Lecture 5
텍스트마이닝

Excerpts the examples from Ilya Katsov "Introduction to Algorithmic Marketing"

텍스트마이닝

웹크롤링

텍스트 전처리

토픽분석

Tokenization

Normalization

Stemming

TF-IDF

LSA

LDA

텍스트 기본 전처리

Maison Kitsuné Men’s Slim Jeans. These premium jeans come in a slim fit for a fashionable look.

토큰화

[Maison] [Kitsune] [Men’s] [Slim] [Jeans] [These] [premium] [jeans] [come] [in] [a] [slim] [fit] [for] [a] [fashionable] [look]

정규화

[maison] [kitsune] [men’s] [slim] [jeans] [these] [premium] [jeans] [come] [in] [a] [slim] [fit] [for] [a] [fashionable] [look]

불용어 제거

[maison] [kitsune] [men’s] [slim] [jeans] [premium] [jeans] [come] [slim] [fit] [fashionable] [look]

Stemming

[maison] [kitsun] [men] [slim] [jean] [premium] [jean] [com] [slim] [fit] [fashion] [look]

Term Document Matrix

말뭉치 (corpus) → 용어 추출 → Term Document Matrix

용어 추출 → Document Term Matrix

- 문서1: 오늘은 바나나를 먹는다
- 문서2: 내일은 콜라를 먹는다

- 문서1: 오늘 , 바나나, 먹다
- 문서2: 내일, 콜라, 먹다

	문서1	문서2
오늘	1	0
내일	0	1
바나나	1	0
콜라	0	1
먹다	1	1

	오늘	내일	바나나	콜라	먹다
문서1	1	0	1	0	1
문서2	0	1	0	1	1

문장 유사도 계산

❖ 두 벡터의 곱

$$d \cdot q = \sum_{i=1}^{n} q_i d_i$$

❖ 벡터의 유클리드 기하학적 수식

$$\| d \| = \sqrt{{d_1}^2 + \cdots + {d_n}^2}$$

❖ 코사인 유사도

$$\cos(q, d) = \frac{d \cdot q}{\| d \| \cdot \| q \|}$$

문장 유사도 계산

❖ 두 문장

```
Product 1: dark blue jeans blue denim fabric
Product 2: skinny jeans in bright blue
```

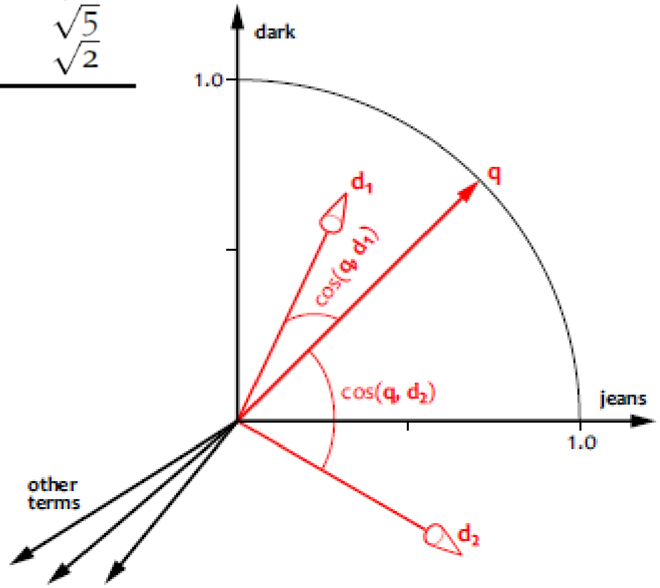
❖ 바이너리 벡터로 표현된 두 문장과 한 검색 질의의 예

	dark	blue	jeans	denim	fabric	skinny	in	bright	$\lVert\cdot\rVert$
$\mathbf{d}_1$	1	1	1	1	1	0	0	0	$\sqrt{5}$
$\mathbf{d}_2$	0	1	1	0	0	1	1	1	$\sqrt{5}$
$\mathbf{q}$	1	0	1	0	0	0	0	0	$\sqrt{2}$

