Module 10.3: SVD for learning word representations

$$\begin{bmatrix} & & \\ & X & \\ & & \end{bmatrix}_{m \times n} =$$

$$\begin{bmatrix} \uparrow & \cdots & \uparrow \\ u_1 & \cdots & u_k \\ \downarrow & \cdots & \downarrow \end{bmatrix}_{m \times k} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{bmatrix}_{k \times k} \begin{bmatrix} \leftarrow & v_1^T & \rightarrow \\ & \vdots & \\ \leftarrow & v_k^T & \rightarrow \end{bmatrix}_{k \times n}$$

- Singular Value Decomposition gives a rank k approximation of the original matrix

$$X = X_{PPMI\, m \times n} = U_{m \times k} \Sigma_{k \times k} V^T_{k \times n}$$

X_{PPMI} (simplifying notation to X) is the co-occurrence matrix with PPMI values

$$\begin{bmatrix} & & \\ & X & \\ & & \end{bmatrix}_{m \times n} =$$

$$\begin{bmatrix} \uparrow & \cdots & \uparrow \\ u_1 & \cdots & u_k \\ \downarrow & \cdots & \downarrow \end{bmatrix}_{m \times k} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{bmatrix}_{k \times k} \begin{bmatrix} \leftarrow & v_1^T & \rightarrow \\ & \vdots & \\ \leftarrow & v_k^T & \rightarrow \end{bmatrix}_{k \times n}$$

- Singular Value Decomposition gives a rank k approximation of the original matrix

$$X = X_{PPMI_{m \times n}} = U_{m \times k} \Sigma_{k \times k} V_{k \times n}^T$$

X_{PPMI} (simplifying notation to X) is the co-occurrence matrix with PPMI values

- SVD gives the best rank-k approximation of the original data (X)

$$\begin{bmatrix} & & \\ & X & \\ & & \end{bmatrix}_{m \times n} =$$

$$\begin{bmatrix} \uparrow & \cdots & \uparrow \\ & & \\ u_1 & \cdots & u_k \\ \downarrow & \cdots & \downarrow \end{bmatrix}_{m \times k} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{bmatrix}_{k \times k} \begin{bmatrix} \leftarrow & v_1^T & \rightarrow \\ & \vdots & \\ \leftarrow & v_k^T & \rightarrow \end{bmatrix}_{k \times n}$$

- Singular Value Decomposition gives a rank k approximation of the original matrix

  $$X = X_{PPMI_{m \times n}} = U_{m \times k} \Sigma_{k \times k} V^T_{k \times n}$$

  X_{PPMI} (simplifying notation to X) is the co-occurrence matrix with PPMI values
- SVD gives the best rank-k approximation of the original data (X)
- Discovers latent semantics in the corpus (let us examine this with the help of an example)

- Notice that the product can be written as a sum of k rank-1 matrices

$$\begin{bmatrix} & & \\ & X & \\ & & \end{bmatrix}_{m\times n} =$$

$$\begin{bmatrix} \uparrow & \cdots & \uparrow \\ & & \\ u_1 & \cdots & u_k \\ \downarrow & \cdots & \downarrow \end{bmatrix}_{m\times k} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{bmatrix}_{k\times k} \begin{bmatrix} \leftarrow & v_1^T & \rightarrow \\ & \vdots & \\ \leftarrow & v_k^T & \rightarrow \end{bmatrix}_{k\times n}$$

$$= \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_k u_k v_k^T$$

$$\begin{bmatrix} & & \\ & X & \\ & & \end{bmatrix}_{m \times n} =$$

$$\begin{bmatrix} \uparrow & \cdots & \uparrow \\ u_1 & \cdots & u_k \\ \downarrow & \cdots & \downarrow \end{bmatrix}_{m \times k} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{bmatrix}_{k \times k} \begin{bmatrix} \leftarrow & v_1^T & \rightarrow \\ & \vdots & \\ \leftarrow & v_k^T & \rightarrow \end{bmatrix}_{k \times n}$$

$$= \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_k u_k v_k^T$$

- Notice that the product can be written as a sum of k rank-1 matrices
- Each $\sigma_i u_i v_i^T \in \mathbb{R}^{m \times n}$ because it is a product of a $m \times 1$ vector with a $1 \times n$ vector

