Distributional Semantics

Jacob Andreas / MIT 6.864 / Spring 2020

Announcements

Check email for exam info (makeups over break)

Listener status (assignments won’t be graded)

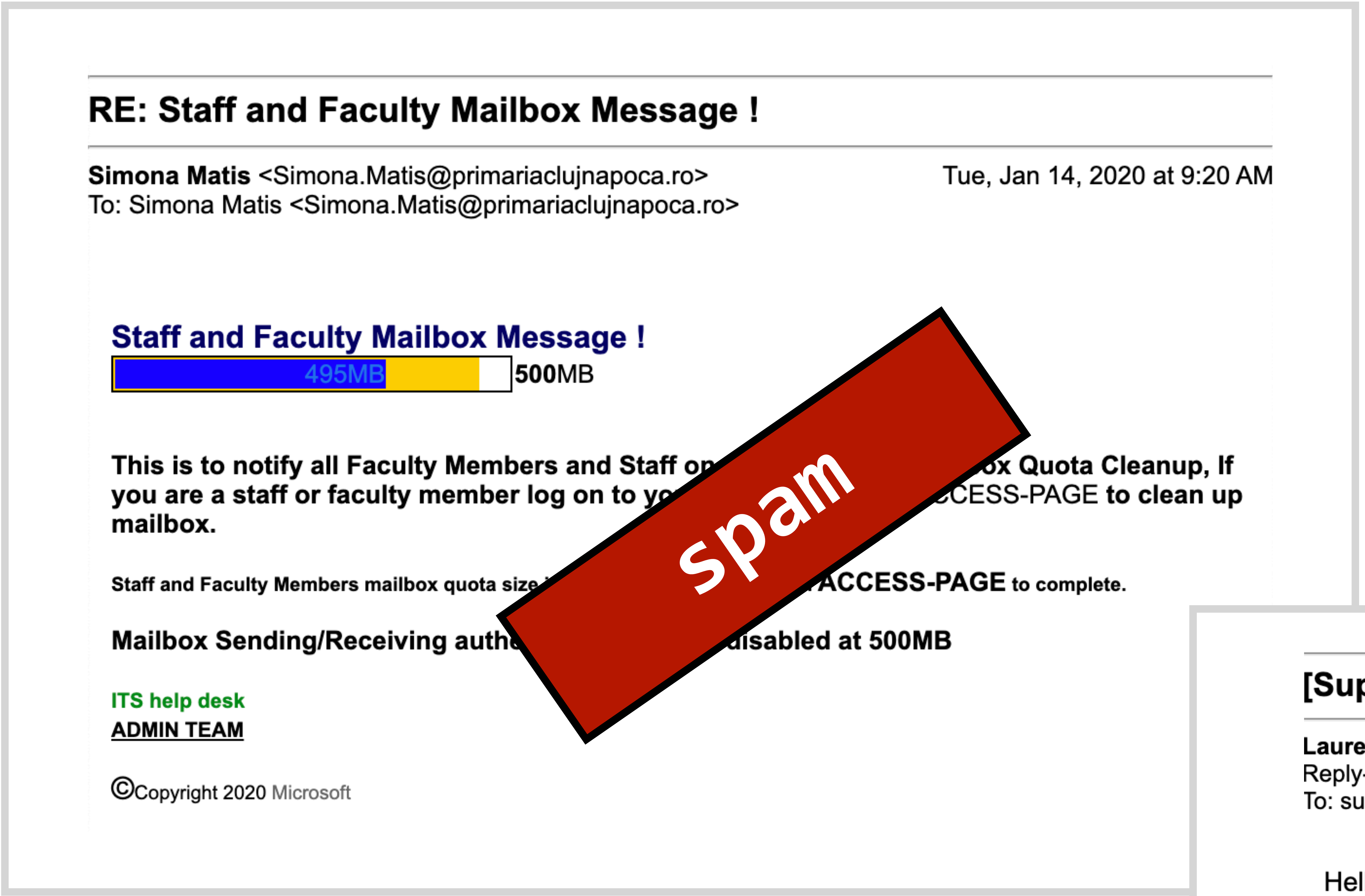
Private piazza posts for project feedback

