

# Vector Semantics

Dense Vectors

# Sparse versus dense vectors

- PPMI vectors are
  - **long** (length  $|V| = 20,000$  to  $50,000$ )
  - **sparse** (most elements are zero)
- Alternative: learn vectors which are
  - **short** (length 200-1000)
  - **dense** (most elements are non-zero)

# Sparse versus dense vectors

- Why dense vectors?
  - Short vectors may be easier to use as features in machine learning (less weights to tune)
  - Dense vectors may generalize better than storing explicit counts
  - They may do better at capturing synonymy:
    - *car* and *automobile* are synonyms; but are represented as distinct dimensions; this fails to capture similarity between a word with *car* as a neighbor and a word with *automobile* as a neighbor

# Three methods for getting short dense vectors

- Singular Value Decomposition (SVD)
  - A special case of this is called LSA – Latent Semantic Analysis
- “Neural Language Model”-inspired predictive models
  - skip-grams and CBOW
- Brown clustering

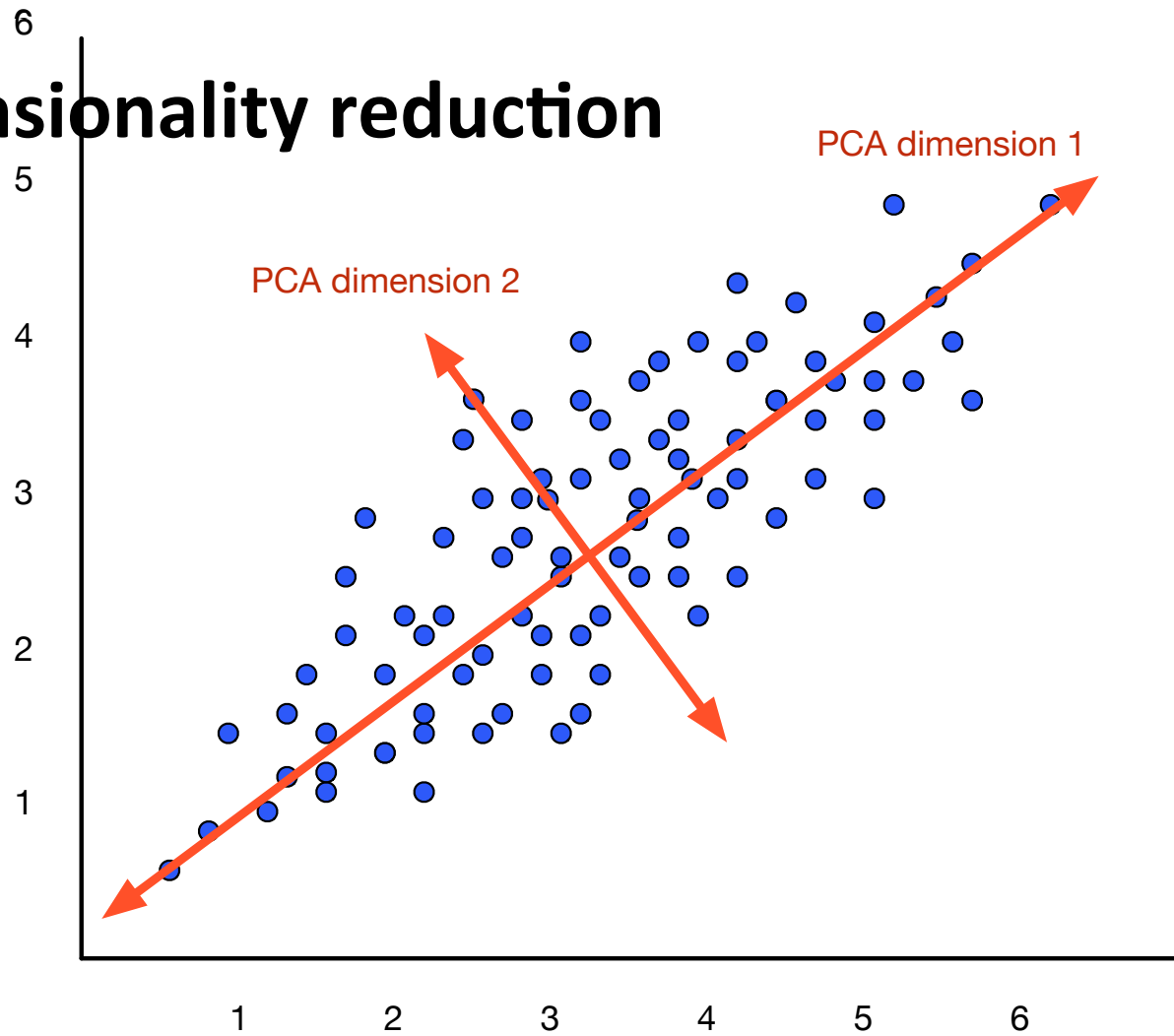
# Vector Semantics

Dense Vectors via SVD

# Intuition

- Approximate an N-dimensional dataset using fewer dimensions
- By first rotating the axes into a new space
- In which the highest order dimension captures the most variance in the original dataset
- And the next dimension captures the next most variance, etc.
- Many such (related) methods:
  - PCA – principle components analysis
  - Factor Analysis
  - SVD

# Dimensionality reduction



# Singular Value Decomposition

*Any rectangular matrix  $X$  equals the product of 3 matrices:*

**W**: rows corresponding to original but  $m$  columns represents a dimension in a new latent space, such that

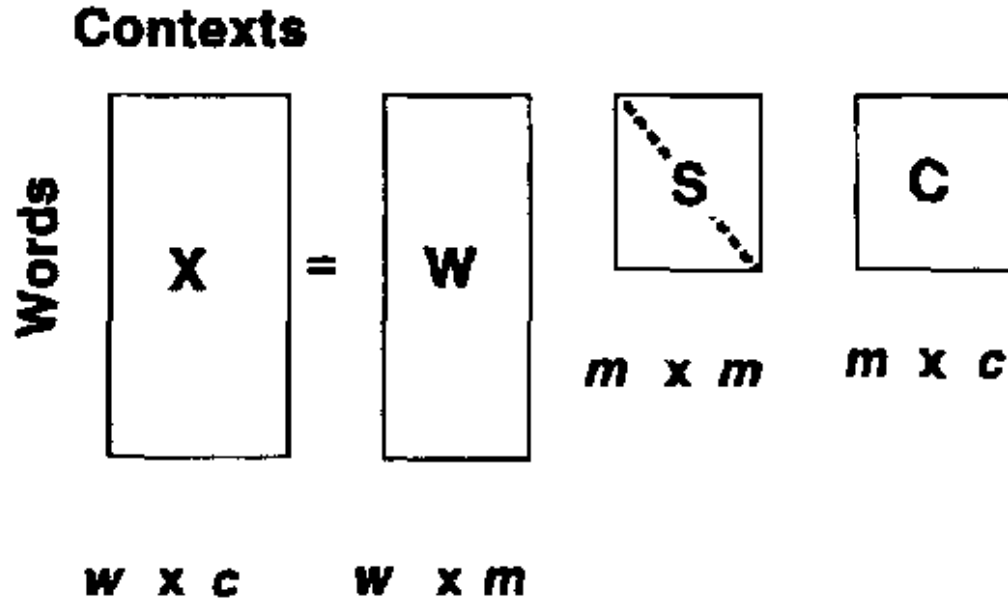
- $M$  column vectors are orthogonal to each other
- Columns are ordered by the amount of variance in the dataset each new dimension accounts for

