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

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

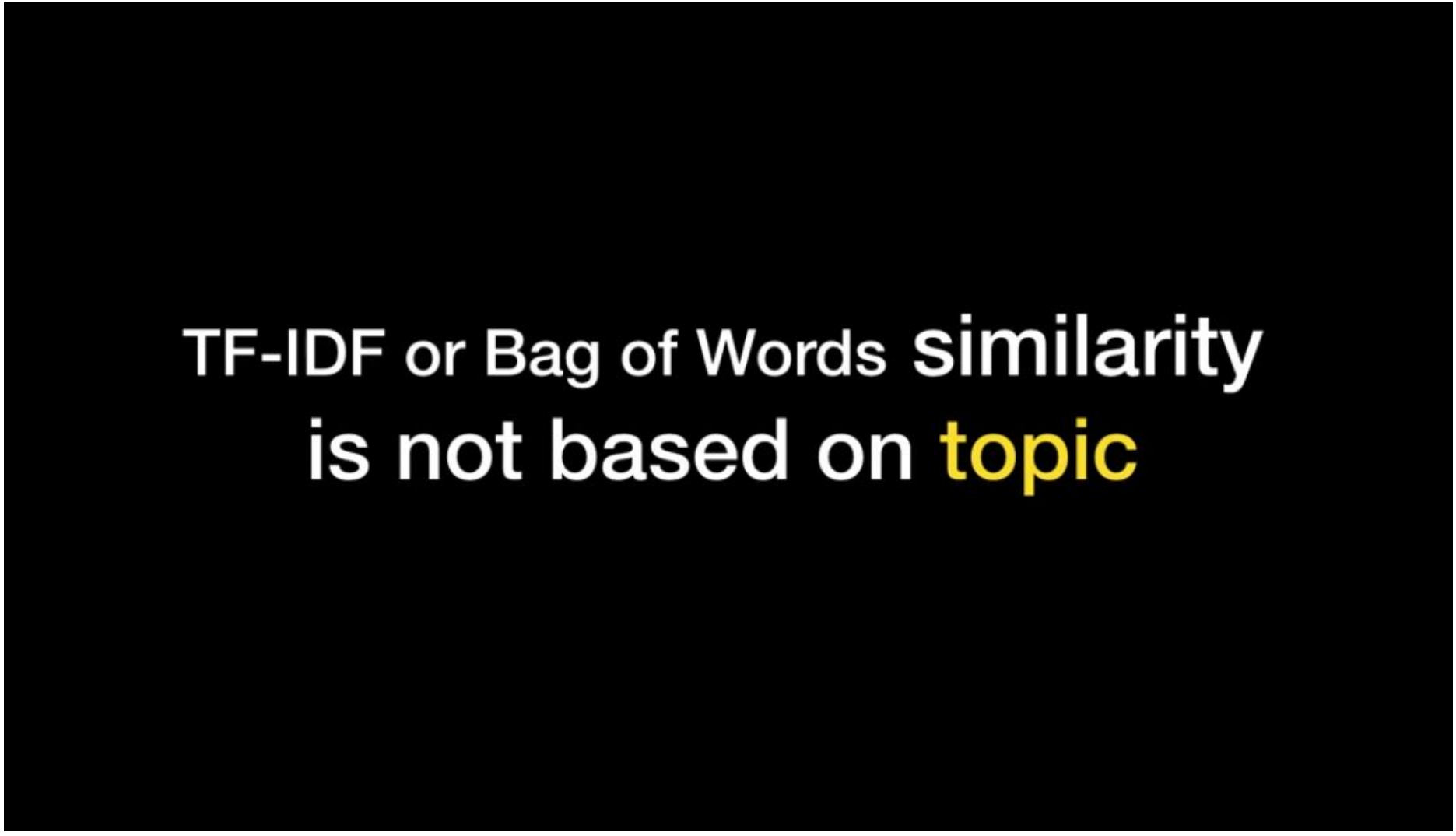
Similarity(pizza, hamburger) = 0

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

	d1 (pizza)	d2 (pizza hamburger cookie)	d3 (hamburger)	d4 (ramen)	d5 (sushi)	d6 (ramen sushi)
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