❖ 각 문장과 검색 질의 사이의 유사도

$$\cos(\mathbf{q}, \mathbf{d}_1) = \frac{1 + 1}{\sqrt{2}\sqrt{5}} = 0.632$$

$$\cos(\mathbf{q}, \mathbf{d}_2) = \frac{1}{\sqrt{2}\sqrt{5}} = 0.316$$



TF IDF Scoring Model

- Bag-of-words model
- TF-IDF

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}} \qquad idf(w) = log(\frac{N}{df_t}) \qquad w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

  - *N* : 문서의 전체 수
  - *t* : 용어
  - *d* : 문서
  - *w* : 가중치
  - $tf_{i,j}$: 용어 *i* 와 용어 *j* 의 단어가 등장하는 횟수
  - $n_{i,j}$: 용어 *i* 와 용어 *j* 의 단어가 현재 문서에서 등장하는 횟수
  - Σ_k : 현재 문서에서 갖고 있는 모든 용어 빈도 수
  - df_i : 용어 *i* 를 포함하는 문서 수 / df_t : 용어 *t* 를 포함하는 문서 수
  - $w_{i,j}$: 용어 *i* 와 용어 *j* 의 단어가 등장하는 TF-IDF 지수(가중치)

TF IDF Scoring Model

- **Term frequency**
  - 문서 내 특정 단어의 출현빈도

Variants of term frequency (tf) weight

weighting scheme	tf weight
binary	$0, 1$
raw count	$f_{t,d}$
term frequency	$f_{t,d} \Big/ \sum_{t' \in d} f_{t',d}$
log normalization	$\log(1 + f_{t,d})$
double normalization 0.5	$0.5 + 0.5 \cdot \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$
double normalization K	$K + (1 - K)\frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$

https://en.wikipedia.org/wiki/Tf%E2%80%93idf

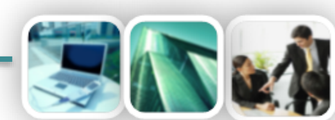
TF IDF Scoring Model

- **Inverse document frequency**
  - **Document frequency: 특정 단어가 출현한 문서의 수**

Variants of inverse document frequency (idf) weight

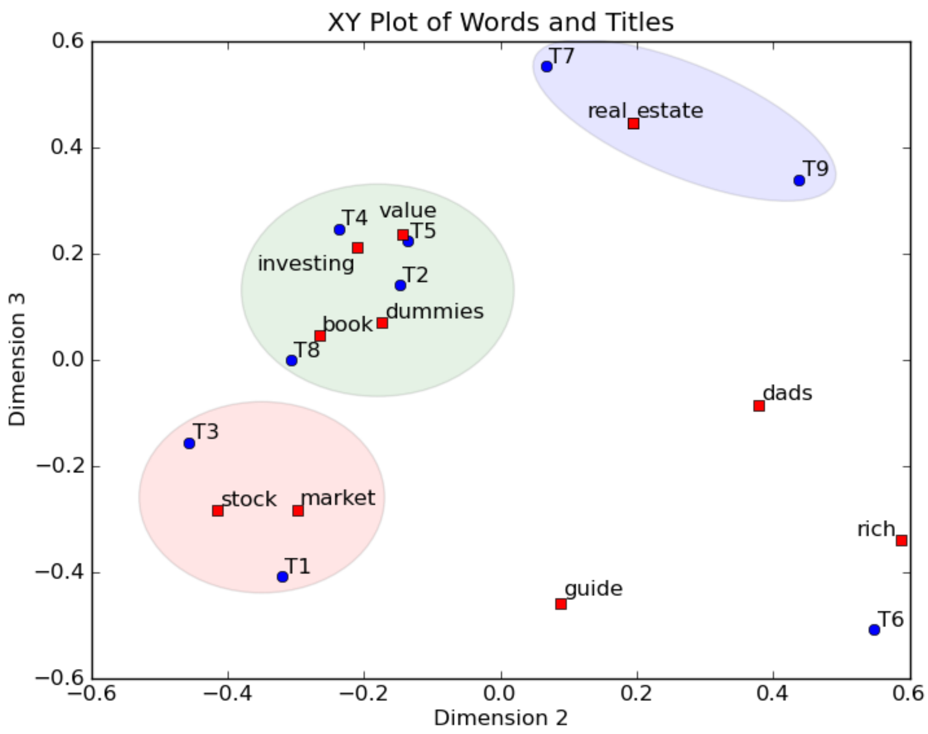
weighting scheme	idf weight ($n_t = \|\{d \in D : t \in d\}\|$)
unary	1
inverse document frequency	$\log \frac{N}{n_t} = -\log \frac{n_t}{N}$
inverse document frequency smooth	$\log \left(\frac{N}{1 + n_t} \right) + 1$
inverse document frequency max	$\log \left(\frac{\max_{\{t' \in d\}} n_{t'}}{1 + n_t} \right)$
probabilistic inverse document frequency	$\log \frac{N - n_t}{n_t}$

