity of two vectors to judge the degree of their similarity [Salton 1971]

- Euclidean distance measures the magnitude of distance between two points
- Cosine similarity measures their orientation

Intrinsic Evaluation

- Relatedness: correlation (Spearman/Pearson) between vector similarity of pair of words and human judgments

word 1	word 2	human score
midday	noon	9.29
journey	voyage	9.29
car	automobile	8.94
…	…	…
professor	cucumber	0.31
king	cabbage	0.23

WordSim-353 (Finkelstein et al. 2002)

Intrinsic Evaluation

- Analogical reasoning (Mikolov et al. 2013). For analogy Germany : Berlin :: France : ???, find closest vector to v("Berlin") - v("Germany") + v("France")

			target
possibly	impossibly	certain	uncertain
generating	generated	shrinking	shrank
think	thinking	look	looking
Baltimore	Maryland	Oakland	California
shrinking	shrank	slowing	slowed
Rabat	Morocco	Astana	Kazakhstan

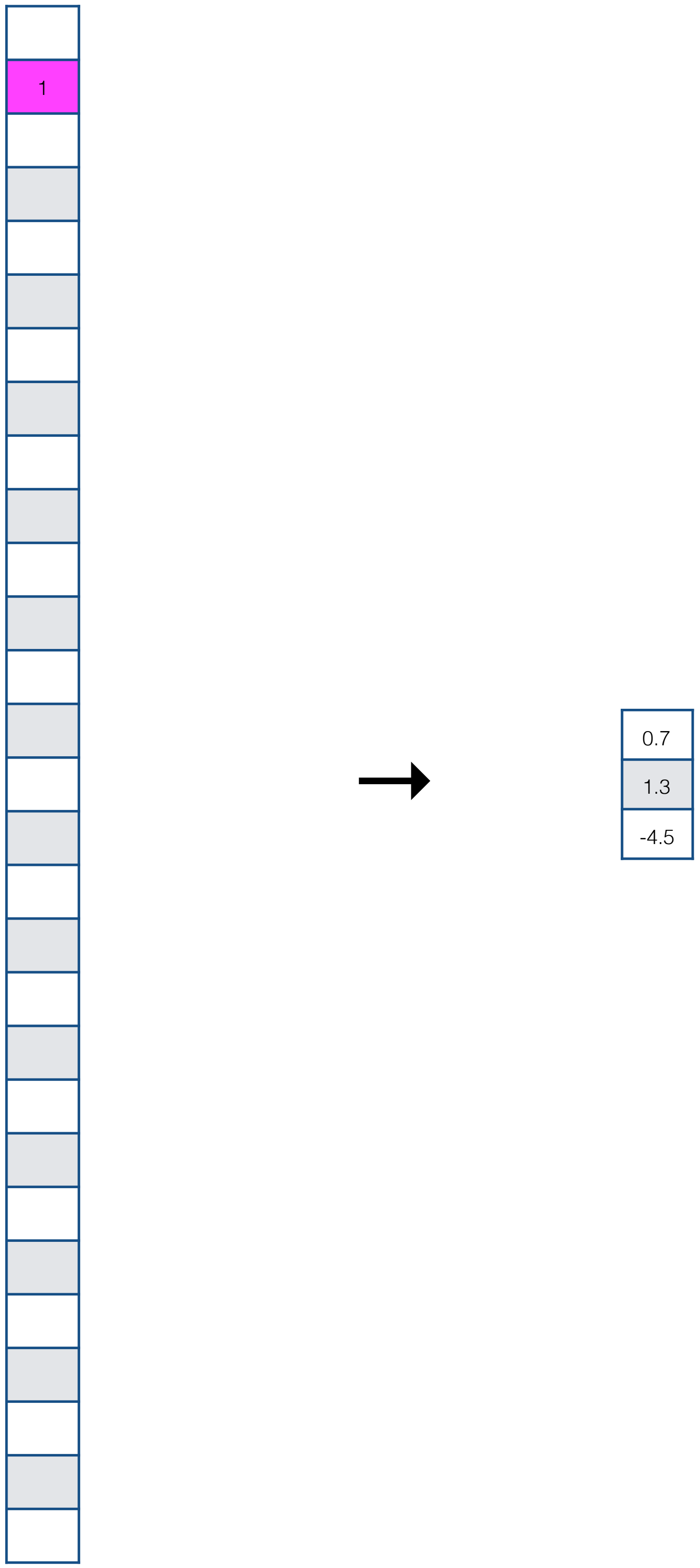
Sparse vectors

“aardvark”

