

# LSA (잠재의미분석)

# Here we have menus from restaurants

- pizza
- pizza hamburger cookie
- hamburger
- ramen
- sushi
- ramen sushi



American food

# If we use bag of words, is pizza more similar to hamburger than ramen?

- pizza
  - pizza hamburger cookie
  - hamburger
  - ramen
  - sushi
  - ramen sushi
- 
- Japanese food

# If we use bag of words, pizza and hamburger doesn't have similarity

	pizza	hamburger	cookie	ramen	sushi
pizza	1	0	0	0	0
pizza hamburger cookie	1	1	1	0	0
hamburger	0	1	0	0	0
ramen	0	0	0	1	0
sushi	0	0	0	0	1
ramen sushi	0	0	0	1	1

**Similarity(pizza, hamburger) = 0**

**Similarity(pizza, ramen) = 0**

**TF-IDF** will have same zero similarity for pizza and hamburger  
since they don't have same words.

**Why?**

TF-IDF or Bag of Words  
similarity is based on **word**

TF-IDF or Bag of Words similarity  
is not based on **topic**



LSA similarity is based on **topic**

# Here we have menus from restaurants

- pizza
- pizza hamburger cookie
- hamburger
- ramen
- sushi
- ramen sushi

# Here is word-document matrix

	pizza	pizza hamburger cookie	hamburger	ramen	sushi	ramen sushi
	d1	d2	d3	d4	d5	d6
pizza	1	1	0	0	0	0
ham burger	0	1	1	0	0	0
cookie	0	1	0	0	0	0
ramen	0	0	0	1	0	1
sushi	0	0	0	0	1	1

= **A**

# After SVD matrix decomposition

$$\mathbf{A} \doteq \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T$$

	t1	t2	t3	t4	t5
w1	0.6	0	0	0.7	-0.3
w2	0.6	0	0	-0.7	-0.3
w3	0.5	0	0	0	0.9
w4	0	0.7	-0.7	0	0
w5	0	0.7	0.7	0	0

	t1	t2	t3	t4	t5	t6
t1	1.9	0	0	0	0	0
t2	0	1.7	0	0	0	0
t3	0	0	1	0	0	0
t4	0	0	0	1	0	0
t5	0	0	0	0	0.5	0

	d1	d2	d3	d4	d5	d6
t1	0.3	0.9	0.3	0	0	0
t2	0	0	0	0.4	0.4	0.8
t3	0	0	0	-0.7	0.7	0
t4	0.7	0	-0.7	0	0	0
t5	-0.6	0.5	-0.6	0	0	0
t6	0	0	0	-0.6	-0.6	0.6

Word matrix  
for topic

# After SVD matrix decomposition

$$\mathbf{A} \hat{=} \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T$$

	t1	t2	t3	t4	t5
w1	0.6	0	0	0.7	-0.3
w2	0.6	0	0	-0.7	-0.3
w3	0.5	0	0	0	0.9
w4	0	0.7	-0.7	0	0
w5	0	0.7	0.7	0	0

	t1	t2	t3	t4	t5	t6
t1	1.9	0	0	0	0	0
t2	0	1.7	0	0	0	0
t3	0	0	1	0	0	0
t4	0	0	0	1	0	0
t5	0	0	0	0	0.5	0

	d1	d2	d3	d4	d5	d6
t1	0.3	0.9	0.3	0	0	0
t2	0	0	0	0.4	0.4	0.8
t3	0	0	0	-0.7	0.7	0
t4	0.7	0	-0.7	0	0	0
t5	-0.6	0.5	-0.6	0	0	0
t6	0	0	0	-0.6	-0.6	0.6

Word matrix  
for topic

Topic Strength

Document matrix  
for topic

# Document Vector

 $\Sigma$ 

	t1	t2	t3	t4	t5	t6
t1	1.9	0	0	0	0	0
t2	0	1.7	0	0	0	0
t3	0	0	1	0	0	0
t4	0	0	0	1	0	0
t5	0	0	0	0	0.5	0

 $\times$  $V^T$ 

	d1	d2	d3	d4	d5	d6
t1	0.3	0.9	0.3	0	0	0
t2	0	0	0	0.4	0.4	0.8
t3	0	0	0	-0.7	0.7	0
t4	0.7	0	-0.7	0	0	0
t5	-0.6	0.5	-0.6	0	0	0
t6	0	0	0	-0.6	-0.6	0.6

