Module 10.8: GloVe representations

- **Count** based methods (SVD) rely on global co-occurrence counts from the corpus for computing word representations

- **Count** based methods (SVD) rely on global co-occurrence counts from the corpus for computing word representations
- Predict based methods **learn** word representations using co-occurrence information

- **Count** based methods (SVD) rely on global co-occurrence counts from the corpus for computing word representations
- Predict based methods **learn** word representations using co-occurrence information
- Why not combine the two (**count** and **learn**) ?

Corpus:

- Human machine interface for computer applications
- User opinion of computer system response time
- User interface management system
- System engineering for improved response time

$$X =$$

	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

$$P(j|i) = \frac{X_{ij}}{\sum X_{ij}} = \frac{X_{ij}}{X_i}$$

$$X_{ij} = X_{ji}$$

- X_{ij} encodes important global information about the co-occurrence between i and j (global: because it is computed for the entire corpus)

Corpus:

- Human machine interface for computer applications
- User opinion of computer system response time
- User interface management system
- System engineering for improved response time

$$X =$$

	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

$$P(j|i) = \frac{X_{ij}}{\sum X_{ij}} = \frac{X_{ij}}{X_i}$$

$$X_{ij} = X_{ji}$$

- X_{ij} encodes important global information about the co-occurrence between i and j (global: because it is computed for the entire corpus)
- Why not learn word vectors which are faithful to this information?

Corpus:

- Human machine interface for computer applications
- User opinion of computer system response time
- User interface management system
- System engineering for improved response time

$$X =$$

	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

$$P(j|i) = \frac{X_{ij}}{\sum X_{ij}} = \frac{X_{ij}}{X_i}$$

$$X_{ij} = X_{ji}$$

- X_{ij} encodes important global information about the co-occurrence between i and j (global: because it is computed for the entire corpus)
- Why not learn word vectors which are faithful to this information?
- For example, enforce

$$\begin{aligned} v_i^T v_j &= \log P(j|i) \\ &= \log X_{ij} - \log(X_i) \end{aligned}$$

Corpus:

- Human machine interface for computer applications
- User opinion of computer system response time
- User interface management system
- System engineering for improved response time

$$X =$$

	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

$$P(j|i) = \frac{X_{ij}}{\sum X_{ij}} = \frac{X_{ij}}{X_i}$$

$$X_{ij} = X_{ji}$$

- X_{ij} encodes important global information about the co-occurrence between i and j (global: because it is computed for the entire corpus)
- Why not learn word vectors which are faithful to this information?
- For example, enforce

$$\begin{aligned} v_i^T v_j &= \log P(j|i) \\ &= \log X_{ij} - \log(X_i) \end{aligned}$$

- Similarly,

$$v_j^T v_i = \log X_{ij} - \log X_j \quad (X_{ij} = X_{ji})$$

Corpus:

- Human machine interface for computer applications
- User opinion of computer system response time
- User interface management system
- System engineering for improved response time

$$X =$$

	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

$$P(j|i) = \frac{X_{ij}}{\sum X_{ij}} = \frac{X_{ij}}{X_i}$$

$$X_{ij} = X_{ji}$$

- X_{ij} encodes important global information about the co-occurrence between i and j (global: because it is computed for the entire corpus)
- Why not learn word vectors which are faithful to this information?
- For example, enforce

$$\begin{aligned} v_i^T v_j &= \log P(j|i) \\ &= \log X_{ij} - \log(X_i) \end{aligned}$$

- Similarly,

$$v_j^T v_i = \log X_{ij} - \log X_j \quad (X_{ij} = X_{ji})$$

- Essentially we are saying that we want word vectors v_i and v_j such that $v_i^T v_j$ is faithful to the globally computed $P(j|i)$

Corpus:

- Human machine interface for computer applications
- User opinion of computer system response time
- User interface management system
- System engineering for improved response time

$$X =$$

	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

$$P(j|i) = \frac{X_{ij}}{\sum X_{ij}} = \frac{X_{ij}}{X_i}$$

$$X_{ij} = X_{ji}$$

- Adding the two equations we get

$$2v_i^T v_j = 2\log X_{ij} - \log X_i - \log X_j$$

$$v_i^T v_j = \log X_{ij} - \frac{1}{2}\log X_i - \frac{1}{2}\log X_j$$

Corpus:

- Human machine interface for computer applications
- User opinion of computer system response time
- User interface management system
- System engineering for improved response time

$$X =$$

	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

$$P(j|i) = \frac{X_{ij}}{\sum X_{ij}} = \frac{X_{ij}}{X_i}$$
$$X_{ij} = X_{ji}$$

- Adding the two equations we get

$$2v_i^T v_j = 2\log X_{ij} - \log X_i - \log X_j$$
$$v_i^T v_j = \log X_{ij} - \frac{1}{2}\log X_i - \frac{1}{2}\log X_j$$

- Note that $\log X_i$ and $\log X_j$ depend only on the words i & j and we can think of them as word specific biases which will be learned

$$v_i^T v_j = \log X_{ij} - b_i - b_j$$
$$v_i^T v_j + b_i + b_j = \log X_{ij}$$

Corpus:

- Human machine interface for computer applications
- User opinion of computer system response time
- User interface management system
- System engineering for improved response time

$$X =$$

	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

$$P(j|i) = \frac{X_{ij}}{\sum X_{ij}} = \frac{X_{ij}}{X_i}$$

$$X_{ij} = X_{ji}$$

- Adding the two equations we get

$$2v_i^T v_j = 2\log X_{ij} - \log X_i - \log X_j$$

$$v_i^T v_j = \log X_{ij} - \frac{1}{2}\log X_i - \frac{1}{2}\log X_j$$

- Note that $\log X_i$ and $\log X_j$ depend only on the words i & j and we can think of them as word specific biases which will be learned

$$v_i^T v_j = \log X_{ij} - b_i - b_j$$

$$v_i^T v_j + b_i + b_j = \log X_{ij}$$

- We can then formulate this as the following optimization problem

$$\min_{v_i, v_j, b_i, b_j} \sum_{i,j} (\underbrace{v_i^T v_j + b_i + b_j}_{\text{predicted value using model parameters}} - \underbrace{\log X_{ij}}_{\text{actual value computed from the given corpus}})^2$$

Corpus:

- Human machine interface for computer applications
- User opinion of computer system response time
- User interface management system
- System engineering for improved response time

$$\min_{v_i, v_j, b_i, b_j} \sum_{i,j} (v_i^T v_j + b_i + b_j - \log X_{ij})^2$$

$$X =$$

	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

$$P(j|i) = \frac{X_{ij}}{\sum X_{ij}} = \frac{X_{ij}}{\sum X_i}$$

$$X_{ij} = X_{ji}$$

Corpus:

- Human machine interface for computer applications
- User opinion of computer system response time
- User interface management system
- System engineering for improved response time

$$X =$$

	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

$$P(j|i) = \frac{X_{ij}}{\sum X_{ij}} = \frac{X_{ij}}{\sum X_i}$$

$$X_{ij} = X_{ji}$$

$$\min_{v_i, v_j, b_i, b_j} \sum_{i,j} (v_i^T v_j + b_i + b_j - \log X_{ij})^2$$

- **Drawback:** weighs all co-occurrences equally

Corpus:

- Human machine interface for computer applications
- User opinion of computer system response time
- User interface management system
- System engineering for improved response time

$$X =$$

	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

$$P(j|i) = \frac{X_{ij}}{\sum X_{ij}} = \frac{X_{ij}}{\sum X_i}$$
$$X_{ij} = X_{ji}$$

$$\min_{v_i, v_j, b_i, b_j} \sum_{i,j} (v_i^T v_j + b_i + b_j - \log X_{ij})^2$$

- **Drawback:** weighs all co-occurrences equally
- **Solution:** add a weighting function

$$\min_{v_i, v_j, b_i, b_j} \sum_{i,j} f(X_{ij})(v_i^T v_j + b_i + b_j - \log X_{ij})^2$$

Corpus:

- Human machine interface for computer applications
- User opinion of computer system response time
- User interface management system
- System engineering for improved response time

$$X =$$

	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

$$P(j|i) = \frac{X_{ij}}{\sum X_{ij}} = \frac{X_{ij}}{\sum X_i}$$
$$X_{ij} = X_{ji}$$

$$\min_{v_i, v_j, b_i, b_j} \sum_{i,j} (v_i^T v_j + b_i + b_j - \log X_{ij})^2$$

- **Drawback:** weighs all co-occurrences equally
- **Solution:** add a weighting function

$$\min_{v_i, v_j, b_i, b_j} \sum_{i,j} f(X_{ij})(v_i^T v_j + b_i + b_j - \log X_{ij})^2$$

- **Wishlist:** $f(X_{ij})$ should be such that neither rare nor frequent words are over-weighted.

$$f(x) = \left\{ \begin{array}{ll} (\frac{x}{x_{max}})^\alpha, & \text{if } x < x_{max} \\ 1, & \text{otherwise} \end{array} \right\}$$

where α can be tuned for a given dataset