**S**: diagonal  $m \times m$  matrix of **singular values** expressing the importance of each dimension.

**C**: columns corresponding to original but  $m$  rows corresponding to singular values



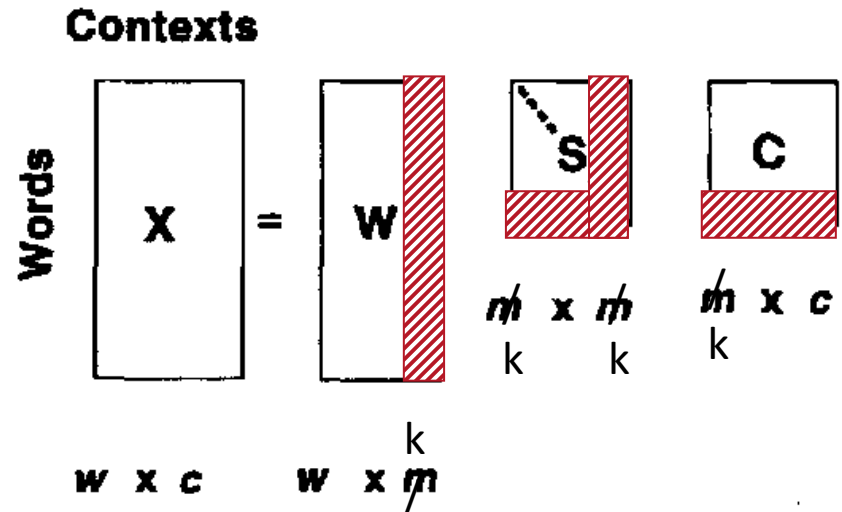
# Singular Value Decomposition



# SVD applied to term-document matrix: Latent Semantic Analysis

Deerwester et al (1988)

- If instead of keeping all  $m$  dimensions, we just keep the top  $k$  singular values. Let's say 300.
- The result is a least-squares approximation to the original  $X$
- But instead of multiplying, we'll just make use of  $W$ .
- Each row of  $W$ :
  - A  $k$ -dimensional vector
  - Representing word  $W$



# Let's return to PPMI word-word matrices

- Can we apply SVD to them?

# SVD applied to term-term matrix

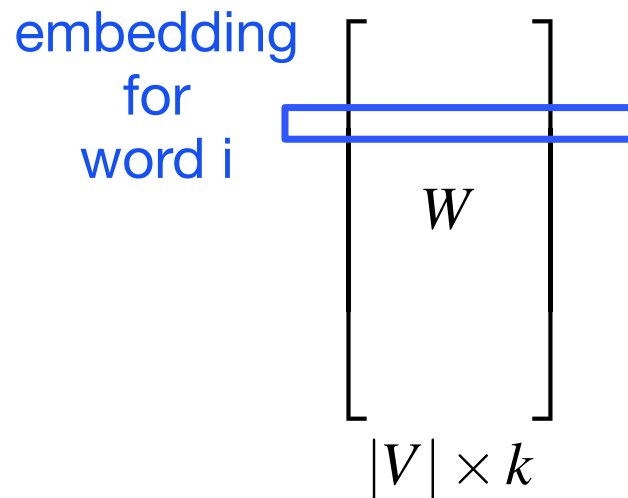
$$\begin{bmatrix} X \\ |V| \times |V| \end{bmatrix} = \begin{bmatrix} W \\ |V| \times |V| \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2 & 0 & \dots & 0 \\ 0 & 0 & \sigma_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma_V \end{bmatrix} \begin{bmatrix} C \\ |V| \times |V| \end{bmatrix}$$

# Truncated SVD on term-term matrix

$$\begin{bmatrix} X \\ |V| \times |V| \end{bmatrix} = \begin{bmatrix} W \\ |V| \times k \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2 & 0 & \dots & 0 \\ 0 & 0 & \sigma_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma_k \end{bmatrix} \begin{bmatrix} C \\ k \times |V| \end{bmatrix}$$

# Truncated SVD produces embeddings

- Each row of  $W$  matrix is a  $k$ -dimensional representation of each word  $w$
- $K$  might range from 50 to 1000
- Generally we keep the top  $k$  dimensions, but some experiments suggest that getting rid of the top 1 dimension or even the top 50 dimensions is helpful (Lapesa and Evert 2014).