V-dimensional vector, single 1 for the identity of the element

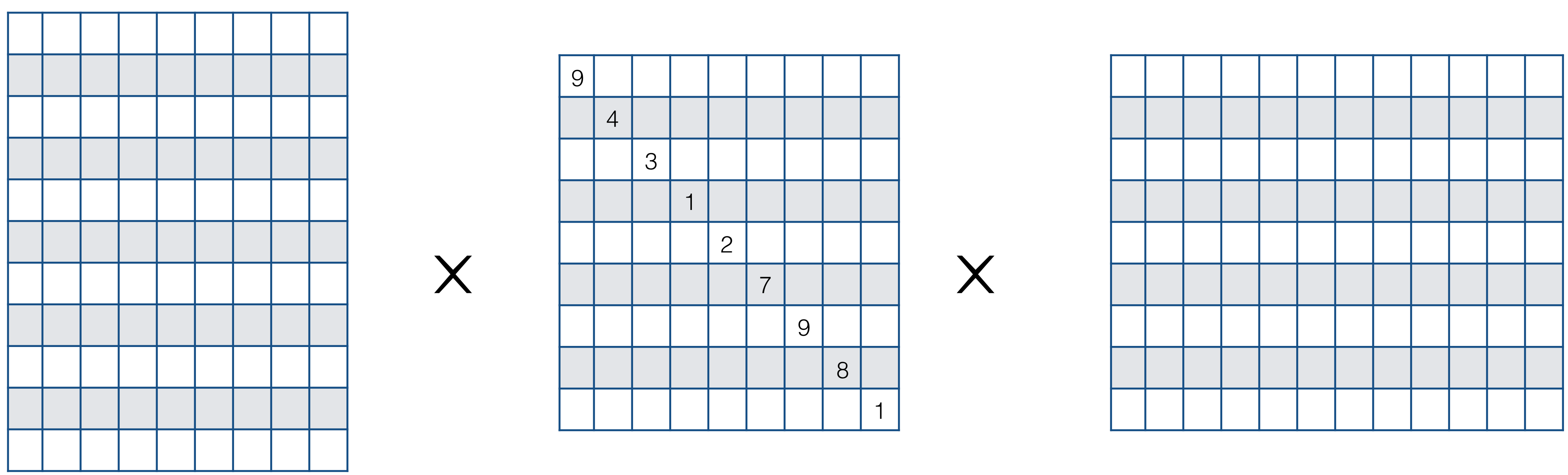
A	0
a	0
aa	0
aal	0
aalii	0
aam	0
Aani	0
aardvark	1
aardwolf	0
...	0
zymotoxic	0
zymurgy	0
Zyrenian	0
Zyrian	0
Zyryan	0
zythem	0
Zythia	0
zythum	0
Zyzomys	0
Zyzzogeton	0

Dense vectors



Singular value decomposition

- Any $n{\times}p$ matrix X can be decomposed into the product of three matrices (where m = the number of linearly independent rows)