https://en.wikipedia.org/wiki/Tf%E2%80%93idf



잠재의미분석

- LSA (Latent semantic analysis)
- LSI (Latent semantic indexing)
- SVD (Singular value decomposition)



잠재의미분석

The Neatest Little Guide to Stock Market Investing

Investing For Dummies, 4th Edition

The Little Book of Common Sense Investing: The Only Way to Guarantee Your Fair Share of Stock Market Returns

The Little Book of Value Investing

Value Investing: From Graham to Buffett and Beyond

Rich Dad's Guide to Investing: What the Rich Invest in, That the Poor and the Middle Class Do Not!

Investing in Real Estate, 5th Edition

Stock Investing For Dummies

Rich Dad's Advisors: The ABC's of Real Estate Investing: The Secrets of Finding Hidden Profits Most Investors Miss

https://technowiki.wordpress.com/2011/08/27/latent-semantic-analysis-lsa-tutorial/

잠재의미분석

Index Words	Titles								
	T1	T2	T3	T4	T5	T6	T7	T8	T9
book			1	1					
dads						1			1
dummies		1						1	
estate							1		1
guide	1					1			
investing	1	1	1	1	1	1	1	1	1
market	1		1						
real							1		1
rich						2			1
stock	1		1					1	
value				1	1				

```
[[ 0.  0.  1.  1.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  1.  0.  0.  1.]
 [ 0.  1.  0.  0.  0.  0.  0.  1.  0.]
 [ 0.  0.  0.  0.  0.  0.  1.  0.  1.]
 [ 1.  0.  0.  0.  0.  1.  0.  0.  0.]
 [ 1.  1.  1.  1.  1.  1.  1.  1.  1.]
 [ 1.  0.  1.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  1.  0.  1.]
 [ 0.  0.  0.  0.  0.  2.  0.  0.  1.]
 [ 1.  0.  1.  0.  0.  0.  0.  1.  0.]
 [ 0.  0.  0.  1.  1.  0.  0.  0.  0.]]
```

book	0.15	-0.27	0.04
dads	0.24	0.38	-0.09
dummies	0.13	-0.17	0.07
estate	0.18	0.19	0.45
guide	0.22	0.09	-0.46
investing	0.74	-0.21	0.21
market	0.18	-0.3	-0.28
real	0.18	0.19	0.45
rich	0.36	0.59	-0.34
stock	0.25	-0.42	-0.28
value	0.12	-0.14	0.23