$A \doteq U \times \Sigma \times V^T$

U

	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

Word matrix for topic

Σ

	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

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

After SVD matrix decomposition

$A \doteq U \times \Sigma \times V^T$

U

	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

Word matrix for topic

Σ

	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

Topic Strength

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

Document matrix for topic

Document Vector

Σ

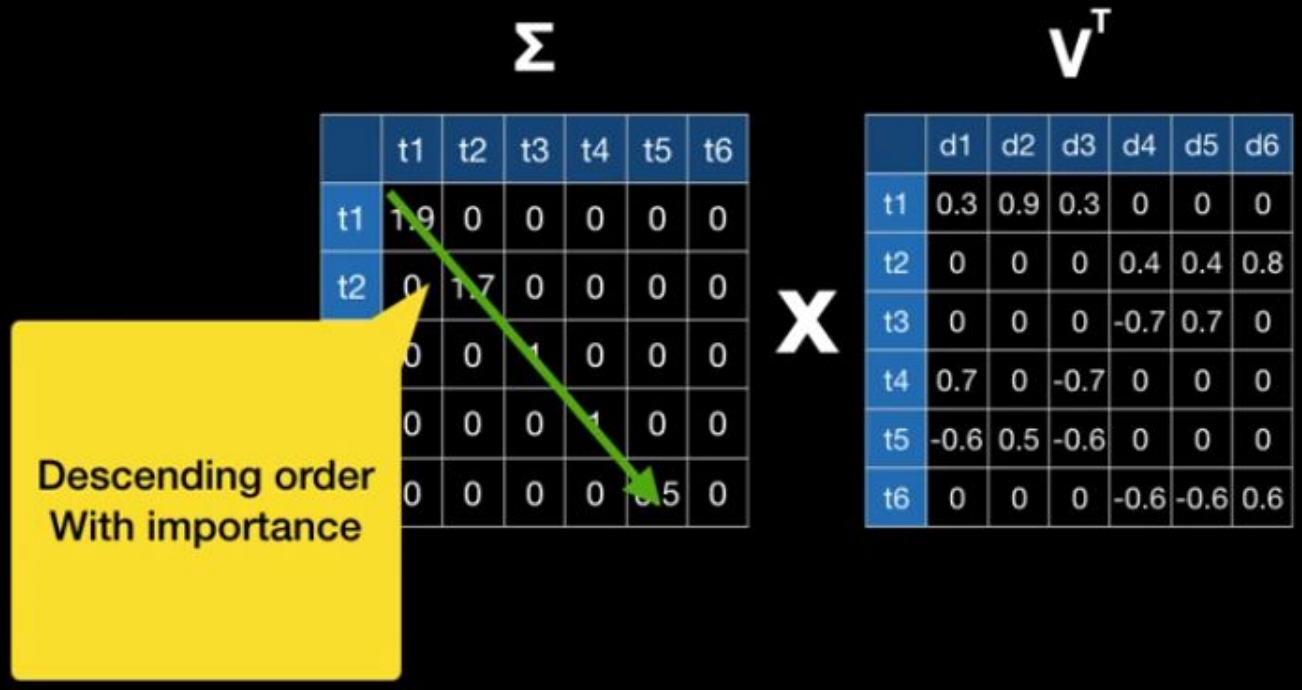
	t1	t2	t3	t4	t5	t6
t1	1.9	0	0	0	0	0
t2	0	1.7	0	0	0	0
t3	0	0	1	0	0	0
t4	0	0	0	1	0	0
t5	0	0	0	0	0.5	0