# Embeddings versus sparse vectors

- Dense SVD embeddings sometimes work better than sparse PPMI matrices at tasks like word similarity
  - Denoising: low-order dimensions may represent unimportant information
  - Truncation may help the models generalize better to unseen data.
  - Having a smaller number of dimensions may make it easier for classifiers to properly weigh the dimensions for the task.
  - Dense models may do better at capturing higher order co-occurrence.

# Vector Semantics

Embeddings inspired by  
neural language models:  
skip-grams and CBOW



# Prediction-based models:

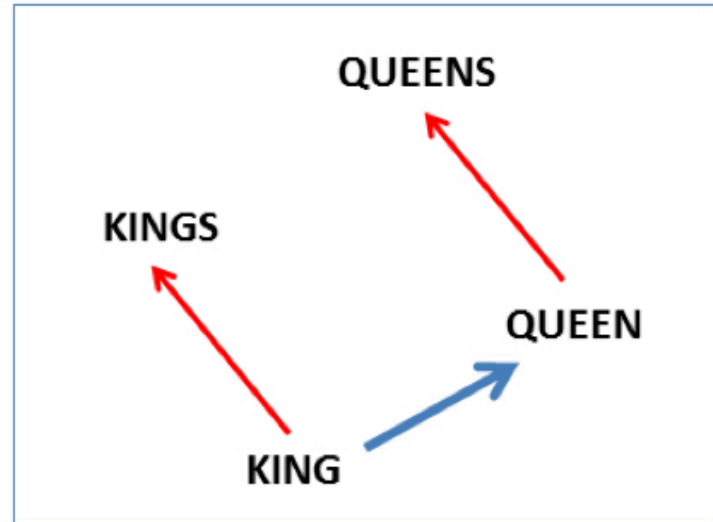
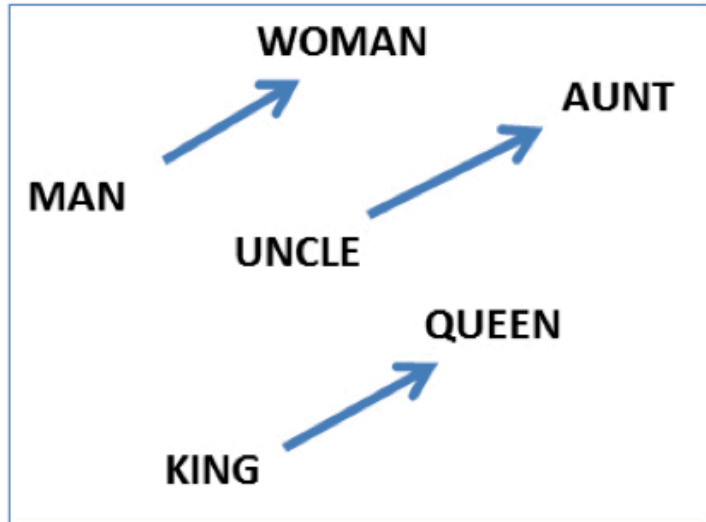
## An alternative way to get dense vectors

- **Skip-gram** (Mikolov et al. 2013a) **CBOW** (Mikolov et al. 2013b)
- Learn embeddings as part of the process of word prediction.
- Train a neural network to predict neighboring words
  - Inspired by **neural net language models**.
  - In so doing, learn dense embeddings for the words in the training corpus.
- Advantages:
  - Fast, easy to train (much faster than SVD)
  - Available online in the `word2vec` package
  - Including sets of pretrained embeddings!

