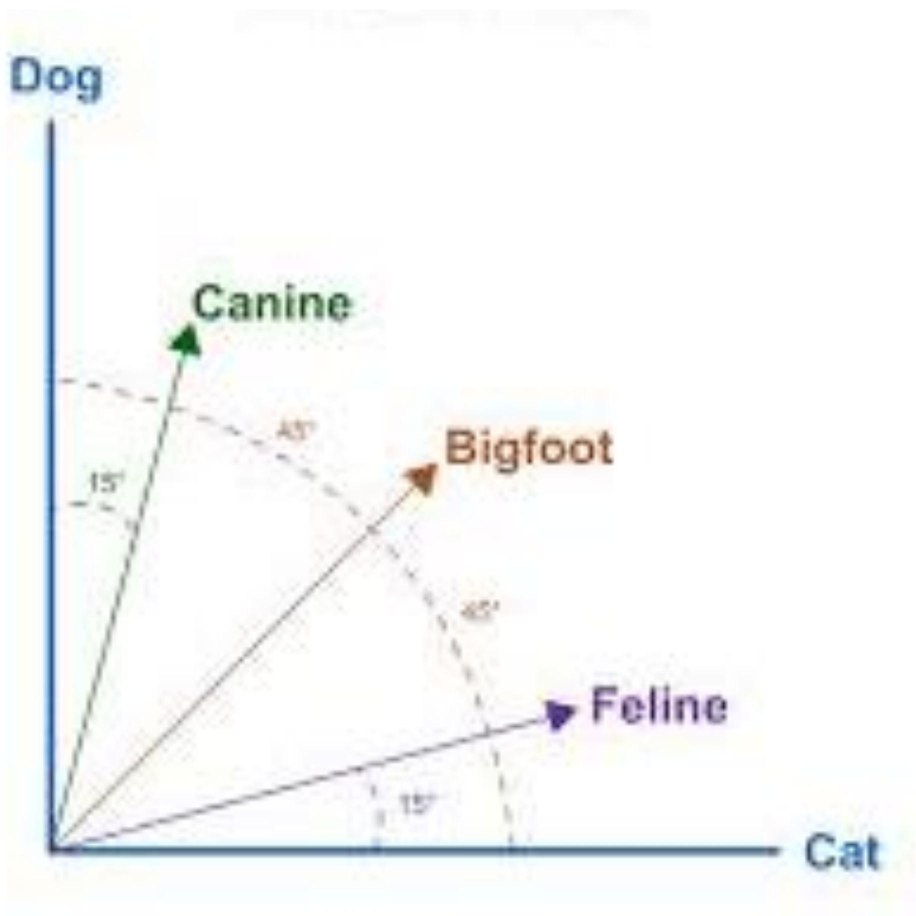


COMP90042 LECTURE 9

DISTRIBUTIONAL SEMANTICS

CO-OCCURRENCE AND SEMANTICS

- "You shall know a word by the company it keeps" (Firth)
- Local context reflects a word's semantic class
  - E.g. *eat a pizza*, *eat a burger*
- Document co-occurrence often indicative of topic
  - E.g. *voting* and *politics*

OUTLINE

- Word-word association using PMI
- The vector space model
  - Weighting for information retrieval: *tf-idf*
  - Dimensionality reduction: SVD
  - Similarity in vector space: cosine

CO-OCCURRENCE MATRIX

	…	the	country	hell	…
…					
state		1973	10	1	
fun		54	2	0	
heaven		55	1	3	
……					

- Lists how often words appear with other words
  - In some predefined context (e.g. sentence in Brown corpus)
- The obvious problem with raw frequency: dominated by common words

POINTWISE MUTUAL INFORMATION

For two events x and y, pointwise mutual information (PMI) comparison between the actual joint probability of the two events (as seen in the data) with the expected probability under the assumption of independence

$$PMI(x, y) = \log_2 \frac{p(x, y)}{p(x)p(y)}$$

CALCULATING PMI

	…	the	country	hell	…		Σ
…							
state		1973	10	1			12786
fun		54	2	0			633
heaven		55	1	3			627
…							
Σ		1047519	3617	780			15871304

$p(x,y) = \text{count}(x,y)/\Sigma$

$p(x) = \Sigma_x / \Sigma$

$p(y) = \Sigma_y / \Sigma$

x= state, y = country

$p(x,y) = 10/15871304 = 6.3 \text{ x } 10^{-7}$

$p(x) = 12786/15871304 = 8.0 \text{ x } 10^{-4}$

$p(y) = 3617/15871304 = 2.3 \text{ x } 10^{-4}$

$PMI(x,y) = \log_2(6.3 \text{ x } 10^{-7})/((8.0 \text{ x } 10^{-4})\ (2.3 \text{ x } 10^{-4}))$

$= 1.78$

PMI MATRIX

	…	the	country	hell	…
…					
state		1.22	1.78	0.63	
fun		0.37	3.79	-inf	
heaven		0.41	2.80	6.60	
……					

- PMI does a better job of capturing interesting semantics
  - E.g. *heaven* and *hell*
- But it is obviously biased towards rare words
- And doesn't handle zeros well

PMI TRICKS

- Zero all negative values (PPMI)
  - Avoid –inf and unreliable negative values
- Counter bias towards rare events
  - Artificially increase marginal probabilities
  - Smooth probabilities

OTHER ASSOCIATION MEASURES

- t-test
- Chi-squared
- Likelihood ratio
- ...