X

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

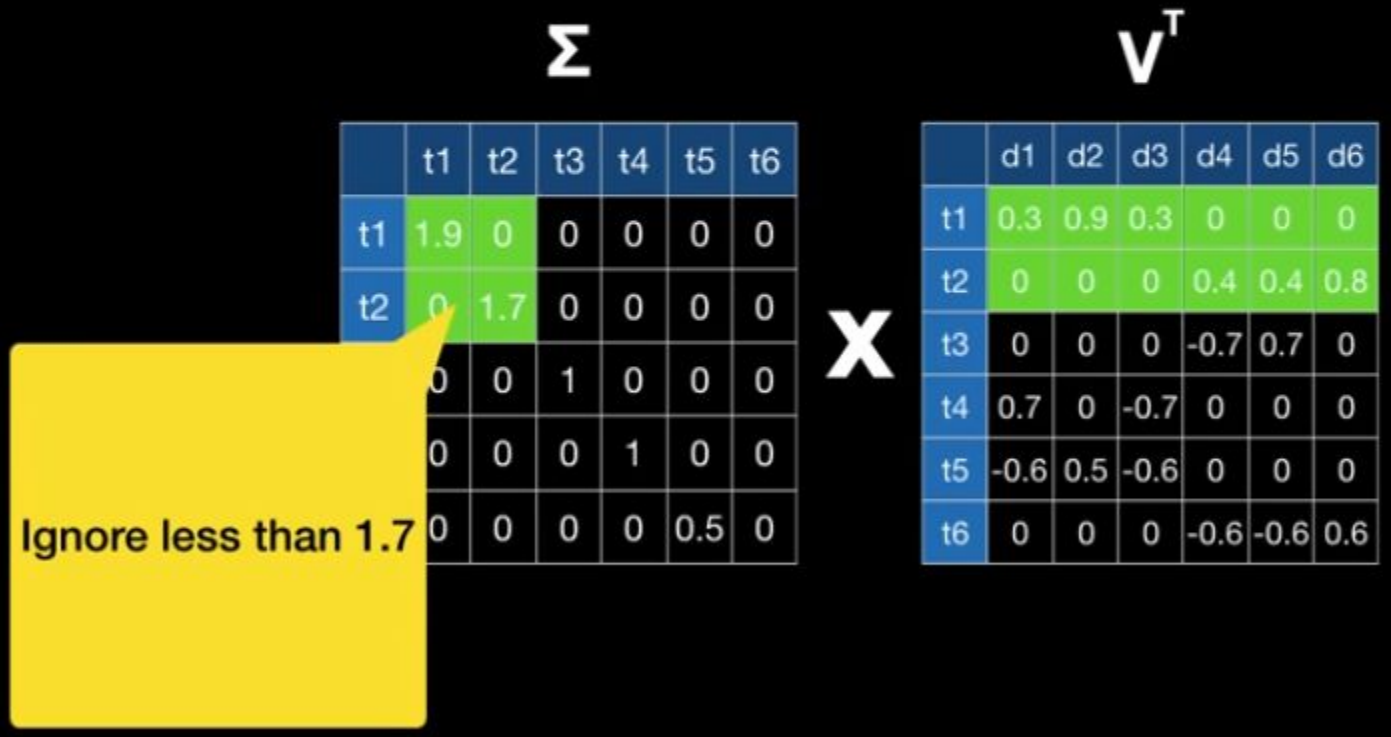
Choose optimal dimension size



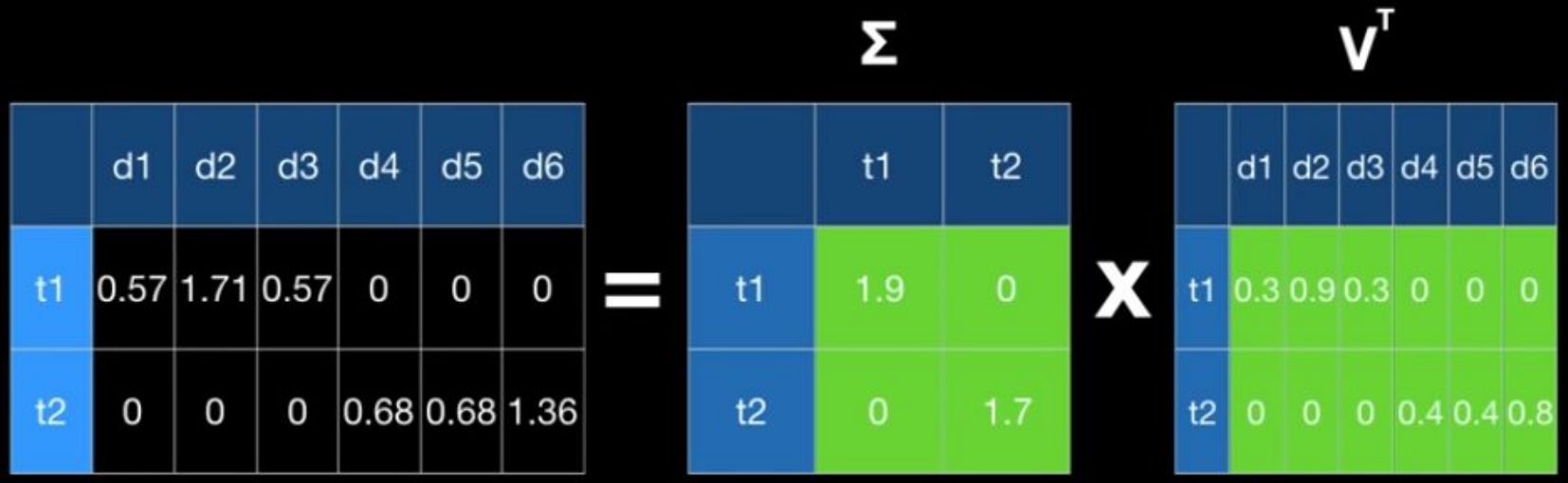
	t1	t2	t3	t4	t5	t6
t1	1.9	0	0	0	0	0
t2	0	1.7	0	0	0	0
	0	0	1	0	0	0
	0	0	0	1	0	0
	0	0	0	0	.5	0