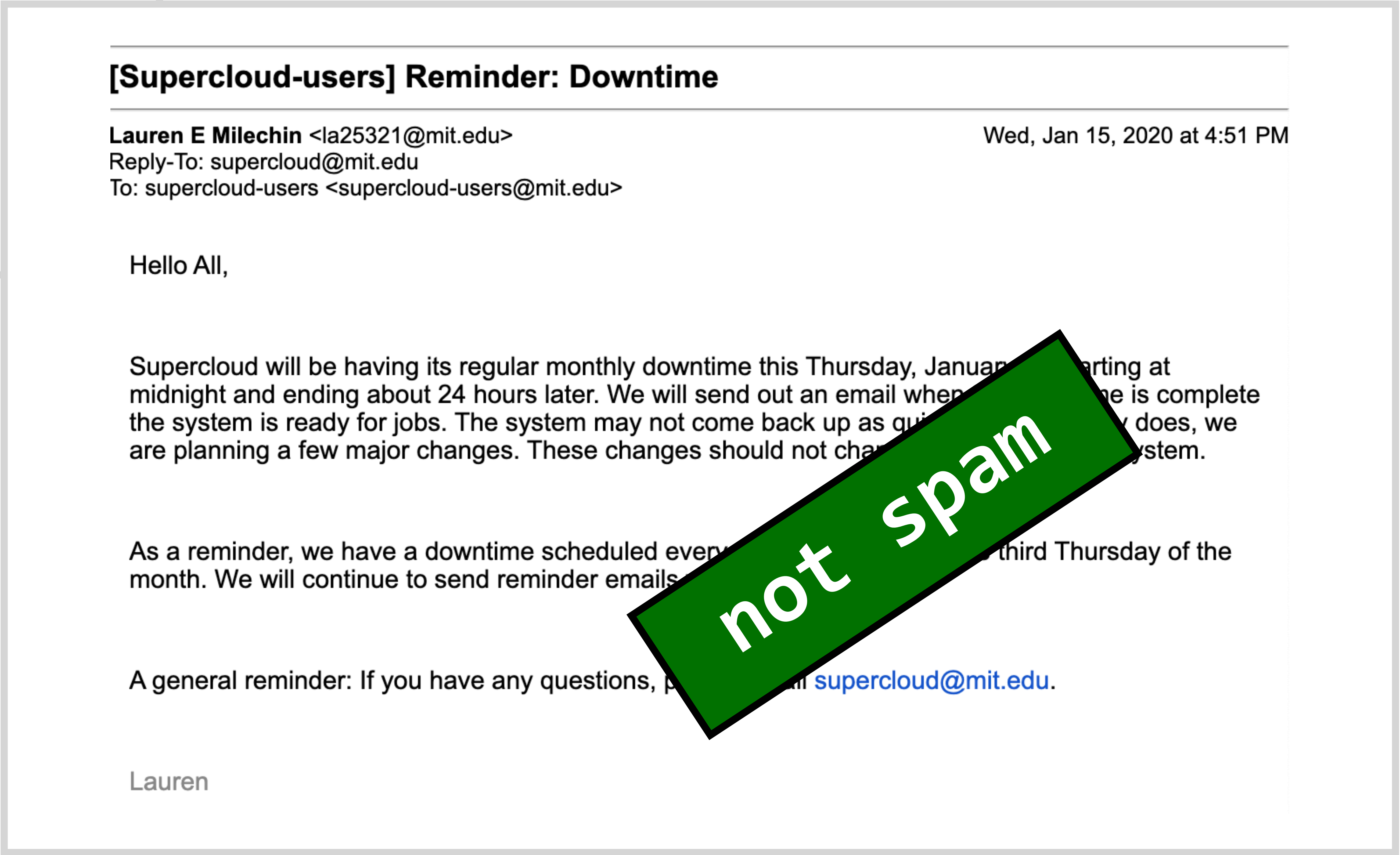
Homework: expected release this weekend

Readings online

Recap: text classification

Spam classification





Linear models

$$s = W^{\mathsf{T}} x$$

$$\begin{array}{l} \text{happy} \\ \text{camper} \\ \text{so} \\ \text{good} \\ \text{horse} \\ \text{saving} \\ \text{delicious} \\ \text{poisoning} \end{array} \quad \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} -2 \\ 0 \\ -1 \\ -1 \\ 1 \\ 0 \\ -2 \\ 4 \end{bmatrix}$$

$$\begin{bmatrix} -3 \\ -2 \\ 0 \\ 2 \\ 3 \end{bmatrix} \quad \begin{array}{l} \text{1 star} \\ \text{2 stars} \\ \text{3 stars} \\ \text{4 stars} \\ \text{5 stars} \end{array}$$

Interpretation: deep bag of words

$$s = W_2^\top f(W_1^\top x)$$

x $\quad f(W_1^\top x) = h_1 \quad W_2^\top h_1 = s$

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \text{input} \qquad \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} \text{“hidden layer”} \qquad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \text{output}$$

Interpretation: deep bag of words

$$s = W_2^\top f(W_1^\top x)$$

$w_{1,i}^\top$

$$\begin{bmatrix}0.1\\1.7\\0.3\end{bmatrix}\begin{bmatrix}0.1\\1.7\\0.3\end{bmatrix}\begin{bmatrix}0.1\\1.7\\0.3\end{bmatrix} \xrightarrow{f(\sum_{i:\ x_i=1} w_i)} \begin{bmatrix}1.2\\1.8\\-0.1\end{bmatrix} \longrightarrow 0.9$$

x *very good, but gave me*

Learning: likelihood

$$L(s, y) = -\log p(y \mid x)$$

$$= -\log \frac{\exp(s_y)}{\sum_i \exp(s_i)}$$

$$= -s_y + \log \sum_i \exp(s_i) := -\log \mathrm{softmax(s)}_y$$

Idea: treat s as a vector of (unnormalized) log-probs, and maximize p(y | x; W).

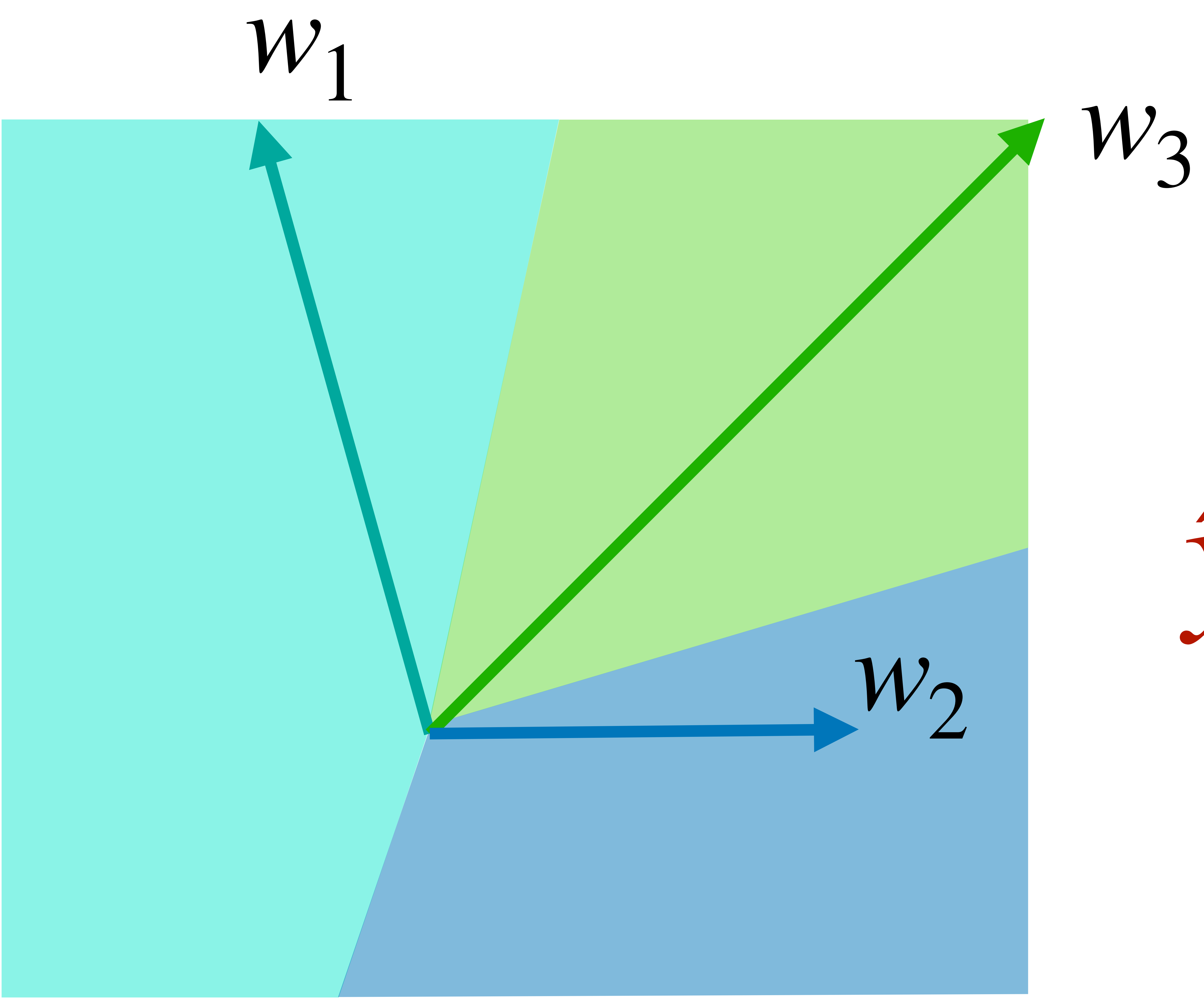
Learning: margin

$$L(s, y) = [s_y - \max(s_{-y}) - c]_+$$

$[x]_+ := \max(x, 0)$

Idea: try to make the score of the right label s_y at least at least *c* greater than the score of every wrong label.