# Choose optimal dimension size

 $\Sigma$ 

	t1	t2	t3	t4	t5	t6
t1	1.9	0	0	0	0	0
t2	0	1.7	0	0	0	0
t3	0	0	1	0	0	0
t4	0	0	0	1	0	0
t5	0	0	0	0	0.5	0
t6	0	0	0	0	0	0

 $\times$  $V^T$ 

	d1	d2	d3	d4	d5	d6
t1	0.3	0.9	0.3	0	0	0
t2	0	0	0	0.4	0.4	0.8
t3	0	0	0	-0.7	0.7	0
t4	0.7	0	-0.7	0	0	0
t5	-0.6	0.5	-0.6	0	0	0
t6	0	0	0	-0.6	-0.6	0.6

Descending order  
With importance

# Dimensionality reduction

 $\Sigma$ 

	t1	t2	t3	t4	t5	t6
t1	1.9	0	0	0	0	0
t2	0	1.7	0	0	0	0
t3	0	0	1	0	0	0
t4	0	0	0	1	0	0
t5	0	0	0	0	0.5	0
t6	0	0	0	0	0	0

 $\times$  $V^T$ 

	d1	d2	d3	d4	d5	d6
t1	0.3	0.9	0.3	0	0	0
t2	0	0	0	0.4	0.4	0.8
t3	0	0	0	-0.7	0.7	0
t4	0.7	0	-0.7	0	0	0
t5	-0.6	0.5	-0.6	0	0	0
t6	0	0	0	-0.6	-0.6	0.6

Ignore less than 1.7



# LSA Document vectors in 2 dimension

$$\begin{array}{|c|c|c|c|c|c|c|} \hline & d1 & d2 & d3 & d4 & d5 & d6 \\ \hline t1 & 0.57 & 1.71 & 0.57 & 0 & 0 & 0 \\ \hline t2 & 0 & 0 & 0 & 0.68 & 0.68 & 1.36 \\ \hline \end{array} = \begin{array}{|c|c|c|} \hline & t1 & t2 \\ \hline t1 & 1.9 & 0 \\ \hline t2 & 0 & 1.7 \\ \hline \end{array} \times \begin{array}{|c|c|c|c|c|c|c|} \hline & d1 & d2 & d3 & d4 & d5 & d6 \\ \hline t1 & 0.3 & 0.9 & 0.3 & 0 & 0 & 0 \\ \hline t2 & 0 & 0 & 0 & 0.4 & 0.4 & 0.8 \\ \hline \end{array}$$

The diagram illustrates the LSA Document vectors in 2 dimension. It shows the relationship between the original document-term matrix, the singular value decomposition (SVD) components, and the resulting 2D document vectors.

The original document-term matrix (left) is a 2x6 matrix with rows t1 and t2, and columns d1 through d6. The values are:

	d1	d2	d3	d4	d5	d6
t1	0.57	1.71	0.57	0	0	0
t2	0	0	0	0.68	0.68	1.36

The SVD components are shown in the middle. The  $\Sigma$  matrix (middle) is a 2x2 matrix with rows t1 and t2, and columns t1 and t2. The values are:

	t1	t2
t1	1.9	0
t2	0	1.7

The  $V^T$  matrix (right) is a 2x6 matrix with rows t1 and t2, and columns d1 through d6. The values are:

	d1	d2	d3	d4	d5	d6
t1	0.3	0.9	0.3	0	0	0
t2	0	0	0	0.4	0.4	0.8

The equation shows that the original document-term matrix is equal to the product of the  $\Sigma$  matrix and the  $V^T$  matrix.

# LSA Document vectors in 2 dimension

$$\begin{array}{|c|c|c|c|c|c|c|} \hline & d1 & d2 & d3 & d4 & d5 & d6 \\ \hline t1 & 0.57 & 1.71 & 0.57 & 0 & 0 & 0 \\ \hline t2 & 0 & 0 & 0 & 0.68 & 0.68 & 1.36 \\ \hline \end{array} = \begin{array}{|c|c|c|} \hline & t1 & t2 \\ \hline t1 & 1.9 & 0 \\ \hline t2 & 0 & 1.7 \\ \hline \end{array} \times \begin{array}{|c|c|c|c|c|c|c|} \hline & d1 & d2 & d3 & d4 & d5 & d6 \\ \hline t1 & 0.3 & 0.9 & 0.3 & 0 & 0 & 0 \\ \hline t2 & 0 & 0 & 0 & 0.4 & 0.4 & 0.8 \\ \hline \end{array}$$

The diagram illustrates the LSA Document vectors in 2 dimensions. It shows the decomposition of a document-term matrix into three components: a term-term matrix ( $\Sigma$ ), a term-term matrix ( $V^T$ ), and a term-term matrix ( $V$ ).