	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

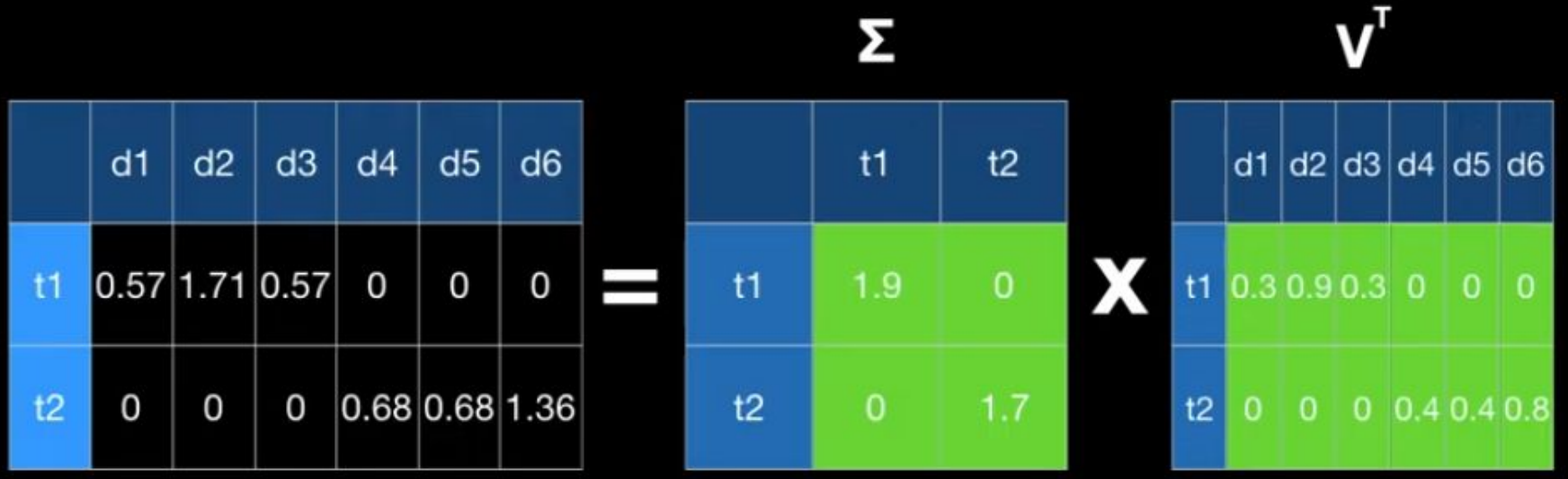
Dimensionality reduction



LSA Document vectors in 2 dimension

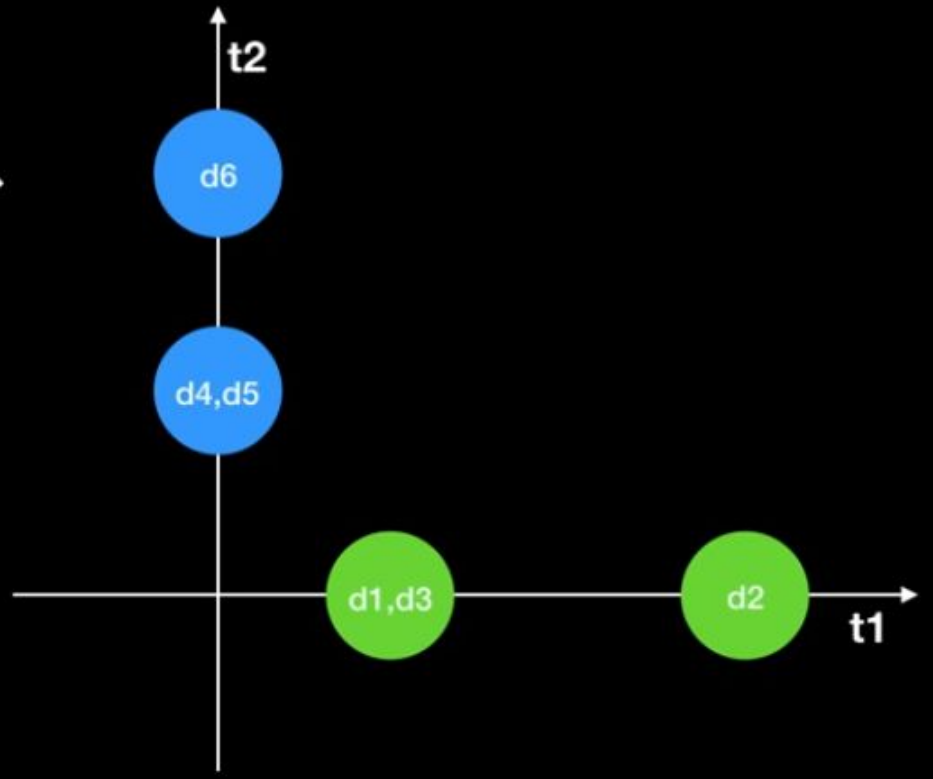