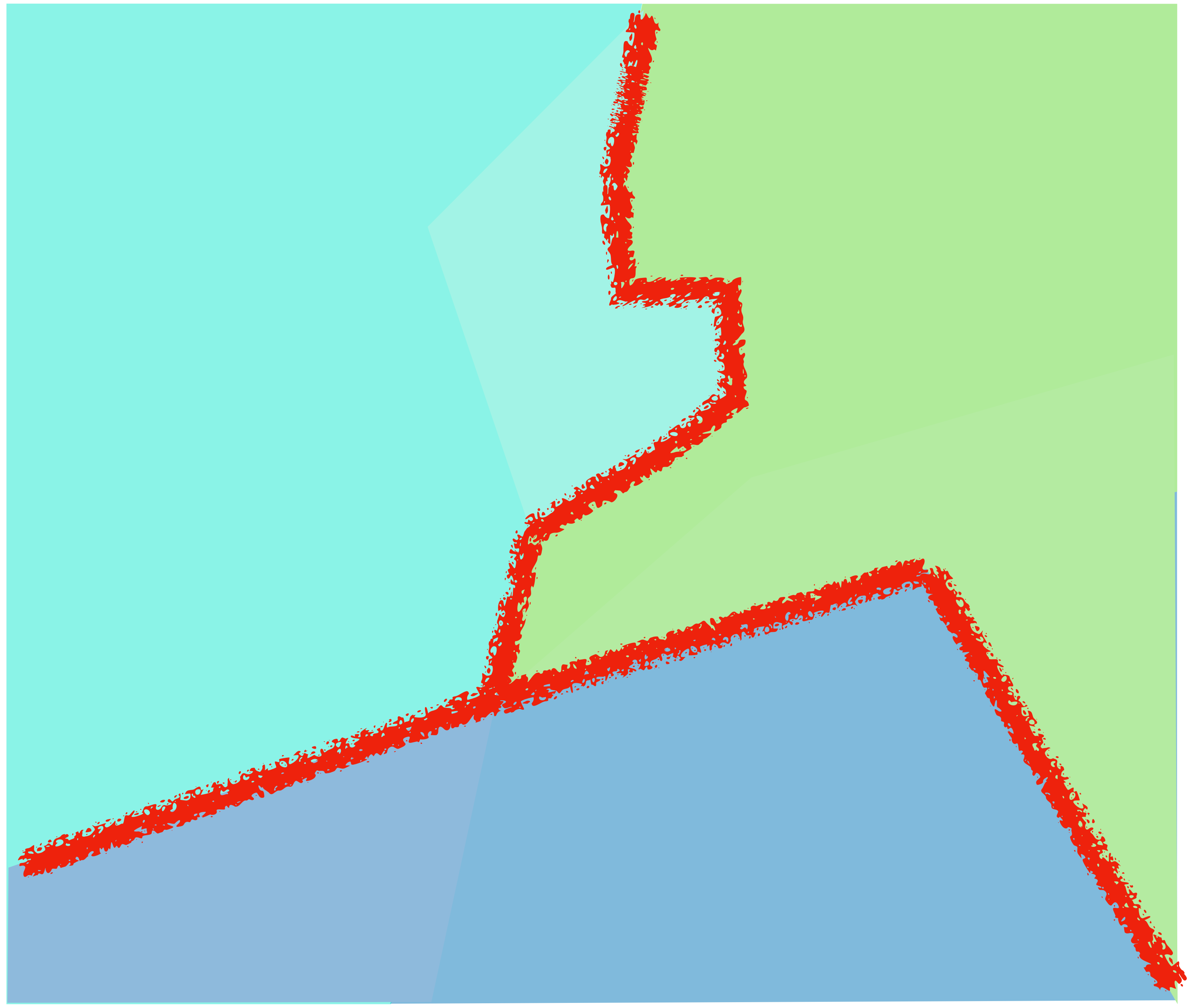
Linear decision boundaries



$$\hat{y} = \text{argmax}_i \; w_i^\top x$$

Multilayer perceptron

$$s = W_2^{\top} f(W_1^{\top} x)$$



Nonlinear decision boundaries!

Challenges: data sparsity^2

$x =$ *perhaps not the most scintillating work in the director's oeuvre*

somewhat...awful	-1	
not...scintillating	0	???
less...terrible	1	

$$y = ?$$

Can we learn that similar words behave similarly in combination?

Interpretation: deep bag of words

More generally, can we learn portable representations of words independent of specific prediction problems?

$$w_{1,i}^{\top}$$

$$\begin{bmatrix}0.1\\1.7\\0.3\end{bmatrix}\begin{bmatrix}0.1\\1.7\\0.3\end{bmatrix}\begin{bmatrix}0.1\\1.7\\0.3\end{bmatrix} \xrightarrow{f(\sum_{i:\ x_i=1} w_i)} \begin{bmatrix}1.2\\1.8\\-0.1\end{bmatrix} \longrightarrow 0.9$$

x *very good, but gave me*

Distributional semantics

Words in context

...it ? its...

...has ? earnings...

...either ? or...

...which ? the...

Words in context

...it ? its...

...has ? earnings...

...either ? or...

...which ? the...

What do we know about the word at "?"

Words in context

...it ? its...

...has ? earnings...

...either ? or...

...which ? the...

What do we know about the word at “?”

not much about meaning, but probably a verb

Wider contexts

...but simple block ? superimposed on the...

...lawyers recently sent ? to growers saying...

...readers' comments in ? to the editor...

Wider contexts

...but simple block ? superimposed on the...

...lawyers recently sent ? to growers saying...

...readers' comments in ? to the editor...

What do we know about the word at "?"

Wider contexts

...but simple block ? superimposed on the...

...lawyers recently sent ? to growers saying...

...readers' comments in ? to the editor...

What do we know about the word at "?"
You send them, they come in a block variety, ...

Unordered contexts

{the, May, since} ? {planning, said, agency}

{future, measure} ? {., performance}

{government's, primary} ? {gauge, forecasting}

Unordered contexts

{the, May, since} ? {planning, said, agency}

{future, measure} ? {., performance}

{government's, primary} ? {gauge, forecasting}

What do we know about the word at "?"

Novel words

Lev stepped closer to the ?, which looked up at him.

The ?'s hand was warm, entirely handlike.

...the recent rental of a ?, one with potential as a weapon.

[Gibson 2014]

What’s a word?

Çekoslovakyalılaştıramadıklarımızdanmışsınız

(you are reportedly one of those who we were not able to turn into a Czechoslovakian)

The distributional hypothesis

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

J.R. Firth, *A Synopsis of Linguistic Theory*, 1957

How can we automate the process of constructing representations of word meaning from information about "company"?

Lexical semantics

Lexical semantics

How can we automate the process of constructing **representations of word meaning** from information about "company"?

What do we want from a representation of word meaning?

Types & syntactic roles

...it ? its...

...which ? the...

...the plate ? the table...

...will arrive ? Tuesday...

Is this word a noun?

A preposition?

What type of entity, event, or relation does it describe?

Selectional restrictions

Pat ate the ?.

The ? dripped down the sides of the bowl.

Pat caught the ?.

The ? smiled.

What sorts of actions can be performed on this word?

Is it animate?
Intelligent?
Solid?

Selectional restrictions

Pat ate the ?.

The ? dripped down the sides of the bowl.

Pat caught the ?.

The ? smiled.

???

What sorts of actions can be performed on this word?

Is it animate?
Intelligent?
Solid?