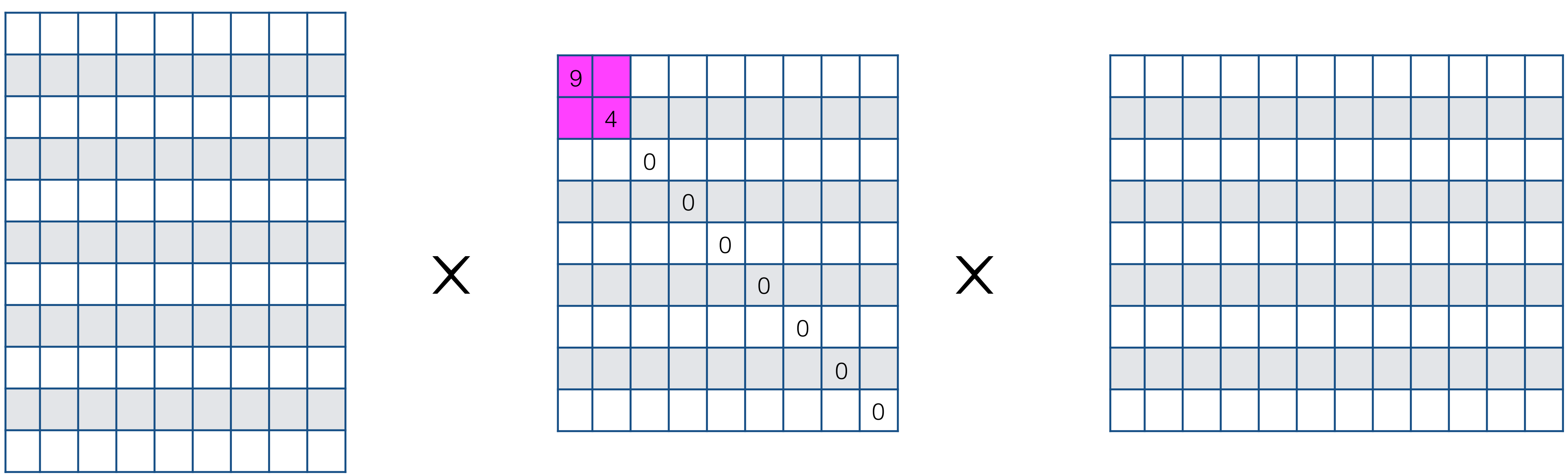
n x m

m x m
(diagonal)

m x p

Singular value decomposition

- We can approximate the full matrix by only considering the leftmost k terms in the diagonal matrix



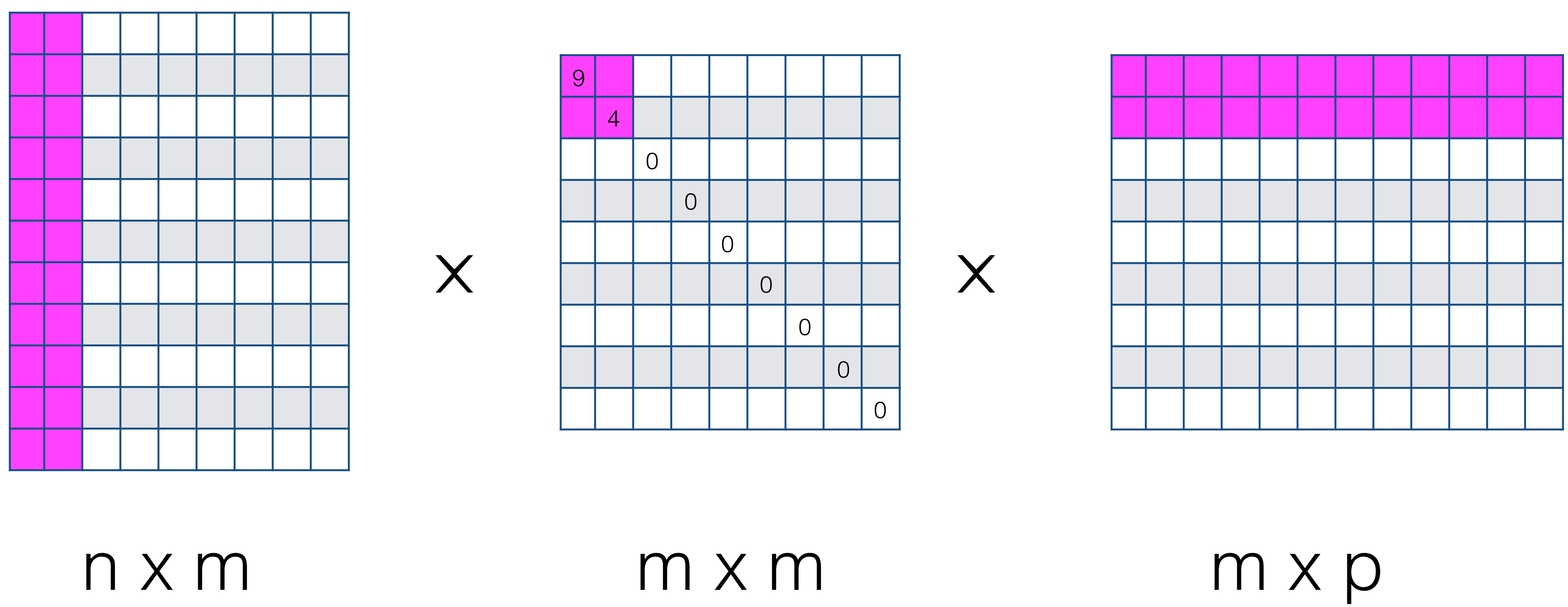
n x m

m x m
(diagonal)