The first matrix (Document-Term matrix) has rows t1 and t2, and columns d1, d2, d3, d4, d5, and d6. The values are:

	d1	d2	d3	d4	d5	d6
t1	0.57	1.71	0.57	0	0	0
t2	0	0	0	0.68	0.68	1.36

The second matrix ( $\Sigma$ ) has rows t1 and t2, and columns t1 and t2. The values are:

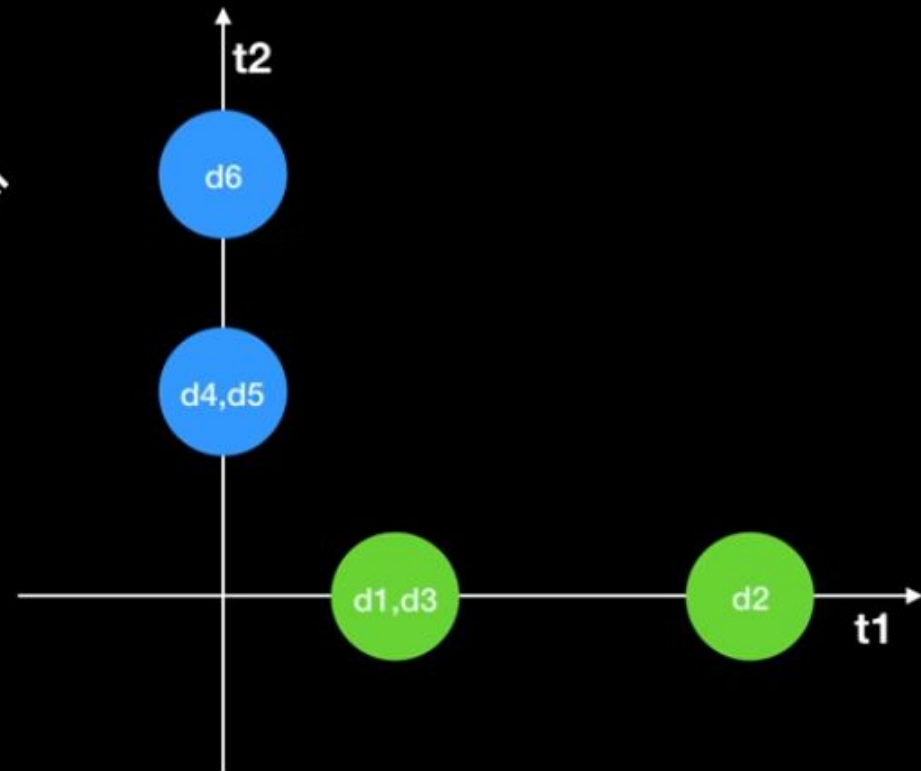
	t1	t2
t1	1.9	0
t2	0	1.7

The third matrix ( $V^T$ ) has rows t1 and t2, and columns d1, d2, d3, d4, d5, and d6. The values are:

	d1	d2	d3	d4	d5	d6
t1	0.3	0.9	0.3	0	0	0
t2	0	0	0	0.4	0.4	0.8

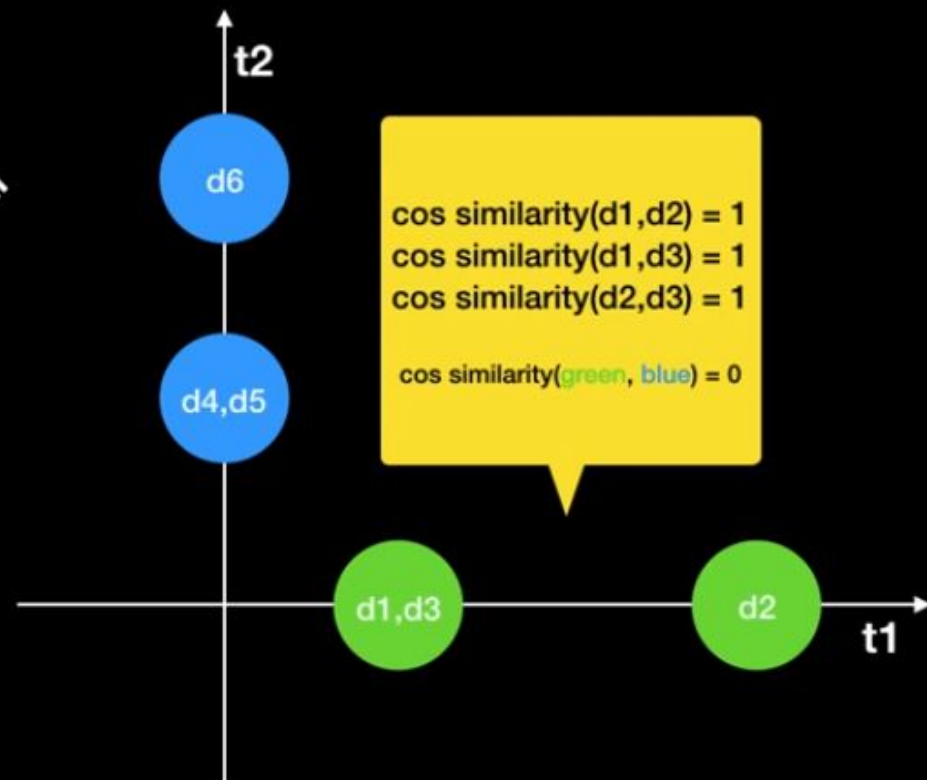
# Topic Similarity

		pizza	pizza hamburger	cookie		
	d1	d2	d3	d4	d5	d6
t1	0.57	1.71	0.57	0	0	0
t2	0	0	0	0.68	0.68	1.36



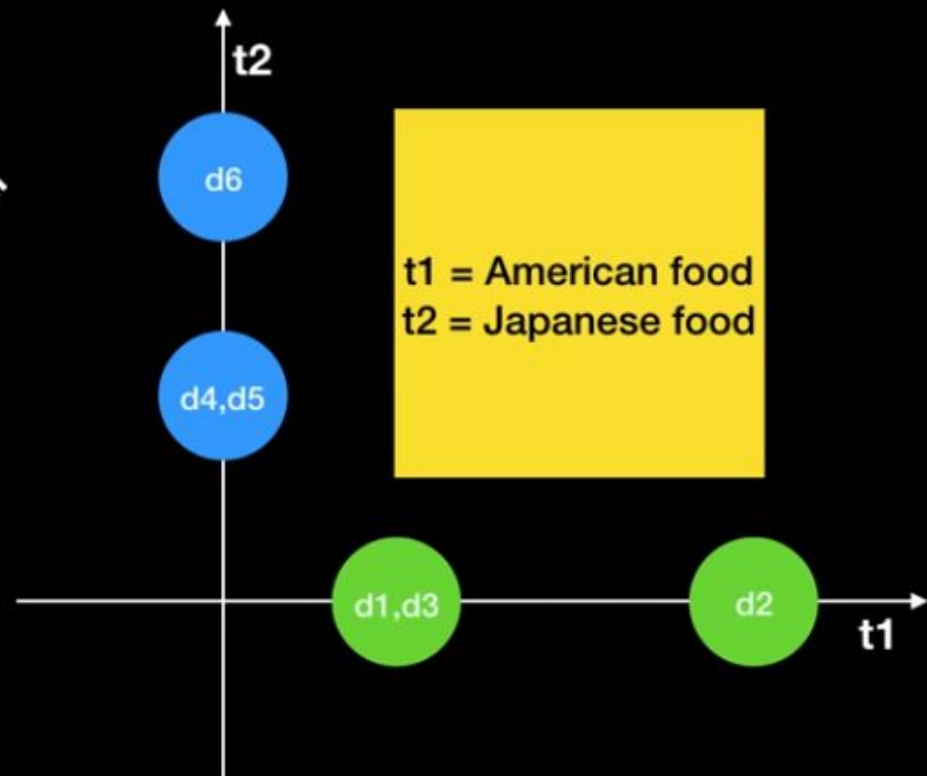
# Topic Similarity

		pizza	pizza hamburger cookie	hamburger	ramen	sushi	ramen sushi
		d1	d2	d3	d4	d5	d6
t1		0.57	1.71	0.57	0	0	0
t2		0	0	0	0.68	0.68	1.36



# Topic Similarity

		pizza	pizza hamburger cookie	hamburger	ramen	sushi	ramen sushi
	d1	d2	d3	d4	d5	d6	
t1	0.57	1.71	0.57	0	0	0	
t2	0	0	0	0.68	0.68	1.36	



출 처: <https://www.youtube.com/watch?v=GVPTGq53H5I>