Lexical relations

antonymy

good / bad, black / white, above / below

synonymy

fear / dread, greeting / welcome, rise / increase

hypernymy

cat / animal, engine / entity, run / move

Perceptual features and grounding

egg, Switzerland, horse
which is biggest?

good, better, best
gray, black
which is most intense?

tomato
which one is it?

[images: Wikipedia]

Summary

Our word representations should capture information about

- types
- constraints on predicate-argument relations
- other relationships (hypernyms, antonyms, meronyms)
- perceptual features

How much of this can we get from context alone?

Co-occurrence statistics

The term-document matrix

Representational idea: construct a matrix where

rows are words
columns are contexts
entries indicate how many times word *i* appears in context *j*

$$W_{td} = \begin{array}{r} \\ cat \\ dog \\ the \end{array} \begin{array}{c} \begin{array}{ccccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 & 1 & 1 & 0 \\ 20 & 13 & 18 & 22 & 15 & 4 & 20 \end{bmatrix} \end{array}$$

The term-document matrix

Representational idea: construct a matrix where

rows are words
columns are contexts
entries indicate how many times word *i* appears in context *j*

$$
W_{td} = \begin{array}{r} \\ cat \\ dog \\ the \end{array} \begin{array}{c} d_1 \\ \left[\begin{array}{c} 1 \\ 0 \\ 20 \end{array}\right. \end{array}
$$

__The__ mouse I saw yesterday was bigger than __the__ biggest __cat__ I've ever seen...

The term-document matrix

Representational idea: construct a matrix where

rows are words
columns are contexts
entries indicate how many times word *i* appears in context *j*

$$W_{td} = \begin{array}{r} \\ cat \\ dog \\ the \end{array} \begin{array}{c} d_1 \\ \left[\begin{array}{c} 1 \\ 0 \\ 20 \end{array}\right. \end{array}$$

*the biggest **<u>cat</u>** I've ever...*

Term-document matrix: rows as word representations

the cat lifted its paw

the dog raised its paw

a nylon dog collar

the paw-shaped tag on the cat's collar

$$W_{td} = \begin{array}{r} \\ cat \\ dog \\ the \end{array} \begin{array}{c} \begin{array}{ccccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 \end{array} \\ \left[\begin{array}{ccccccc} 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 & 1 & 1 & 0 \\ 20 & 13 & 18 & 22 & 15 & 4 & 20 \end{array} \right] \end{array}$$

Term-document matrix: rows as word representations

Related words appear together!

*the **cat** lifted its **paw*** | *a nylon dog collar*

*the dog raised its **paw*** | *the **paw**-shaped tag on the **cat**'s collar*

$$W_{td} = \begin{array}{cc} & \begin{array}{cccc} d_1 & d_2 & d_3 & d_7 \end{array} \\ \begin{array}{c} \textit{cat} \\ \textit{paw} \end{array} & \begin{bmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix} \end{array}$$

Term-document matrix: rows as word representations

Related words (sometimes) appear together!

*the **cat** lifted its paw*

*a nylon **dog** collar*

*the **dog** raised its paw*

*the paw-shaped tag on the **cat**'s collar*

$$W_{td} = \begin{array}{r} \\ cat \\ dog \\ \\ \end{array} \begin{array}{c} \begin{array}{cccc} d_1 & d_2 & d_3 & \quad d_7 \end{array} \\ \left[\begin{array}{cccc} 1 & 0 & 0 & \quad 1 \\ 0 & 1 & 1 & \quad 0 \\ & & & \end{array} \right. \end{array}$$

Term-document matrix: rows as word representations

Related words **(sometimes)** appear together!

but document co-occurrence alone isn't a sufficient signal for semantic similarity.

Term-document matrix: rows as word representations

Related words **(sometimes)** appear together!

but document co-occurrence alone isn't a sufficient signal for semantic similarity.

words might be in strict alternation:

*the **cat** lifted its paw*

*a nylon **dog** collar*

*the **dog** raised its paw*

*the paw-shaped tag on the **cat**'s collar*

Term-document matrix: rows as word representations

Related words **(sometimes)** appear together!

but document co-occurrence alone isn't a sufficient signal for semantic similarity.

or co-occurrence statistics might be sparse:

$$W_{td} = \begin{array}{r} \\ cat \\ dog \end{array} \begin{array}{c} \begin{array}{ccccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 \end{array} \\ \left[\begin{array}{ccccccc} 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 & 1 & 1 & 0 \end{array} \right] \end{array}$$

Term-document matrix: rows as word representations

Related words **(sometimes)** appear together!

but document co-occurrence alone isn't a sufficient signal for semantic similarity.

or co-occurrence statistics might be sparse:

$$
\begin{array}{r c}
 & \begin{array}{ccccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 \end{array} \\
\begin{array}{r} \textit{cat} \\ \textit{scintillating} \end{array} & \left[\begin{array}{ccccccc} 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & 1 & \mathbf{0} \end{array} \right.
\end{array}
$$

The word co-occurrence matrix

Solution 1: low-rank approximation

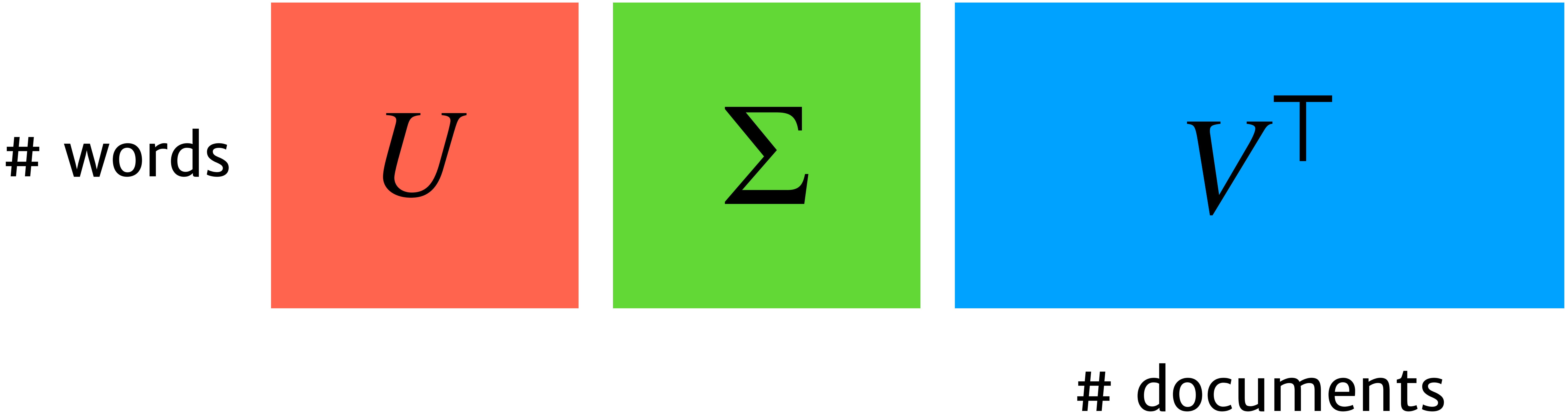
Theorem: for every $m \times n$ matrix A, there exists a factorization

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

with U and V orthonormal and Σ diagonal.