LSA Document vectors in 2 dimension



Topic Similarity

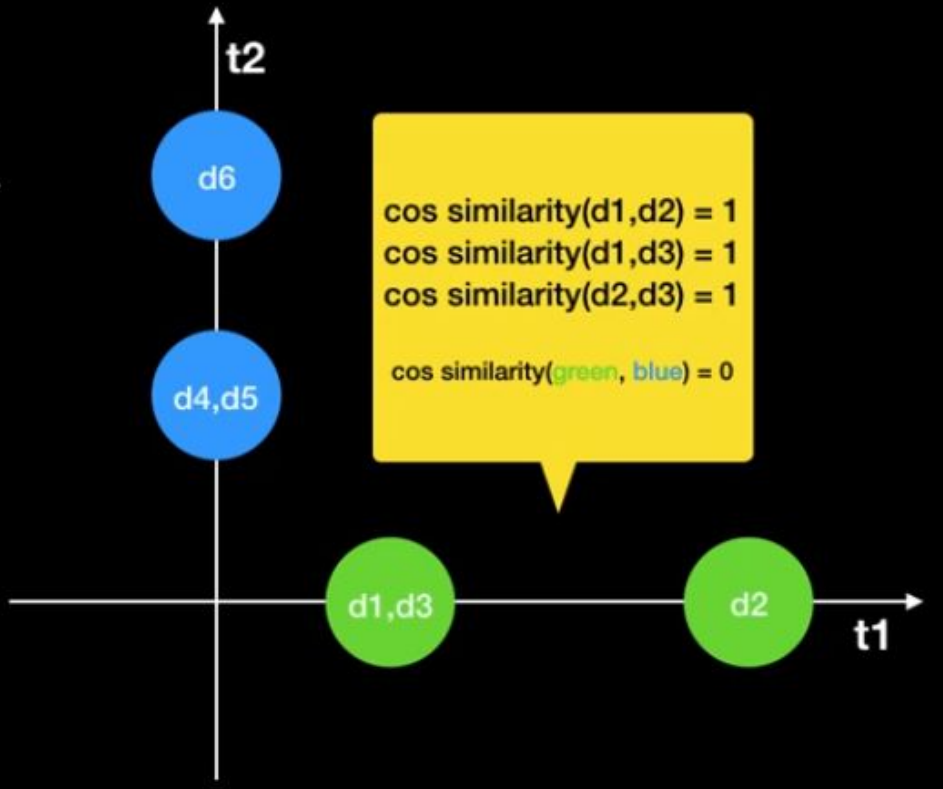
	pizza	pizza hamburger cookie	hamburger	ramen	sushi	ramen sushi
	d1	d2	d3	d4	d5	d6
t1	0.57	1.71	0.57	0	0	0
t2	0	0	0	0.68	0.68	1.36