$$\begin{bmatrix} & & \\ & X & \\ & & \end{bmatrix}_{m\times n} =$$

$$\begin{bmatrix} \uparrow & \cdots & \uparrow \\ u_1 & \cdots & u_k \\ \downarrow & \cdots & \downarrow \end{bmatrix}_{m\times k} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{bmatrix}_{k\times k} \begin{bmatrix} \leftarrow & v_1^T & \rightarrow \\ & \vdots & \\ \leftarrow & v_k^T & \rightarrow \end{bmatrix}_{k\times n}$$

$$= \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_k u_k v_k^T$$

- Notice that the product can be written as a sum of k rank-1 matrices
- Each $\sigma_i u_i v_i^T \in \mathbb{R}^{m\times n}$ because it is a product of a $m \times 1$ vector with a $1 \times n$ vector
- If we truncate the sum at $\sigma_1 u_1 v_1^T$ then we get the best rank-1 approximation of X

$$\begin{bmatrix} & & \\ & X & \\ & & \end{bmatrix}_{m \times n} =$$

$$\begin{bmatrix} \uparrow & \cdots & \uparrow \\ u_1 & \cdots & u_k \\ \downarrow & \cdots & \downarrow \end{bmatrix}_{m \times k} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{bmatrix}_{k \times k} \begin{bmatrix} \leftarrow & v_1^T & \rightarrow \\ & \vdots & \\ \leftarrow & v_k^T & \rightarrow \end{bmatrix}_{k \times n}$$

$$= \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_k u_k v_k^T$$

- Notice that the product can be written as a sum of k rank-1 matrices
- Each $\sigma_i u_i v_i^T \in \mathbb{R}^{m \times n}$ because it is a product of a $m \times 1$ vector with a $1 \times n$ vector
- If we truncate the sum at $\sigma_1 u_1 v_1^T$ then we get the best rank-1 approximation of X (By SVD theorem! But what does this mean? We will see on the next slide)

$$\begin{bmatrix} & & \\ & X & \\ & & \end{bmatrix}_{m\times n} =$$

$$\begin{bmatrix} \uparrow & \cdots & \uparrow \\ u_1 & \cdots & u_k \\ \downarrow & \cdots & \downarrow \end{bmatrix}_{m\times k} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{bmatrix}_{k\times k} \begin{bmatrix} \leftarrow & v_1^T & \rightarrow \\ & \vdots & \\ \leftarrow & v_k^T & \rightarrow \end{bmatrix}_{k\times n}$$

$$= \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_k u_k v_k^T$$

- Notice that the product can be written as a sum of k rank-1 matrices
- Each $\sigma_i u_i v_i^T \in \mathbb{R}^{m\times n}$ because it is a product of a $m \times 1$ vector with a $1 \times n$ vector
- If we truncate the sum at $\sigma_1 u_1 v_1^T$ then we get the best rank-1 approximation of X (By SVD theorem! But what does this mean? We will see on the next slide)
- If we truncate the sum at $\sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T$ then we get the best rank-2 approximation of X and so on

- What do we mean by approximation here?

$$\begin{bmatrix} & & \\ & X & \\ & & \end{bmatrix}_{m\times n} =$$

$$\begin{bmatrix} \uparrow & \cdots & \uparrow \\ u_1 & \cdots & u_k \\ \downarrow & \cdots & \downarrow \end{bmatrix}_{m\times k} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{bmatrix}_{k\times k} \begin{bmatrix} \leftarrow & v_1^T & \rightarrow \\ & \vdots & \\ \leftarrow & v_k^T & \rightarrow \end{bmatrix}_{k\times n}$$

$$= \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_k u_k v_k^T$$

- What do we mean by approximation here?
- Notice that X has $m \times n$ entries

$$\begin{bmatrix} & & \\ & X & \\ & & \end{bmatrix}_{m \times n} =$$

$$\begin{bmatrix} \uparrow & \cdots & \uparrow \\ u_1 & \cdots & u_k \\ \downarrow & \cdots & \downarrow \end{bmatrix}_{m \times k} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{bmatrix}_{k \times k} \begin{bmatrix} \leftarrow & v_1^T & \rightarrow \\ & \vdots & \\ \leftarrow & v_k^T & \rightarrow \end{bmatrix}_{k \times n}$$