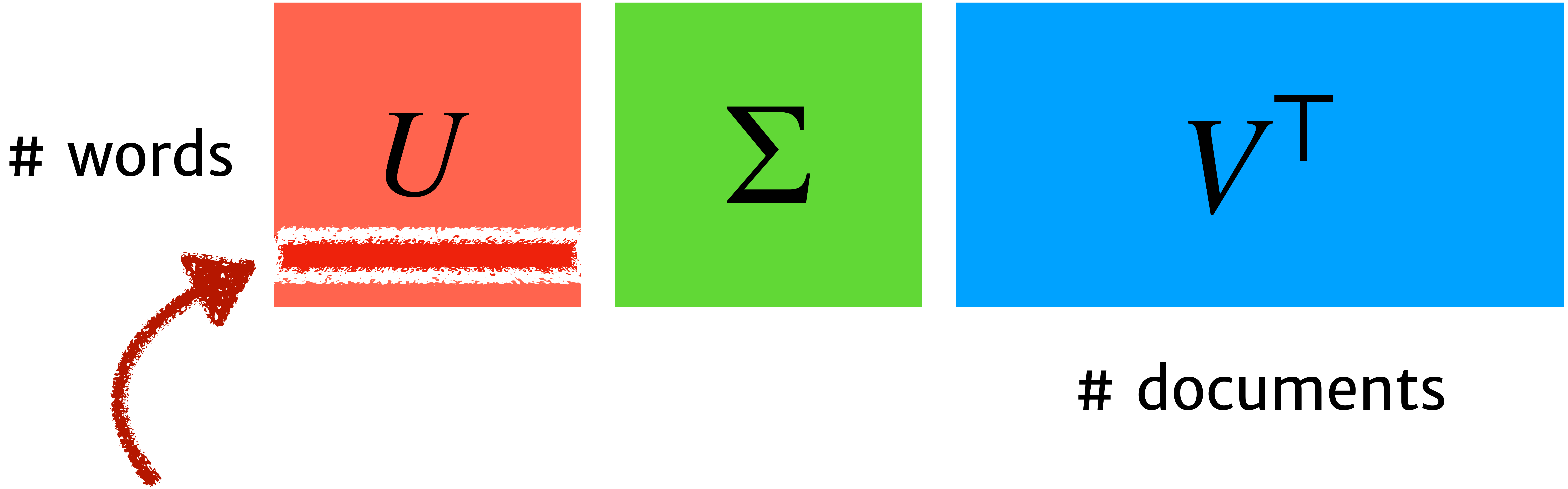
Latent Semantic Analysis

Solution 1: low-rank approximation



Latent Semantic Analysis

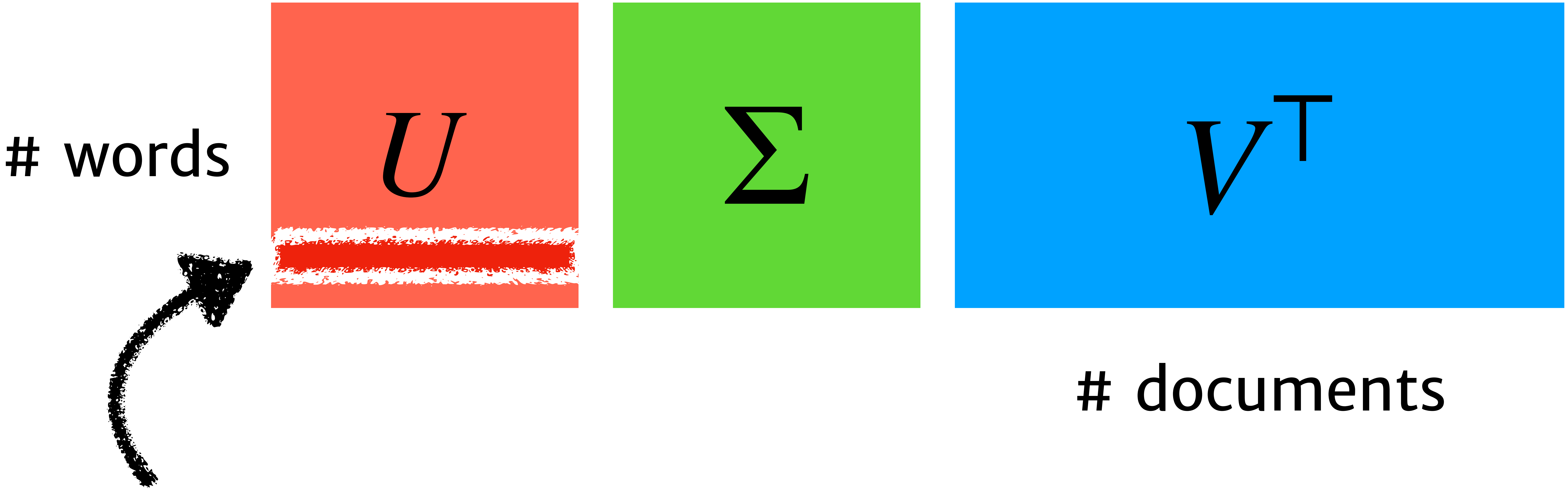
Solution 1: low-rank approximation



this is a word representation

Latent Semantic Analysis

Solution 1: low-rank approximation



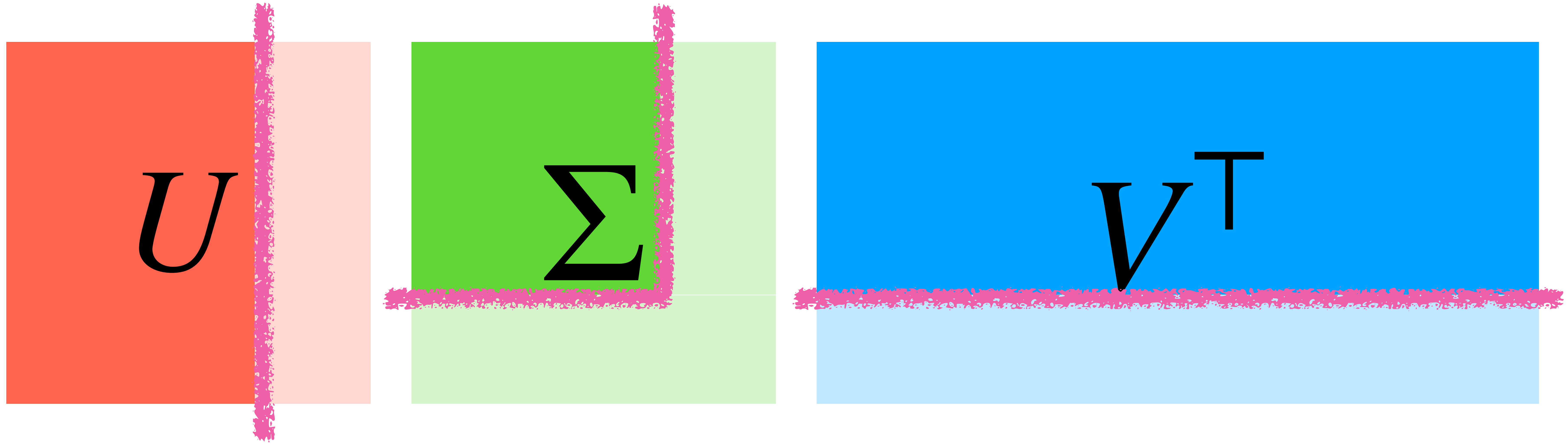
this is a word representation

vectors are all orthogonal!