Topic Similarity

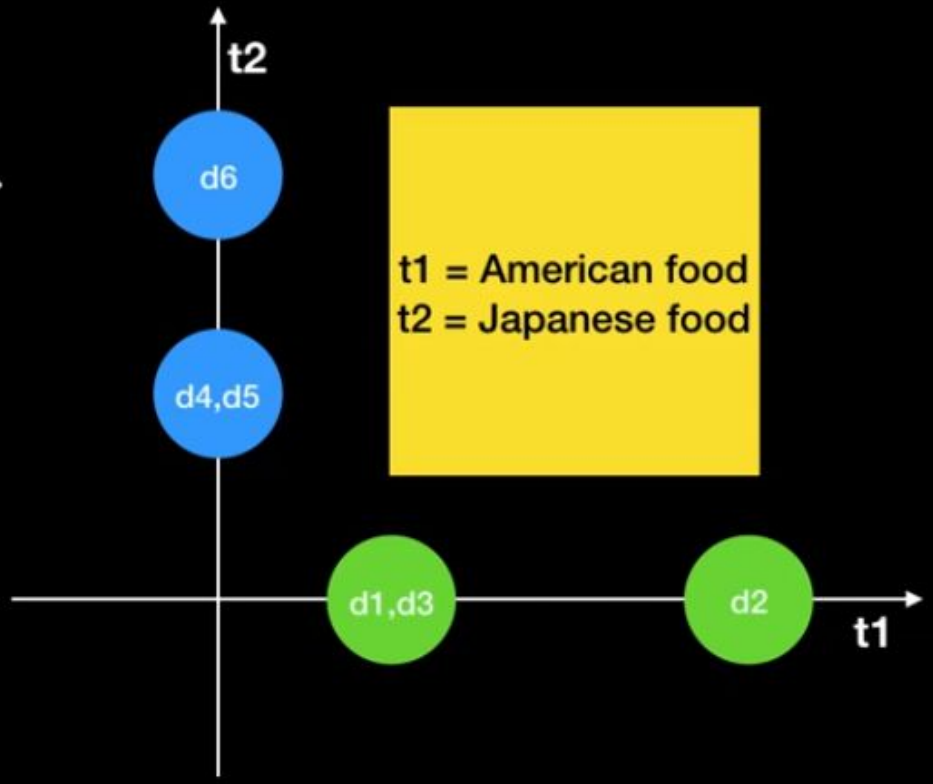
	d1	d2	d3	d4	d5	d6
t1	0.57	1.71	0.57	0	0	0
t2	0	0	0	0.68	0.68	1.36

pizza
pizza hamburger cookie
hamburger
ramen
sushi
ramen sushi



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



출처: https://www.youtube.com/watch?v=GVPTGq53H5I