$$= \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_k u_k v_k^T$$

$$\begin{bmatrix} & & \\ & X & \\ & & \end{bmatrix}_{m \times n} =$$

$$\begin{bmatrix} \uparrow & \cdots & \uparrow \\ u_1 & \cdots & u_k \\ \downarrow & \cdots & \downarrow \end{bmatrix}_{m \times k} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{bmatrix}_{k \times k} \begin{bmatrix} \leftarrow & v_1^T & \rightarrow \\ & \vdots & \\ \leftarrow & v_k^T & \rightarrow \end{bmatrix}_{k \times n}$$

$$= \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_k u_k v_k^T$$

- What do we mean by approximation here?
- Notice that X has $m \times n$ entries
- When we use he rank-1 approximation we are using only $n + m + 1$ entries to reconstruct $[u \in \mathbb{R}^m, v \in \mathbb{R}^n, \sigma \in \mathbb{R}^1]$

$$\begin{bmatrix} & & \\ & X & \\ & & \end{bmatrix}_{m \times n} =$$

$$\begin{bmatrix} \uparrow & \cdots & \uparrow \\ u_1 & \cdots & u_k \\ \downarrow & \cdots & \downarrow \end{bmatrix}_{m \times k} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{bmatrix}_{k \times k} \begin{bmatrix} \leftarrow & v_1^T & \rightarrow \\ & \vdots & \\ \leftarrow & v_k^T & \rightarrow \end{bmatrix}_{k \times n}$$

$$= \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_k u_k v_k^T$$

- What do we mean by approximation here?
- Notice that X has $m \times n$ entries
- When we use he rank-1 approximation we are using only $n + m + 1$ entries to reconstruct $[u \in \mathbb{R}^m, v \in \mathbb{R}^n, \sigma \in \mathbb{R}^1]$
- But SVD theorem tells us that u_1,v_1 and σ_1 store the most information in X (akin to the principal components in X)

$$
\begin{bmatrix} & & \\ & X & \\ & & \end{bmatrix}_{m\times n} =
$$

$$
\begin{bmatrix} \uparrow & \cdots & \uparrow \\ u_1 & \cdots & u_k \\ \downarrow & \cdots & \downarrow \end{bmatrix}_{m\times k}
\begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{bmatrix}_{k\times k}
\begin{bmatrix} \leftarrow & v_1^T & \rightarrow \\ & \vdots & \\ \leftarrow & v_k^T & \rightarrow \end{bmatrix}_{k\times n}
$$

$$
= \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_k u_k v_k^T
$$

- What do we mean by approximation here?
- Notice that X has $m \times n$ entries
- When we use he rank-1 approximation we are using only $n + m + 1$ entries to reconstruct $[u \in \mathbb{R}^m, v \in \mathbb{R}^n, \sigma \in \mathbb{R}^1]$
- But SVD theorem tells us that u_1,v_1 and σ_1 store the most information in X (akin to the principal components in X)
- Each subsequent term $(\sigma_2 u_2 v_2^T,\ \sigma_3 u_3 v_3^T,\ \ldots)$ stores less and less important information

$$\overbrace{\begin{array}{|c|c|c|c|}\hline 0 & 0 & 0 & 1 \\ \hline\end{array}}^{verylight}\overbrace{\begin{array}{|c|c|c|c|}\hline 1 & 0 & 1 & 1 \\ \hline\end{array}}^{green}$$

$$\overbrace{\begin{array}{|c|c|c|c|}\hline 0 & 0 & 1 & 0 \\ \hline\end{array}}^{light}\overbrace{\begin{array}{|c|c|c|c|}\hline 1 & 0 & 1 & 1 \\ \hline\end{array}}^{green}$$

$$\overbrace{\begin{array}{|c|c|c|c|}\hline 0 & 1 & 0 & 0 \\ \hline\end{array}}^{dark}\overbrace{\begin{array}{|c|c|c|c|}\hline 1 & 0 & 1 & 1 \\ \hline\end{array}}^{green}$$

$$\overbrace{\begin{array}{|c|c|c|c|}\hline 1 & 0 & 0 & 0 \\ \hline\end{array}}^{verydark}\overbrace{\begin{array}{|c|c|c|c|}\hline 1 & 0 & 1 & 1 \\ \hline\end{array}}^{green}$$