Latent Semantic Analysis

Solution 1: low-rank approximation: truncate cols. of U and V

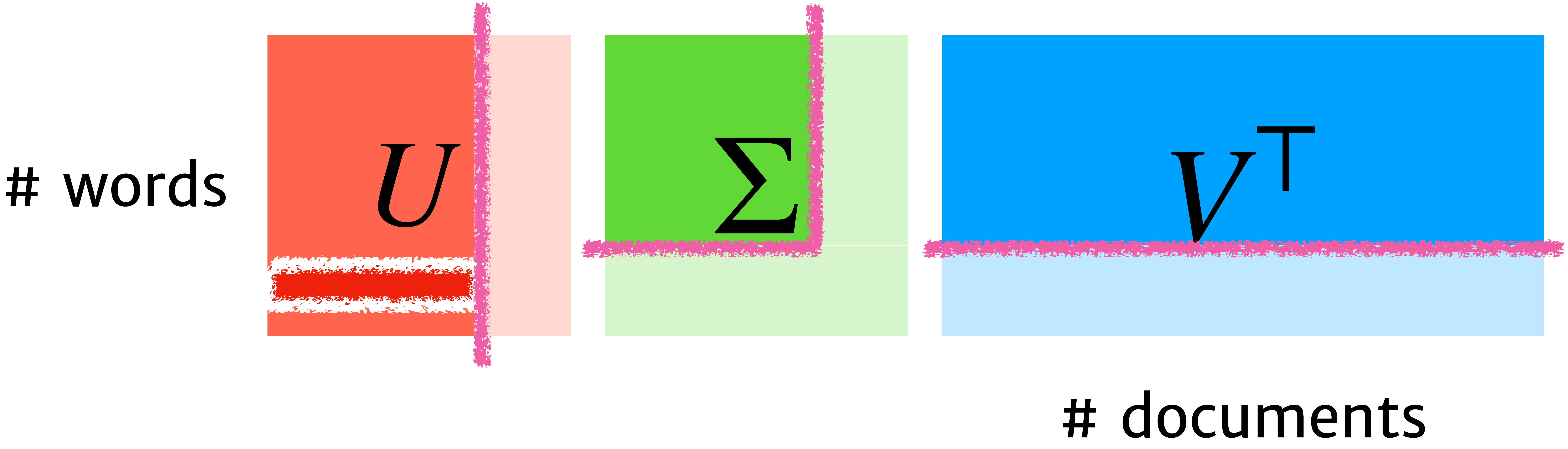
words



documents

Latent Semantic Analysis

Solution 1: low-rank approximation: truncate cols. of U and V



(**Theorem**: this is the best rank-k approx. to the original matrix)

Latent Semantic Analysis

$$\begin{array}{r} \\ cat \\ paw \\ algorithm \end{array} \begin{array}{c} \begin{array}{ccccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 \end{array} \\ \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \end{array}$$

Latent Semantic Analysis

$$\begin{array}{rc} & \begin{array}{ccccccc} \overbrace{d_1 \quad d_2 \quad d_3}^{\text{documents about animals}} & \overbrace{d_4 \quad d_5 \quad d_6 \quad d_7}^{\text{documents about computers}} \end{array} \\ \begin{array}{r} \textit{cat} \\ \textit{paw} \\ \textit{algorithm} \end{array} & \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \end{array}$$

Latent Semantic Analysis

$$U = \begin{array}{c} \begin{bmatrix} .7 & 0 & -.7 \\ .7 & 0 & .7 \\ 0 & 1 & 0 \end{bmatrix} \end{array} \begin{array}{l} cat \\ paw \\ algo. \end{array} \qquad V = \begin{bmatrix} .4 & 0 & -.7 \\ & \vdots & \\ 0 & 1 & 0 \end{bmatrix} \begin{array}{l} d_1 \\ \\ d_7 \end{array}$$

useless word reps!

$$\begin{array}{r} \\ cat \\ paw \\ algorithm \end{array} \begin{array}{c} \begin{array}{ccccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 \end{array} \\ \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \end{array}$$

Latent Semantic Analysis

$$U = \begin{array}{c l} \begin{bmatrix} .7 & 0 & \cancel{-.7} \\ .7 & 0 & \cancel{.7} \\ 0 & 1 & \cancel{0} \end{bmatrix} & \begin{array}{l} \textit{cat} \\ \\ \textit{algo.} \end{array} \end{array} \qquad V = \begin{array}{c l} \begin{bmatrix} .4 & 0 & \cancel{-.7} \\ & \vdots & \\ 0 & 1 & \cancel{0} \end{bmatrix} & \begin{array}{l} d_1 \\ \\ d_7 \end{array} \end{array}$$

useful word reps!

$$\begin{array}{r c} & \begin{array}{ccccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 \end{array} \\ \begin{array}{r} \textit{cat} \\ \textit{paw} \\ \textit{algorithm} \end{array} & \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \end{array}$$

Latent Semantic Analysis: Intuition

$$U = \begin{array}{cc} \begin{bmatrix} .7 & 0 & \cancel{-.7} \\ .7 & 0 & \cancel{.7} \\ 0 & 1 & \cancel{0} \end{bmatrix} & \begin{array}{l} \textit{cat} \\ \\ \textit{algo.} \end{array} \end{array} \qquad V = \begin{array}{cc} \begin{bmatrix} .4 & 0 & \cancel{-.7} \\ & \vdots & \\ 0 & 1 & \cancel{0} \end{bmatrix} & \begin{array}{l} d_1 \\ \\ d_7 \end{array} \end{array}$$

Most words don't appear in most documents, so dimensionality reduction techniques cluster words with similar contexts even when they don't co-occur.

Frequency effects

These counts are way bigger!

vector space similarity thinks they're more important
dimensionality reduction cares more about them

$$W_{td} = \begin{array}{r} \\ cat \\ dog \\ the \end{array} \begin{array}{c} \begin{array}{ccccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 & 1 & 1 & 0 \\ 20 & 13 & 18 & 22 & 15 & 4 & 20 \end{bmatrix} \end{array}$$

TF-IDF normalization

term frequency (tf):
\# of times word *i* appears in document *j*

inverse document frequency (idf):
log (# of documents / # of documents containing word *i*)

count'(i, j) = tf · idf

TF–IDF normalization

term frequency (tf):
\# of times word *w* appears in document *d*

inverse document frequency (idf):
log (# of documents / # of documents containing word *w*)

count'(i, j) = tf · idf

close to 0 if word *i* appears in almost every document

Pointwise mutual info. normalization

p(w) = # of times w appears in any document / word count

p(d) = fraction of documents identical to doc d

p(w, d) = # of times w and d appear together / (# words x # docs)

PMI(i, j) = p(w, d) / (p(w) p(d))