USES OF ASSOCIATION MEASURES

- Collocation extraction
- Word similarity
- Lexicon creation
- Weighting for vector space models

THE VECTOR SPACE MODEL

- Fundamental idea: represent meaning as a vector
- One matrix, two viewpoints
  - Documents represented by their words (web search)
  - Words represented by their documents (text analysis)

	…	425	426	427	…
…					
state		0	3		
fun		1	0		
heaven		0	0		
……					

	…	state	fun	heaven	…
…					
425		0	1	0	
426		3	0	0	
427		0	0	0	
……					

MANIPULATING THE VSM

- Weighting the values
- Creating low-dimensional dense vectors
- Comparing vectors

TF-IDF

- Standard weighting scheme for information retrieval
- Also discounts common words

$$idf_w = \log \frac{|D|}{df_w}$$

tf matrix

	…	the	country	hell	…
…					
425		43	5	1	
426		24	1	0	
427		37	0	3	
…					
df		500	14	7	

tf-idf matrix

	…	the	country	hell	…
…					
425		0	25.8	6.2	
426		0	5.2	0	
427		0	0	18.5	
…					

DIMENSIONALITY REDUCTION

- Term-document matrices are very *sparse*
- Dimensionality reduction: create shorter, denser vectors
- More practical (less features)
- More generalizable (less overfitting)
  - Captures synonymy
- Remove noise

SINGULAR VALUE DECOMPOSITION

$$\begin{array}{cc} & |\mathrm{D}| \\ |\mathrm{V}| & \begin{bmatrix} 0 & 1 & 0 & & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & & 0 \\ & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \end{array}$$

A
(term-document matrix)

$$A = U\Sigma V^T$$

$$m = Rank(A)$$

U
(new term matrix)

Σ
(singular values)

V^T
(new document matrix)

$$\begin{array}{cc} & m \\ |\mathrm{V}| & \begin{bmatrix} 2.2 & 0.3 & & 8.7 \\ 5.5 & -2.8 & \cdots & 0.1 \\ -1.3 & 3.7 & & 3.5 \\ & \vdots & \ddots & \vdots \\ 2.9 & -2.1 & \cdots & -1.9 \end{bmatrix} \end{array} \quad \begin{array}{cc} & m \\ m & \begin{bmatrix} 9.1 & 0 & 0 & & 0 \\ 0 & 4.4 & 0 & \cdots & 0 \\ 0 & 0 & 2.3 & & 0 \\ & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0.1 \end{bmatrix} \end{array} \quad \begin{array}{cc} & |\mathrm{D}| \\ m & \begin{bmatrix} -0.2 & 4.0 & & -1.3 \\ -4.1 & 0.6 & \cdots & -0.2 \\ 2.6 & 6.1 & & 1.4 \\ \vdots & & \ddots & \vdots \\ -1.9 & -1.8 & \cdots & 0.3 \end{bmatrix} \end{array}$$

TRUNCATING

- Truncating U, Σ, and V^T to k dimensions produces best possible k rank approximation of original matrix
- So truncated, U_k (or $V_k{}^T$) is a new low dimensional representation of the word (or document)
- Typical values for k are 100-5000
- When applied to words in documents, often called LSA