m x p

Singular value decomposition

- We can approximate the full matrix by only considering the leftmost k terms in the diagonal matrix (the k largest singular values)



	Hamlet	Macbeth	Romeo & Juliet	Richard III	Julius Caesar	Tempest	Othello	King Lear
knife	1	1	4	2		2		2
dog	2		6	6		2		12
sword	17	2	7	12		2		17
love	64		135	63		12		48
like	75	38	34	36	34	41	27	44

knife		
dog		
sword		
love		
like		

Hamlet	Macbeth	Romeo & Juliet	Richard III	Julius Caesar	Tempest	Othello	King Lear

Low-dimensional representation for terms (here 2-dim)

Low-dimensional representation for documents (here 2-dim)

knife		
dog		
sword		
love		
like		

Hamle t	Macbet h	Romeo & Juliet	Richar d III	Julius Caesar	Tempe st	Othello	King Lear

Latent semantic analysis

- Latent Semantic Analysis/Indexing (Deerwester et al. 1998) is this process of applying SVD to the term-document co-occurence matrix
- Terms typically weighted by tf-idf
- This is a form of dimensionality reduction (for terms, from a D-dimensionsal sparse vector to a K-dimensional dense one), $K << D$.

dist sim + dist rep

- Term-term co-occurrence matrix

- SVD to yield low-dimensional representation

	bank	*interest*	*finals*
cash	300	210	133
sport	75	140	200

Figure 1: A collocation matrix.

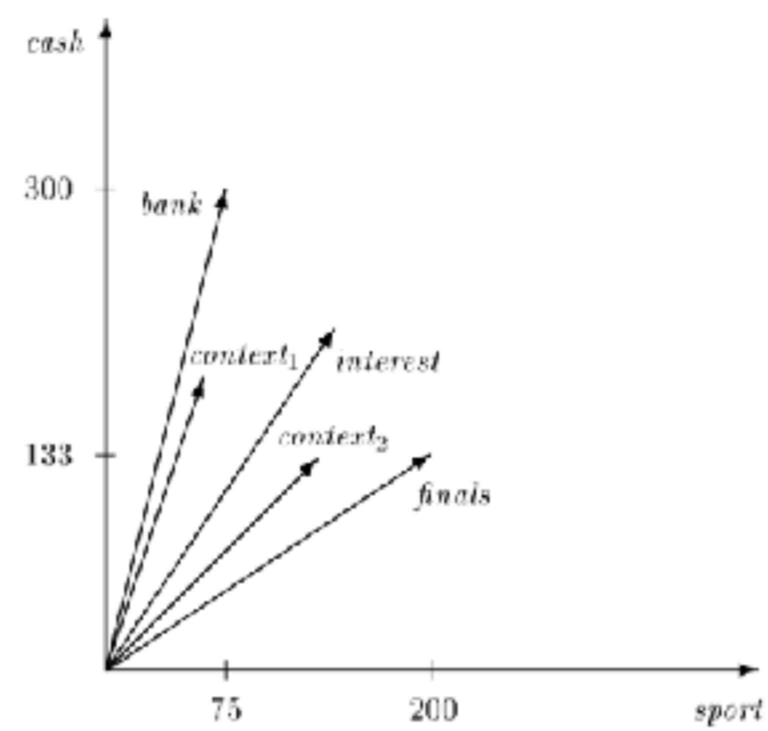


Figure 2: A vector model for context.

Schütze (1992), "Dimensions of Meaning"

7.embeddings/ DistributionalSimilarity.ipynb

- Explore context choices for distributional vectors
- Find more similar words
- Understands which contexts were important for determining similarity.