≈ p(d | w) if p(d) is roughly constant

Frequency effects

After weighting, these counts don't really matter

$$W_{td} = \begin{array}{r} \\ cat \\ dog \\ the \end{array} \begin{array}{c} \begin{array}{ccccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 \end{array} \\ \left[\begin{array}{ccccccc} .8 & .8 & 0 & .8 & 0 & .8 & 0 \\ 0 & 2.4 & 0 & 1.1 & 1.1 & 1.1 & 0 \\ .02 & .01 & .02 & .02 & .01 & 0 & .02 \end{array}\right] \end{array}$$

The word co-occurrence matrix

Solution 2: work in the word co-occurrence matrix

$$W_{tt} = \begin{array}{c} \\ cat \\ dog \\ the \end{array} \begin{array}{c} \begin{array}{ccc} cat & dog & the \end{array} \\ \begin{bmatrix} 10 & 8 & 103 \\ 8 & 20 & 97 \\ 103 & 97 & 995 \end{bmatrix} \end{array}$$

rows are words
columns are words
entries indicate how many times word *i* appears in the same context as word *j*

The word co-occurrence matrix

Solution 2: word co-occurrence matrix

$$W_{tt} = \begin{array}{c} \\ cat \\ dog \\ the \end{array} \begin{array}{c} \begin{array}{ccc} cat & dog & the \end{array} \\ \begin{bmatrix} 10 & 8 & 103 \\ 8 & 20 & 97 \\ 103 & 97 & 995 \end{bmatrix} \end{array}$$

Much smaller and less sparse.

The word co-occurrence matrix

Notice:

\# of times word *i* occurs in the same document as word *j*

$$= \sum_{\text{doc. } d} (\text{\# of times } i \text{ occurs in } d) \times (\text{\# of times } j \text{ occurs in } d)$$

$$= W_{td}[i, :]\ W_{td}[j, :]^\top$$

*i*th row of W_{td}

words that co-occur frequently have large row dot products in T-D matrix!

The word co-occurrence matrix

Notice:

# of times word i occurs in the same document as word j

$$= \sum_{\text{doc. } d} (\text{\# of times } i \text{ occurs in } d) \times (\text{\# of times } j \text{ occurs in } d)$$

$$= W_{td}[i, :]\ W_{td}[j, :]^\top \qquad W_{tt} = W_{td} W_{td}^\top$$

ith row of W_{td}

The word co-occurrence matrix

But: this matrix is still sparse at rare words.

$$W_{tt} = \begin{array}{r} \\ cat \\ the \\ loris \end{array} \begin{array}{c} \begin{array}{ccc} cat & the & loris \end{array} \\ \begin{bmatrix} 10 & 98 & \mathbf{0} \\ 98 & 20 & 1 \\ \mathbf{0} & 1 & 1 \end{bmatrix} \end{array}$$

But: this r

The word co-occurrence matrix

But: this matrix is still sparse at rare words.

Frequent words will continue to dominate similarity measurements.

$$W_{tt} = \begin{array}{r} \\ \textit{cat} \\ \textit{the} \\ \textit{loris} \end{array} \begin{array}{c} \begin{array}{ccc} \textit{cat} & \textit{the} & \textit{loris} \end{array} \\ \begin{bmatrix} 10 & \mathbf{98} & 0 \\ \mathbf{98} & 20 & \mathbf{1} \\ 0 & \mathbf{1} & 1 \end{bmatrix} \end{array}$$

The word co-occurrence matrix

But: this matrix is still sparse at rare words.

Frequent words will continue to dominate similarity measurements.

Still a good idea to do normalization (tf-idf / PMI) and rank reduction!

PMI revisited

p(i) = # of times *i* appears in any document / word count

p(i, j) = # of times *i* and *j* appear together / (# words)^2

PMI(i, j) = p(i, j) / (p(i) p(j))

$\cancel{\propto p(j \mid i)}$

PMI revisited

p(i) = # of times *i* appears in any document / word count

p(i, j) = # of times *i* and *j* appear together / (# words)^2

PMI(i, j) = p(i, j) / (p(i) p(j))

$\cancel{\approx p(j \mid i)}$

Do I see these words together more often than if they were independent?

Summary: constructing distributional word vectors

1. Estimate a matrix of co-occurrence statistics

term-document matrix

$$W_{td} = \begin{array}{r} \\ cat \\ dog \\ the \end{array} \begin{array}{c} \begin{array}{ccccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 & 1 & 1 & 0 \\ 20 & 13 & 18 & 22 & 15 & 4 & 20 \end{bmatrix} \end{array}$$

Summary: constructing distributional word vectors

1. Estimate a matrix of co-occurrence statistics

word co-occurrence matrix

$$W_{tt} = \begin{array}{c c} & \begin{array}{ccc} \textit{cat} & \textit{dog} & \textit{the} \end{array} \\ \begin{array}{c} \textit{cat} \\ \textit{dog} \\ \textit{the} \end{array} & \begin{bmatrix} 10 & 8 & 103 \\ 8 & 20 & 97 \\ 103 & 97 & 995 \end{bmatrix} \end{array}$$

Summary: constructing distributional word vectors

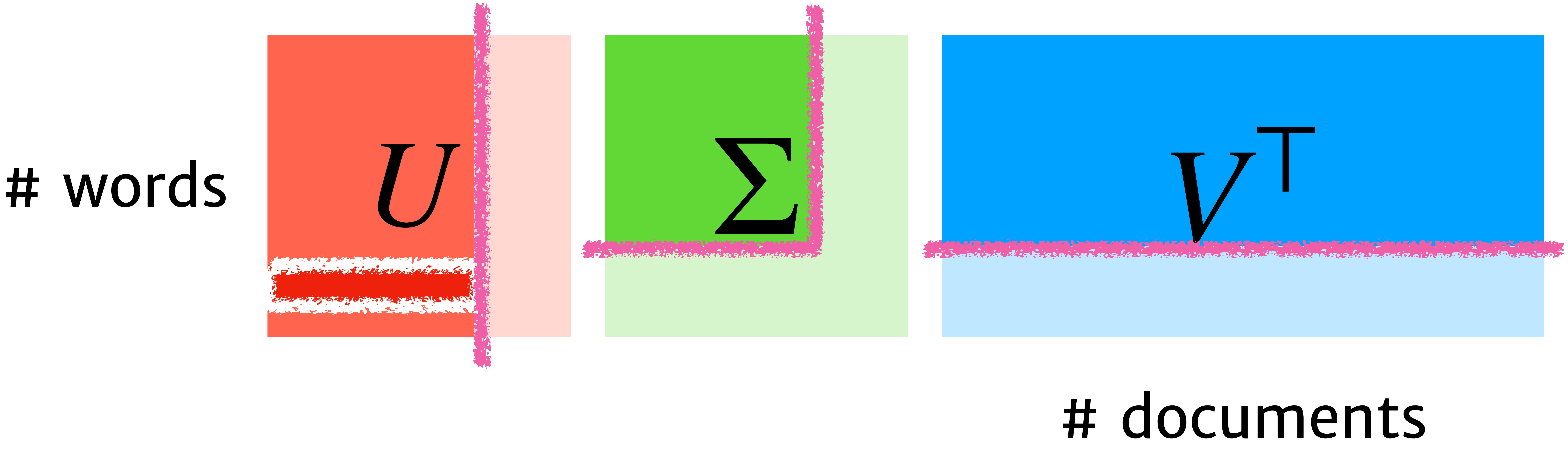
1. Estimate a matrix of co-occurrence statistics
2. Normalize / smooth counts

TF-IDF: # occurrences of word i
$\times$ log (# documents / # documents containing i)

PMI: p(word i, word j) / [p(word i) p(word j)]

Summary: constructing distributional word vectors

1. Estimate a matrix of co-occurrence statistics
2. Normalize / smooth counts
3. Take a low-rank approximation



Summary: constructing distributional word vectors

1. Estimate a matrix of co-occurrence statistics
2. Normalize / smooth counts
3. Take a low-rank approximation
4. ???
5. Profit!