# Embeddings capture relational meaning!

$\text{vector}('king') - \text{vector}('man') + \text{vector}('woman') \approx \text{vector}('queen')$

$\text{vector}('Paris') - \text{vector}('France') + \text{vector}('Italy') \approx \text{vector}('Rome')$



# Vector Semantics

Brown clustering

# Brown clustering

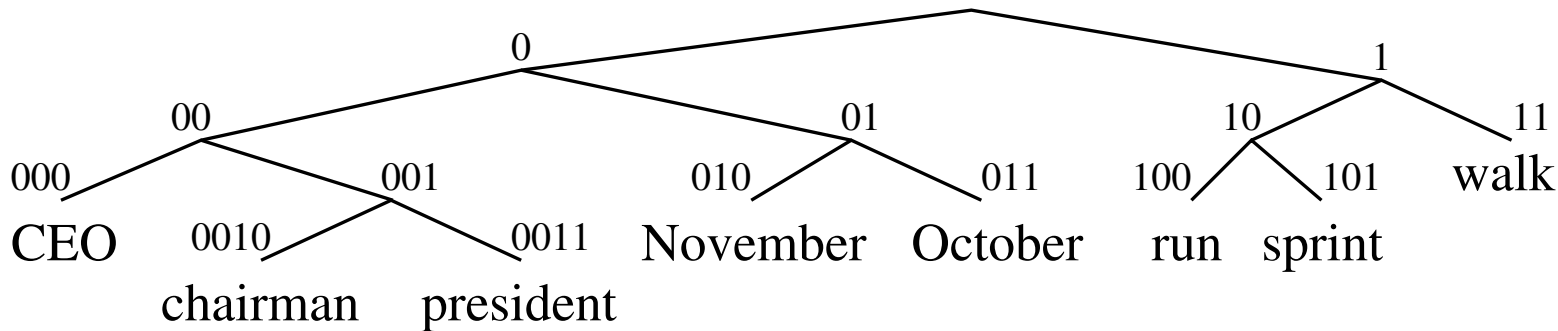
- An agglomerative clustering algorithm that clusters words based on which words precede or follow them
- These word clusters can be turned into a kind of vector
- We'll give a very brief sketch here.

# Brown clustering algorithm

- Each word is initially assigned to its own cluster.
- We now consider merging each pair of clusters. Highest quality merge is chosen.
  - Quality = merges two words that have similar probabilities of preceding and following words
  - (More technically quality = smallest decrease in the likelihood of the corpus according to a class-based language model)
- Clustering proceeds until all words are in one big cluster.

# Brown Clusters as vectors

- By tracing the order in which clusters are merged, the model builds a binary tree from bottom to top.
- Each word represented by binary string = path from root to leaf
- Each intermediate node is a cluster
- Chairman is 0010, “months” = 01, and verbs = 1



# Brown cluster examples

Friday Monday Thursday Wednesday Tuesday Saturday Sunday weekends Sundays Saturdays  
June March July April January December October November September August  
pressure temperature permeability density porosity stress velocity viscosity gravity tension  
anyone someone anybody somebody  
had hadn't hath would've could've should've must've might've  
asking telling wondering instructing informing kidding reminding bothering thanking deposing  
mother wife father son husband brother daughter sister boss uncle  
great big vast sudden mere sheer gigantic lifelong scant colossal  
down backwards ashore sideways southward northward overboard aloft downwards adrift

# Vector Semantics

Evaluating similarity



# Evaluating similarity

- Extrinsic (task-based, end-to-end) Evaluation:
  - Question Answering
  - Spell Checking
  - Essay grading
- Intrinsic Evaluation:
  - Correlation between algorithm and human word similarity ratings
    - Wordsim353: 353 noun pairs rated 0-10.  $sim(plane, car)=5.77$
  - Taking TOEFL multiple-choice vocabulary tests
    - Levied is closest in meaning to:  
imposed, believed, requested, correlated

# Summary

- Distributional (vector) models of meaning
  - **Sparse** (PPMI-weighted word-word co-occurrence matrices)
  - **Dense:**
    - Word-word SVD 50-2000 dimensions
    - Skip-grams and CBOW
    - Brown clusters 5-20 binary dimensions.