- As an analogy consider the case when we are using 8 bits to represent colors

$$\overbrace{\boxed{0}\boxed{0}\boxed{0}\boxed{1}}^{verylight}\overbrace{\boxed{1}\boxed{0}\boxed{1}\boxed{1}}^{green}$$

$$\overbrace{\boxed{0}\boxed{0}\boxed{1}\boxed{0}}^{light}\overbrace{\boxed{1}\boxed{0}\boxed{1}\boxed{1}}^{green}$$

$$\overbrace{\boxed{0}\boxed{1}\boxed{0}\boxed{0}}^{dark}\overbrace{\boxed{1}\boxed{0}\boxed{1}\boxed{1}}^{green}$$

$$\overbrace{\boxed{1}\boxed{0}\boxed{0}\boxed{0}}^{verydark}\overbrace{\boxed{1}\boxed{0}\boxed{1}\boxed{1}}^{green}$$

- As an analogy consider the case when we are using 8 bits to represent colors
- The representation of very light, light, dark and very dark green would look different

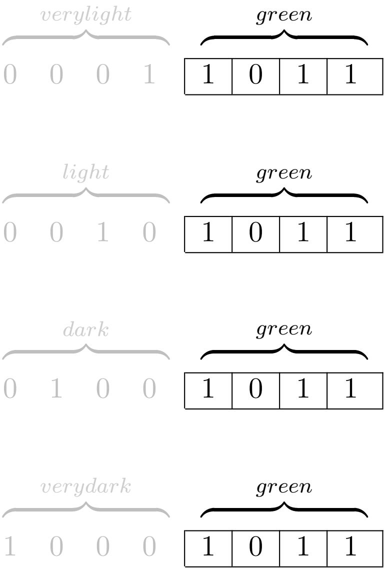
$$\overbrace{\boxed{0}\boxed{0}\boxed{0}\boxed{1}}^{verylight}\ \overbrace{\boxed{1}\boxed{0}\boxed{1}\boxed{1}}^{green}$$

$$\overbrace{\boxed{0}\boxed{0}\boxed{1}\boxed{0}}^{light}\ \overbrace{\boxed{1}\boxed{0}\boxed{1}\boxed{1}}^{green}$$

$$\overbrace{\boxed{0}\boxed{1}\boxed{0}\boxed{0}}^{dark}\ \overbrace{\boxed{1}\boxed{0}\boxed{1}\boxed{1}}^{green}$$

$$\overbrace{\boxed{1}\boxed{0}\boxed{0}\boxed{0}}^{verydark}\ \overbrace{\boxed{1}\boxed{0}\boxed{1}\boxed{1}}^{green}$$

- As an analogy consider the case when we are using 8 bits to represent colors
- The representation of very light, light, dark and very dark green would look different
- But now what if we were asked to compress this into 4 bits? (akin to compressing $m \times m$ values into $m + m + 1$ values on the previous slide)



- As an analogy consider the case when we are using 8 bits to represent colors
- The representation of very light, light, dark and very dark green would look different
- But now what if we were asked to compress this into 4 bits? (akin to compressing $m \times m$ values into $m + m + 1$ values on the previous slide)
- We will retain the most important 4 bits and now the previously (slightly) latent similarity between the colors now becomes very obvious

$$\overbrace{0\quad 0\quad 0\quad 1}^{verylight}\quad \overbrace{\boxed{1}\boxed{0}\boxed{1}\boxed{1}}^{green}$$

$$\overbrace{0\quad 0\quad 1\quad 0}^{light}\quad \overbrace{\boxed{1}\boxed{0}\boxed{1}\boxed{1}}^{green}$$

$$\overbrace{0\quad 1\quad 0\quad 0}^{dark}\quad \overbrace{\boxed{1}\boxed{0}\boxed{1}\boxed{1}}^{green}$$

$$\overbrace{1\quad 0\quad 0\quad 0}^{verydark}\quad \overbrace{\boxed{1}\boxed{0}\boxed{1}\boxed{1}}^{green}$$

- As an analogy consider the case when we are using 8 bits to represent colors
- The representation of very light, light, dark and very dark green would look different
- But now what if we were asked to compress this into 4 bits? (akin to compressing $m \times m$ values into $m + m + 1$ values on the previous slide)
- We will retain the most important 4 bits and now the previously (slightly) latent similarity between the colors now becomes very obvious
- Something similar is guaranteed by SVD (retain the most important information and discover the latent similarities between words)

	human	**machine**	**system**	**for**	...	**user**
human	0	2.944	0	2.25	...	**0**
machine	2.944	0	0	2.25	...	0
system	0	**0**	0	1.15	...	1.84
for	2.25	2.25	1.15	0	...	0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	0	0	1.84	0	...	0