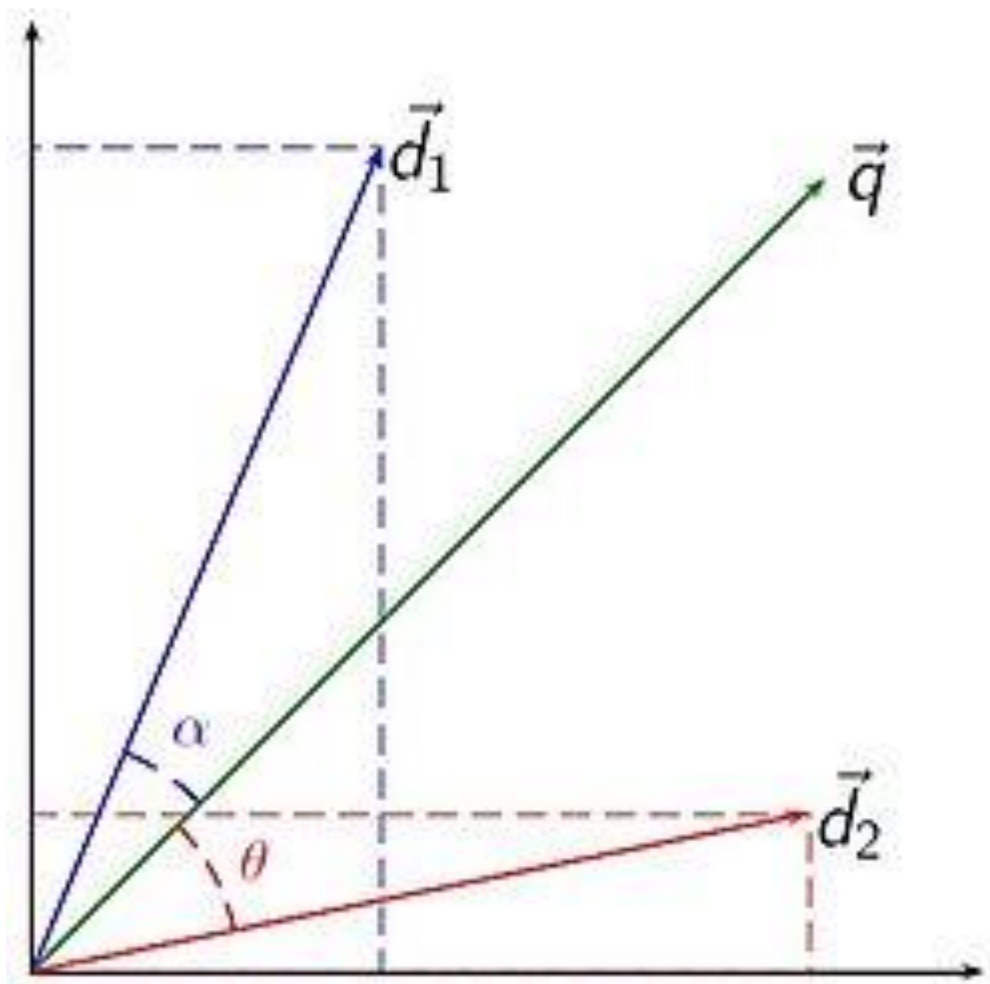
$$U \qquad \begin{array}{cc} & \begin{array}{cccc} & & m & k \end{array} \\ |\mathrm{V}| & \begin{bmatrix} 2.2 & 0.3 & & -2.4 & & 8.7 \\ 5.5 & -2.8 & \cdots & 1.1 & \cdots & 0.1 \\ -1.3 & 3.7 & & 4.7 & & 3.5 \\ & \vdots & \ddots & \vdots & \ddots & \vdots \\ 2.9 & -2.1 & \cdots & -3.3 & \cdots & -1.9 \end{bmatrix} \end{array} \Rightarrow \quad U_k \qquad \begin{array}{cc} & k \\ |\mathrm{V}| & \begin{bmatrix} 2.2 & 0.3 & & -2.4 \\ 5.5 & -2.8 & \cdots & 1.1 \\ -1.3 & 3.7 & & 4.7 \\ & \vdots & \ddots & \vdots \\ 2.9 & -2.1 & \cdots & -3.3 \end{bmatrix} \end{array}$$

MORE DIMENSIONALITY REDUCTION

- Principal components analysis (PCA)
- Independent components analysis
- Factor analysis
- Nonnegative matrix factorization
- Latent Dirichlet allocation
- …

SIMILARITY

- Regardless of vector representation, classic use of vector is comparison with other vector
  - Though vectors can also be used directly as features
- For IR: find documents most similar to query



COSINE SIMILARITY

The cosine of the angle between two vectors is the dot product of the two vectors divided by the product of their norms:

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|}$$

Where

$$\vec{a} \cdot \vec{b} = \sum_{i=1}^{N} a_i b_i$$

And

$$|\vec{a}| = \sqrt{\sum_{i=1}^{N} {a_i}^2}$$

COSINE SIMILARITY EXAMPLE

$$\vec{a} = [0,0,1,0,1,1,1]$$

$$\vec{b} = [0,1,1,0,1,1,0]$$

$$\vec{a} \cdot \vec{b} = 1 + 1 + 1 = 3$$

$$|\vec{a}| = \left|\vec{b}\right| = \sqrt{1 + 1 + 1 + 1} = 2$$

$$\cos\theta = \frac{3}{2*2} = 0.75$$

OTHER METRICS FOR VECTOR SIMILARITY

- Euclidean
- Jaccard
- Dice
- Kullback-Leibler (KL) Divergence
- Jenson-Shannon Divergence
- …

A FINAL WORD

- Distributional semantics is fundamental to both fields covered by this course
- Basic methodology presented here is not new, but still very much used
- In the next lecture we will continue with distributional semantics, looking at a recent, state-of-the-art approach to dimensionality reduction using neural networks

FURTHER READING

- J&M3, Ch 15, J&M3 Ch 16.1
- (Optional) From Frequency to Meaning: Vector Space Models of Semantics