Did we win?

Our word representations should capture information about

- types
- constraints on predicate-argument relations
- other relationships (hypernyms, antonyms, meronyms)
- perceptual features

Did we win?

Our word representations should capture information about

- types
  - yes! if context information preserves ordering, words with similar types appear in similar contexts (verbs come after nouns etc.)

Did we win?

Our word representations should capture information about

- types
- constraints on predicate-argument relations
- yes! dot products between word vectors predict frequency of co-occurrence

Did we win?

Our word representations should capture information about

- types
- constraints on predicate-argument relations
- other relationships (hypernyms, antonyms, meronyms)
- maybe? e.g. antonyms have the same type but appear in disjoint contexts

Did we win?

Our word representations should capture information about

- types
- constraints on predicate-argument relations
- other relationships (hypernyms, antonyms, meronyms)
- perceptual features
- ??? *plant* is close to *green*, but what does *green* look like?

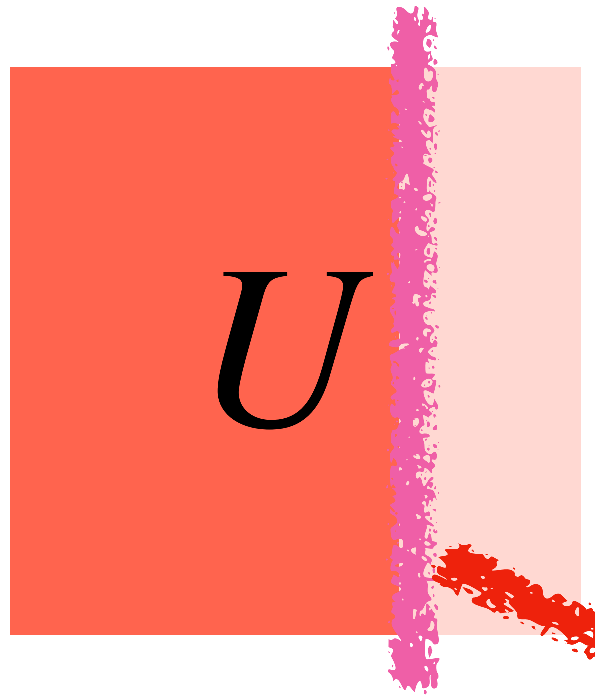
In models

$$s = W_2^\top f(W_1^\top x)$$

x $\qquad$ $f(W_1^\top x) = h_1$ $\qquad$ $W_2^\top h_1 = s$

$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$ input $\qquad$ $\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$ "hidden layer" $\qquad$ $\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$ output

In models

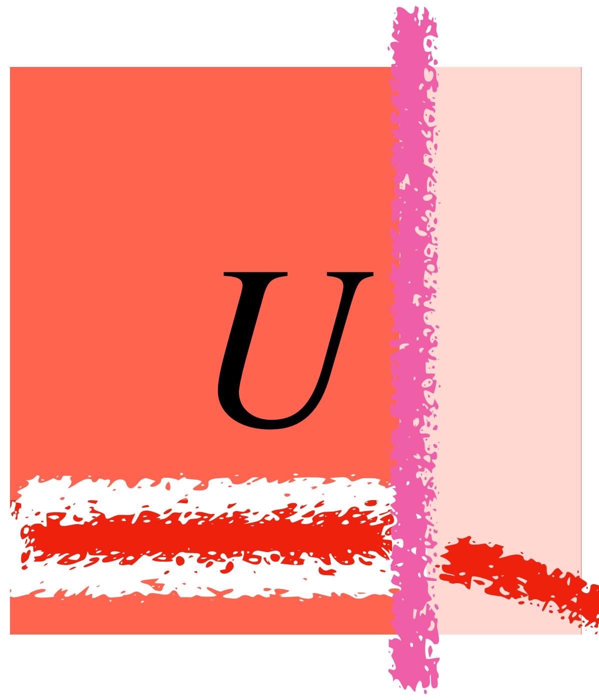


$$s = W_2^\top f(W_1^\top x)$$

x $\qquad$ $f(W_1^\top x) = h_1$ $\qquad$ $W_2^\top h_1 = s$

$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$ input $\qquad$ $\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$ “hidden layer” $\qquad$ $\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$ output

In models



$$s = W_2^\top f(W_1^\top x)$$

$$w_{1,i}^\top$$

$$\begin{bmatrix} 0.1 \\ 1.7 \\ 0.3 \end{bmatrix}\begin{bmatrix} 0.1 \\ 1.7 \\ 0.3 \end{bmatrix}\begin{bmatrix} 0.1 \\ 1.7 \\ 0.3 \end{bmatrix} \xrightarrow{f(\sum_{i:\ x_i=1} w_i)} \begin{bmatrix} 1.2 \\ 1.8 \\ -0.1 \end{bmatrix} \longrightarrow 0.9$$

x *very good, but gave me*

Bias in distributional representations

$$
\begin{array}{c|c}
 & \begin{array}{cccc} \textit{man} & \textit{woman} & \textit{doctor} & \textit{nurse} \end{array} \\
\begin{array}{r} \textit{man} \\ \textit{woman} \\ \textit{doctor} \\ \textit{nurse} \end{array} & \begin{bmatrix} & 100 & 45 & 23 \\ 100 & & 18 & 48 \\ 45 & 18 & & 80 \\ 23 & 48 & 80 & \end{bmatrix}
\end{array}
$$

Bias in distributional representations

$$
\begin{array}{r c}
 & \begin{array}{cccc} \textit{man} & \textit{woman} & \textit{doctor} & \textit{nurse} \end{array} \\
\begin{array}{r} \textit{man} \\ \textit{woman} \\ \textit{doctor} \\ \textit{nurse} \end{array} &
\begin{bmatrix} & 100 & 45 & 23 \\ 100 & & 18 & 48 \\ 45 & 18 & & 80 \\ 23 & 48 & 80 & \end{bmatrix}
\end{array}
$$

What word is most similar to *man*?
What’s next?

Bias in distributional representations

$$
\begin{array}{c|c}
 & \begin{matrix} \textit{man} & \textit{woman} & \textit{doctor} & \textit{nurse} \end{matrix} \\
\begin{matrix} \textit{man} \\ \textit{woman} \\ \textit{doctor} \\ \textit{nurse} \end{matrix} &
\begin{bmatrix}
 & 100 & 45 & 23 \\
100 & & 18 & 48 \\
45 & 18 & & 80 \\
23 & 48 & 80 &
\end{bmatrix}
\end{array}
$$

It's usually most helpful to think of bias as a property of decisions, not parameters.

But how will these parameters influence decisions?

Next class: more word embeddings