Co-occurrence Matrix (X)

	human	**machine**	**system**	**for**	...	**user**
human	2.01	2.01	0.23	2.14	...	**0.43**
machine	2.01	2.01	0.23	2.14	...	0.43
system	0.23	**0.23**	1.17	0.96	...	1.29
for	2.14	2.14	0.96	1.87	...	-0.13
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	0.43	0.43	1.29	-0.13	...	1.71

Low rank $X \rightarrow$ Low rank $\hat{X}$

- Notice that after low rank reconstruction with SVD, the latent co-occurrence between $\{system, machine\}$ and $\{human, user\}$ has become visible

$$X =$$

	human	machine	system	for	...	user
human	0	2.944	0	2.25	...	0
machine	2.944	0	0	2.25	...	0
system	0	0	0	1.15	...	1.84
for	2.25	2.25	1.15	0	...	0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	0	0	1.84	0	...	0

- Recall that earlier each row of the original matrix X served as the representation of a word

$$X =$$

	human	**machine**	**system**	**for**	...	**user**
human	0	2.944	0	2.25	...	0
machine	2.944	0	0	2.25	...	0
system	0	0	0	1.15	...	1.84
for	2.25	2.25	1.15	0	...	0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	0	0	1.84	0	...	0

$$XX^T =$$

	human	**machine**	**system**	**for**	...	**user**
human	32.5	23.9	7.78	20.25	...	7.01
machine	23.9	32.5	7.78	20.25	...	7.01
system	7.78	7.78	0	17.65	...	21.84
for	20.25	20.25	17.65	36.3	...	11.8
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	7.01	7.01	21.84	11.8	...	28.3

$$cosine\_sim(human, user) = 0.21$$

- Recall that earlier each row of the original matrix X served as the representation of a word
- Then XX^T is a matrix whose ij-th entry is the dot product between the representation of word i ($X[i:]$) and word j ($X[j:]$)

$$X =$$

	human	machine	system	for	...	user
human	0	2.944	0	2.25	...	0
machine	2.944	0	0	2.25	...	0
system	0	0	0	1.15	...	1.84
for	2.25	2.25	1.15	0	...	0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	0	0	1.84	0	...	0

$$XX^T =$$

	human	machine	system	for	...	user
human	32.5	23.9	7.78	20.25	...	7.01
machine	23.9	32.5	7.78	20.25	...	7.01
system	7.78	7.78	0	17.65	...	21.84
for	20.25	20.25	17.65	36.3	...	11.8
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	7.01	7.01	21.84	11.8	...	28.3

$$cosine\_sim(human, user) = 0.21$$

- Recall that earlier each row of the original matrix X served as the representation of a word
- Then XX^T is a matrix whose ij-th entry is the dot product between the representation of word i ($X[i:]$) and word j ($X[j:]$)

$$X[i:]$$

$$X[j:]$$

$$X =$$

	human	machine	system	for	...	user
human	0	2.944	0	2.25	...	0
machine	2.944	0	0	2.25	...	0
system	0	0	0	1.15	...	1.84
for	2.25	2.25	1.15	0	...	0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	0	0	1.84	0	...	0

$$XX^T =$$

	human	machine	system	for	...	user
human	32.5	23.9	7.78	20.25	...	7.01
machine	23.9	32.5	7.78	20.25	...	7.01
system	7.78	7.78	0	17.65	...	21.84
for	20.25	20.25	17.65	36.3	...	11.8
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	7.01	7.01	21.84	11.8	...	28.3

$$cosine\_sim(human, user) = 0.21$$

- Recall that earlier each row of the original matrix X served as the representation of a word
- Then XX^T is a matrix whose ij-th entry is the dot product between the representation of word i ($X[i:]$) and word j ($X[j:]$)

$$\begin{array}{r} X[i:] \\ \\ X[j:] \end{array} \underbrace{\begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 0 \\ 1 & 3 & 5 \end{bmatrix}}_{X}$$

$$X =$$

	human	machine	system	for	...	user
human	0	2.944	0	2.25	...	0
machine	2.944	0	0	2.25	...	0
system	0	0	0	1.15	...	1.84
for	2.25	2.25	1.15	0	...	0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	0	0	1.84	0	...	0