*

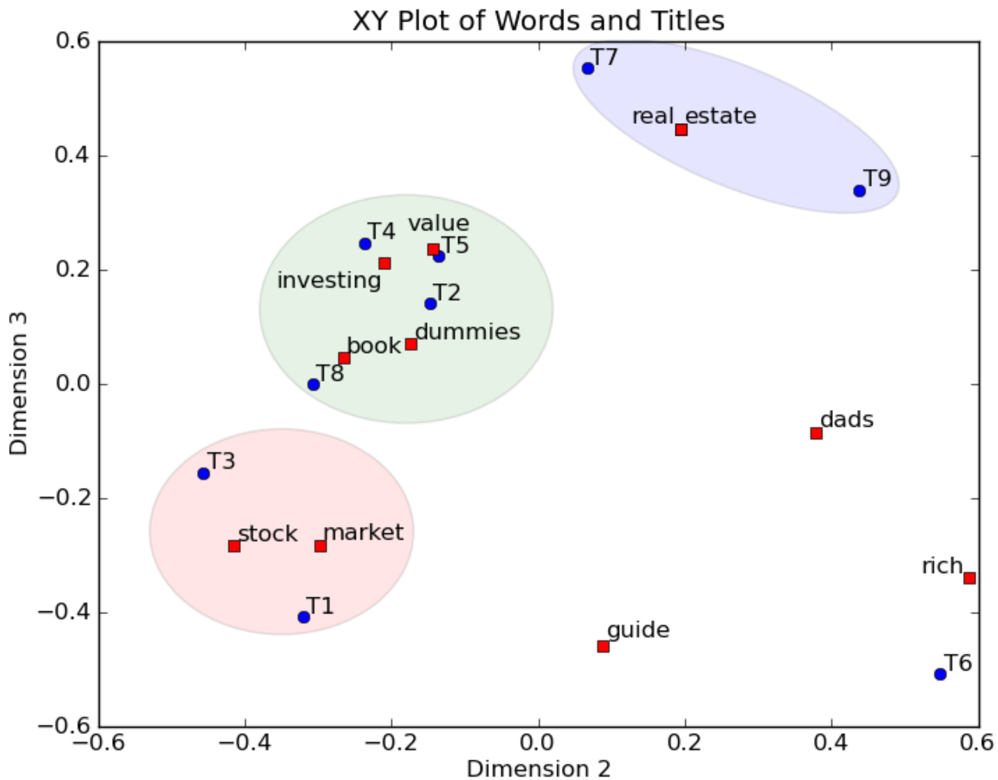
3.91	0	0
0	2.61	0
0	0	2

*

T1	T2	T3	T4	T5	T6	T7	T8	T9
0.35	0.22	0.34	0.26	0.22	0.49	0.28	0.29	0.44
-0.32	-0.15	-0.46	-0.24	-0.14	0.55	0.07	-0.31	0.44
-0.41	0.14	-0.16	0.25	0.22	-0.51	0.55	0	0.34

https://technowiki.wordpress.com/2011/08/27/latent-semantic-analysis-lsa-tutorial/

잠재의미분석



https://technowiki.wordpress.com/2011/08/27/latent-semantic-analysis-lsa-tutorial/

잠재의미분석

❖ 전치행렬 (transposed matrix)

$$M = \begin{bmatrix} 1\,2 \\ 3\,4 \\ 5\,6 \end{bmatrix} \quad M^{\mathrm{T}} = \begin{bmatrix} 1\,3\,5 \\ 2\,4\,6 \end{bmatrix}$$

❖ 단위행렬 (identity matrix)

$$I = \begin{bmatrix} 1\,0\,0 \\ 0\,1\,0 \\ 0\,0\,1 \end{bmatrix}$$

❖ 역행렬 (inverse matrix)

$$A \times A^{-1} = I$$

$$\begin{bmatrix} 1\,2\,3 \\ 4\,5\,6 \\ 7\,8\,9 \end{bmatrix} \times \begin{bmatrix} & & \\ & ? & \\ & & \end{bmatrix} = \begin{bmatrix} 1\,0\,0 \\ 0\,1\,0 \\ 0\,0\,1 \end{bmatrix}$$

출처: https://wikidocs.net/24949

잠재의미분석

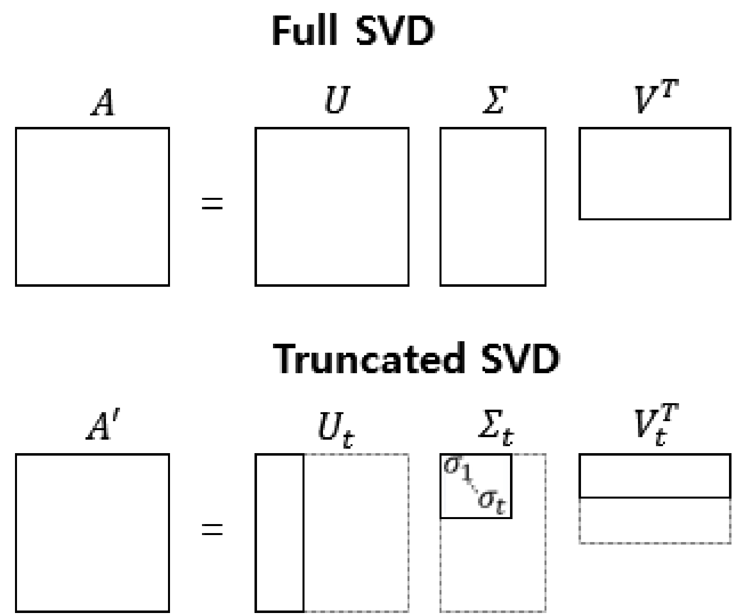
- 직교행렬 (orthogonal matrix)

