Language Modeling: Documents

EECS 395/495 Fall 2013

Language Modeling

- Modeling Documents
  - "Bag of words"
  - Latent Semantic Analysis, Latent Dirichlet Allocation
- Modeling sequences of words
  - N-gram Models
  - HMMs
  - Neural Network Language Models

Latent Semantic Analysis (LSA)

- The technique that started it all
  - ca. 1989
- Idea: automatically find similar words, docs
  - If two words tend to occur in similar documents, the words are similar
  - If two documents tend to include similar words, the documents are similar

LSA: Linear Algebra Formulation (1)

- *X* = term-frequency matrix (*t* x *d*)

	d1	d2	d3	d4	d5	d6
airplane	2	21	0	0	0	10
does	3	20	13	10	2	12
elephant	2	3	0	2	0	0
found	12	4	2	3	12	1
house	1	1	0	0	1	0

LSA: Linear Algebra Formulation (2)

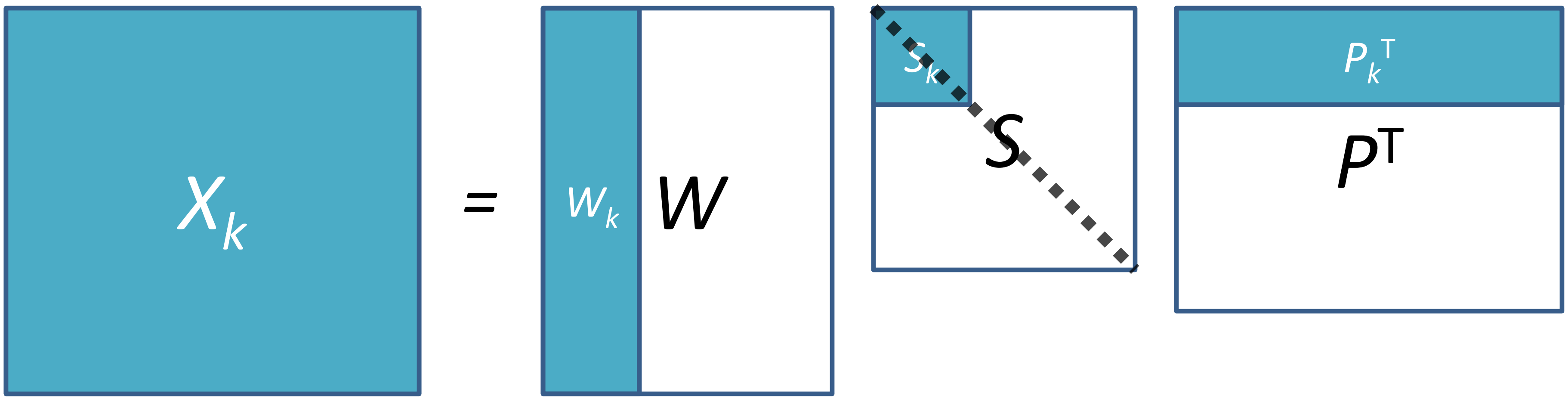
- Write X (t x d) as

$$X = W\,S\,P^{\mathrm{T}}$$

- Where
  - W (t x r) and P (r x t) are orthonormal matrices
  - S (r x r) is a diagonal matrix, entries sorted in decreasing order
  - ….where r = min(t, r)

LSA: Linear Algebra Formulation (3)

- $X = W\,S\,P^{T}$



- $X_k = W_k\,S_k\,P_k^{T}$ (all but first *k* entries of *S* => 0)
- **Key:** X_k is the *best* rank-k approx. to *X*
  (in mean squared error)

Rank-*k* approximation

- $X_k = W_k \, S_k \, P_k^{\mathsf{T}}$
- Put another way:
  - Represent each *term* as k-vector of numbers
  - Represent each *document* as another k-vector
- Vectors represent "semantics" in the sense that
  - Entry in X_k is dot product of term & doc vectors (with dims weighted by S_k)
  - No other length *k* vectorization approximates *X* better

Example

Example of text data: Titles of Some Technical Memos

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* of paths in *trees*
m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4: *Graph minors*: A *survey*

From:

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, **25**, 259-284.

$$\{X\} =$$

	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 (human.user) = -.38

r (human.minors) = -.29

$\{W\} =$

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

$\{S\} =$

3.34								
	2.54							
		2.35						
			1.64					
				1.50				
					1.31			
						0.85		
							0.56	
								0.36

$\{P\} =$

0.20	0.61	0.46	0.54	0.28	0.00	0.01	0.02	0.08
-0.06	0.17	-0.13	-0.23	0.11	0.19	0.44	0.62	0.53
0.11	-0.50	0.21	0.57	-0.51	0.10	0.19	0.25	0.08
-0.95	-0.03	0.04	0.27	0.15	0.02	0.02	0.01	-0.03
0.05	-0.21	0.38	-0.21	0.33	0.39	0.35	0.15	-0.60
-0.08	-0.26	0.72	-0.37	0.03	-0.30	-0.21	0.00	0.36
0.18	-0.43	-0.24	0.26	0.67	-0.34	-0.15	0.25	0.04
-0.01	0.05	0.01	-0.02	-0.06	0.45	-0.76	0.45	-0.07
-0.06	0.24	0.02	-0.08	-0.26	-0.62	0.02	0.52	-0.45

X_k

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-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	0.70	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	-0.20	-0.11
survey	0.10	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	-0.30	0.20	0.31	0.69	0.98	0.85
minors	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

r (human.user) = .94

r (human.minors) = -.83

To Review

- Represent each *term* as k-vector of numbers
  - "human" = [0.22, -0.11]
- Represent each *document* as another k-vector
  - d1 = [0.2, 0.61]
- Entry in X_k for "human" appearing in p1 is dot product of vectors, weighted by entries of S_k $= S_{k\,(1,1)} * 0.22 * 0.2 + S_{k\,(2,2)} * -0.11 * 0.61 = -0.02$
- So using k = 1, what does the word vector signify? The document vector?

Examples

- Alternative of *stemming* often conflates meanings, e.g. *flower* becomes *flow* – as compared to LSA:
  - *flower-flow have cos = -.01,*
  - *dish-dishes cos = .68.*
- More examples of *cos* in latent space capturing word meaning:
  - Flower: petals .93, gymnosperms 0.47
  - Flow: flows .84, opens 0.46
  - Dish: sauce 0.70, bowl 0.63
  - Dishes: kitchen 0.75, cup 0.57

Thomas K Landauer and Susan Dumais (2008), Scholarpedia, 3(11):4356
http://www.scholarpedia.org/article/Latent_semantic_analysis