$$XX^T =$$

	human	machine	system	for	...	user
human	32.5	23.9	7.78	20.25	...	7.01
machine	23.9	32.5	7.78	20.25	...	7.01
system	7.78	7.78	0	17.65	...	21.84
for	20.25	20.25	17.65	36.3	...	11.8
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	7.01	7.01	21.84	11.8	...	28.3

$$cosine\_sim(human, user) = 0.21$$

- Recall that earlier each row of the original matrix X served as the representation of a word
- Then XX^T is a matrix whose ij-th entry is the dot product between the representation of word i ($X[i:]$) and word j ($X[j:]$)

$$\begin{array}{r} X[i:] \\ \\ X[j:] \end{array} \underbrace{\begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 0 \\ 1 & 3 & 5 \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} 1 & 2 & 1 \\ 2 & 1 & 3 \\ 3 & 0 & 5 \end{bmatrix}}_{X^T}$$

$$X =$$

	human	machine	system	for	...	user
human	0	2.944	0	2.25	...	0
machine	2.944	0	0	2.25	...	0
system	0	0	0	1.15	...	1.84
for	2.25	2.25	1.15	0	...	0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	0	0	1.84	0	...	0

$$XX^T =$$

	human	machine	system	for	...	user
human	32.5	23.9	7.78	20.25	...	7.01
machine	23.9	32.5	7.78	20.25	...	7.01
system	7.78	7.78	0	17.65	...	21.84
for	20.25	20.25	17.65	36.3	...	11.8
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	7.01	7.01	21.84	11.8	...	28.3

$$cosine\_sim(human, user) = 0.21$$

- Recall that earlier each row of the original matrix X served as the representation of a word
- Then XX^T is a matrix whose ij-th entry is the dot product between the representation of word i ($X[i:]$) and word j ($X[j:]$)

$$\begin{array}{c} X[i:] \\ \\ X[j:] \end{array} \underbrace{\begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 0 \\ 1 & 3 & 5 \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} 1 & 2 & 1 \\ 2 & 1 & 3 \\ 3 & 0 & 5 \end{bmatrix}}_{X^T}$$

$$= \underbrace{\begin{bmatrix} . & . & 22 \\ . & . & . \\ . & . & . \end{bmatrix}}_{XX^T}$$

$$X =$$

	human	machine	system	for	...	user
human	0	2.944	0	2.25	...	0
machine	2.944	0	0	2.25	...	0
system	0	0	0	1.15	...	1.84
for	2.25	2.25	1.15	0	...	0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	0	0	1.84	0	...	0

$$XX^T =$$

	human	machine	system	for	...	user
human	32.5	23.9	7.78	20.25	...	7.01
machine	23.9	32.5	7.78	20.25	...	7.01
system	7.78	7.78	0	17.65	...	21.84
for	20.25	20.25	17.65	36.3	...	11.8
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	7.01	7.01	21.84	11.8	...	28.3

$$cosine\_sim(human, user) = 0.21$$

- Recall that earlier each row of the original matrix X served as the representation of a word
- Then XX^T is a matrix whose ij-th entry is the dot product between the representation of word i ($X[i:]$) and word j ($X[j:]$)

$$\begin{array}{c} X[i:] \\ \\ X[j:] \end{array} \underbrace{\begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 0 \\ 1 & 3 & 5 \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} 1 & 2 & 1 \\ 2 & 1 & 3 \\ 3 & 0 & 5 \end{bmatrix}}_{X^T}$$

$$= \underbrace{\begin{bmatrix} . & . & 22 \\ . & . & . \\ . & . & . \end{bmatrix}}_{XX^T}$$

- The ij-th entry of XX^T thus (roughly) captures the cosine similarity between $word_i, word_j$

- Once we do an SVD what is a good choice for the representation of $word_i$?

$$\hat{X} =$$

	human	**machine**	**system**	**for**	...	**user**
human	2.01	2.01	0.23	2.14	...	0.43
machine	2.01	2.01	0.23	2.14	...	0.43
system	0.23	0.23	1.17	0.96	...	1.29
for	2.14	2.14	0.96	1.87	...	-0.13
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	0.43	0.43	1.29	-0.13	...	1.71

- Once we do an SVD what is a good choice for the representation of $word_i$?
- Obviously, taking the i-th row of the reconstructed matrix does not make sense because it is still high dimensional

$$\hat{X} =$$

	human	machine	system	for	...	user
human	2.01	2.01	0.23	2.14	...	0.43
machine	2.01	2.01	0.23	2.14	...	0.43
system	0.23	0.23	1.17	0.96	...	1.29
for	2.14	2.14	0.96	1.87	...	-0.13
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	0.43	0.43	1.29	-0.13	...	1.71

$$\hat{X}\hat{X}^T =$$

	human	machine	system	for	...	user
human	25.4	25.4	7.6	21.9	...	6.84
machine	25.4	25.4	7.6	21.9	...	6.84
system	7.6	7.6	24.8	18.03	...	20.6
for	21.9	21.9	0.96	24.6	...	15.32
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	6.84	6.84	20.6	15.32	...	17.11

$$cosine\_sim(human, user) = 0.33$$

- Once we do an SVD what is a good choice for the representation of $word_i$?
- Obviously, taking the i-th row of the reconstructed matrix does not make sense because it is still high dimensional
- But we saw that the reconstructed matrix $\hat{X} = U\Sigma V^T$ discovers latent semantics and its word representations are more meaningful

$$\hat{X} =$$

	human	machine	system	for	...	user
human	2.01	2.01	0.23	2.14	...	0.43
machine	2.01	2.01	0.23	2.14	...	0.43
system	0.23	0.23	1.17	0.96	...	1.29
for	2.14	2.14	0.96	1.87	...	-0.13
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	0.43	0.43	1.29	-0.13	...	1.71

$$\hat{X}\hat{X}^T =$$

	human	machine	system	for	...	user
human	25.4	25.4	7.6	21.9	...	6.84
machine	25.4	25.4	7.6	21.9	...	6.84
system	7.6	7.6	24.8	18.03	...	20.6
for	21.9	21.9	0.96	24.6	...	15.32
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	6.84	6.84	20.6	15.32	...	17.11

$$cosine\_sim(human, user) = 0.33$$

- Once we do an SVD what is a good choice for the representation of $word_i$?
- Obviously, taking the i-th row of the reconstructed matrix does not make sense because it is still high dimensional
- But we saw that the reconstructed matrix $\hat{X} = U\Sigma V^T$ discovers latent semantics and its word representations are more meaningful
- **Wishlist:** We would want representations of words (i, j) to be of smaller dimensions but still have the same similarity (dot product) as the corresponding rows of $\hat{X}$

$$\hat{X} =$$

	human	machine	system	for	...	user
human	2.01	2.01	0.23	2.14	...	0.43
machine	2.01	2.01	0.23	2.14	...	0.43
system	0.23	0.23	1.17	0.96	...	1.29
for	2.14	2.14	0.96	1.87	...	-0.13
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	0.43	0.43	1.29	-0.13	...	1.71

$$\hat{X}\hat{X}^T =$$

	human	machine	system	for	...	user
human	25.4	25.4	7.6	21.9	...	6.84
machine	25.4	25.4	7.6	21.9	...	6.84
system	7.6	7.6	24.8	18.03	...	20.6
for	21.9	21.9	0.96	24.6	...	15.32
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	6.84	6.84	20.6	15.32	...	17.11

$$similarity = 0.33$$

- Notice that the dot product between the rows of the the matrix $W_{word} = U\Sigma$ is the same as the dot product between the rows of $\hat{X}$

$$\hat{X}\hat{X}^T = (U\Sigma V^T)(U\Sigma V^T)^T$$

$\hat{X} =$

	human	machine	system	for	...	user
human	2.01	2.01	0.23	2.14	...	0.43
machine	2.01	2.01	0.23	2.14	...	0.43
system	0.23	0.23	1.17	0.96	...	1.29
for	2.14	2.14	0.96	1.87	...	-0.13
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	0.43	0.43	1.29	-0.13	...	1.71

$\hat{X}\hat{X}^T =$

	human	machine	system	for	...	user
human	25.4	25.4	7.6	21.9	...	6.84
machine	25.4	25.4	7.6	21.9	...	6.84
system	7.6	7.6	24.8	18.03	...	20.6
for	21.9	21.9	0.96	24.6	...	15.32
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	6.84	6.84	20.6	15.32	...	17.11

$$similarity = 0.33$$

- Notice that the dot product between the rows of the the matrix $W_{word} = U\Sigma$ is the same as the dot product between the rows of $\hat{X}$

$$\hat{X}\hat{X}^T = (U\Sigma V^T)(U\Sigma V^T)^T$$
$$= (U\Sigma V^T)(V\Sigma U^T)$$

$$\hat{X} =$$

	human	machine	system	for	...	user
human	2.01	2.01	0.23	2.14	...	0.43
machine	2.01	2.01	0.23	2.14	...	0.43
system	0.23	0.23	1.17	0.96	...	1.29
for	2.14	2.14	0.96	1.87	...	-0.13
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	0.43	0.43	1.29	-0.13	...	1.71

$$\hat{X}\hat{X}^T =$$

	human	machine	system	for	...	user
human	25.4	25.4	7.6	21.9	...	6.84
machine	25.4	25.4	7.6	21.9	...	6.84
system	7.6	7.6	24.8	18.03	...	20.6
for	21.9	21.9	0.96	24.6	...	15.32
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	6.84	6.84	20.6	15.32	...	17.11

$$similarity = 0.33$$

- Notice that the dot product between the rows of the the matrix $W_{word} = U\Sigma$ is the same as the dot product between the rows of $\hat{X}$

$$\begin{aligned} \hat{X}\hat{X}^T &= (U\Sigma V^T)(U\Sigma V^T)^T \\ &= (U\Sigma V^T)(V\Sigma U^T) \\ &= U\Sigma\Sigma^T U^T \quad (\because V^T V = I) \end{aligned}$$

$$\hat{X} =$$

	human	machine	system	for	...	user
human	2.01	2.01	0.23	2.14	...	0.43
machine	2.01	2.01	0.23	2.14	...	0.43
system	0.23	0.23	1.17	0.96	...	1.29
for	2.14	2.14	0.96	1.87	...	-0.13
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	0.43	0.43	1.29	-0.13	...	1.71

$$\hat{X}\hat{X}^T =$$

	human	machine	system	for	...	user
human	25.4	25.4	7.6	21.9	...	6.84
machine	25.4	25.4	7.6	21.9	...	6.84
system	7.6	7.6	24.8	18.03	...	20.6
for	21.9	21.9	0.96	24.6	...	15.32
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	6.84	6.84	20.6	15.32	...	17.11

$$similarity = 0.33$$

- Notice that the dot product between the rows of the the matrix $W_{word} = U\Sigma$ is the same as the dot product between the rows of $\hat{X}$

$$\begin{aligned}
\hat{X}\hat{X}^T &= (U\Sigma V^T)(U\Sigma V^T)^T \\
&= (U\Sigma V^T)(V\Sigma U^T) \\
&= U\Sigma\Sigma^T U^T \quad (\because V^T V = I) \\
&= U\Sigma(U\Sigma)^T = W_{word}W_{word}^T
\end{aligned}$$

$$\hat{X} =$$

	human	machine	system	for	...	user
human	2.01	2.01	0.23	2.14	...	0.43
machine	2.01	2.01	0.23	2.14	...	0.43
system	0.23	0.23	1.17	0.96	...	1.29
for	2.14	2.14	0.96	1.87	...	-0.13
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	0.43	0.43	1.29	-0.13	...	1.71

$$\hat{X}\hat{X}^T =$$

	human	machine	system	for	...	user
human	25.4	25.4	7.6	21.9	...	6.84
machine	25.4	25.4	7.6	21.9	...	6.84
system	7.6	7.6	24.8	18.03	...	20.6
for	21.9	21.9	0.96	24.6	...	15.32
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
user	6.84	6.84	20.6	15.32	...	17.11

$$similarity = 0.33$$

- Notice that the dot product between the rows of the the matrix $W_{word} = U\Sigma$ is the same as the dot product between the rows of $\hat{X}$

$$\begin{aligned}\hat{X}\hat{X}^T &= (U\Sigma V^T)(U\Sigma V^T)^T \\ &= (U\Sigma V^T)(V\Sigma U^T) \\ &= U\Sigma\Sigma^T U^T \quad (\because V^T V = I) \\ &= U\Sigma(U\Sigma)^T = W_{word}W_{word}^T\end{aligned}$$

- Conventionally,

$$W_{word} = U\Sigma \in \mathbb{R}^{m \times k}$$

is taken as the representation of the m words in the vocabulary and

$$W_{context} = V$$

is taken as the representation of the context words