$$A^{-1} = A^{T}$$

- 대각행렬 (diagonal matrix)

$$\Sigma = \begin{bmatrix} a & 0 & 0 \\ 0 & a & 0 \\ 0 & 0 & a \end{bmatrix}$$

- 절단된 SVD (truncated SVD)



출처: https://wikidocs.net/24949

잠재의미분석

B

$$\mathrm{U}_k$$

$$\mathrm{U} = \begin{array}{c} \\ T_1 \\ T_2 \\ T_3 \\ T_4 \\ T_5 \\ T_6 \\ \vdots \\ T_m \end{array} \begin{array}{c} \begin{array}{cccccc} C_1 & C_2 & C_3 & \cdots & C_m \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1m} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2m} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3m} \\ a_{41} & a_{42} & a_{43} & \cdots & a_{4m} \\ a_{51} & a_{52} & a_{53} & \cdots & a_{5m} \\ a_{61} & a_{62} & a_{63} & \cdots & a_{6m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mm} \end{pmatrix} \end{array}$$

$$\Sigma_k$$

$$\Sigma = \begin{array}{c} \\ T_1 \\ T_2 \\ T_3 \\ T_4 \\ \vdots \\ T_m \end{array} \begin{array}{c} \begin{array}{ccccc} D_1 & D_2 & D_3 & \cdots & D_n \end{array} \\ \begin{pmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ 0 & a_{22} & 0 & \cdots & 0 \\ 0 & 0 & a_{33} & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_{mm} \end{pmatrix} \end{array}$$

$$\mathrm{V}_k^T$$

$$\mathrm{V}^T = \begin{array}{c} \\ C_1 \\ C_2 \\ C_3 \\ C_4 \\ \vdots \\ C_n \end{array} \begin{array}{c} \begin{array}{ccccc} D_1 & D_2 & D_3 & \cdots & D_n \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ a_{41} & a_{42} & a_{43} & \cdots & a_{4n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{pmatrix} \end{array}$$

잠재의미분석

$$
\mathbf{X} = \begin{array}{c} \\ t_1 \\ t_2 \\ \\ t_m \end{array}
\begin{array}{c}
\begin{array}{cccc} d_1 & d_2 & & d_n \end{array} \\
\begin{bmatrix}
tf(t_1, d_1) & tf(t_1, d_2) & \cdots & tf(t_1, d_n) \\
tf(t_2, d_1) & tf(t_2, d_2) & \cdots & tf(t_2, d_n) \\
\vdots & \vdots & \ddots & \vdots \\
tf(t_m, d_1) & tf(t_m, d_2) & \cdots & tf(t_m, d_n)
\end{bmatrix}
\end{array}
$$

	문서1	문서2
오늘	2	0
내일	0	1
바나나	1	0
콜라	0	5
먹다	1	1

$$
\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T
$$

$$
= \underbrace{\begin{bmatrix} | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_r \\ | & & | \end{bmatrix}}_{m \times r}
\underbrace{\begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r \end{bmatrix}}_{r \times r}
\underbrace{\begin{bmatrix} - & \mathbf{v}_1 & - \\ & \vdots & \\ - & \mathbf{v}_n & - \end{bmatrix}}_{n \times r}
$$

$$
\mathbf{X}_k = \mathbf{U}_k\Sigma_k\mathbf{V}_k^T
$$

$$
= \underbrace{\begin{bmatrix} | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_k \\ | & & | \end{bmatrix}}_{m \times k}
\underbrace{\begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_k \end{bmatrix}}_{k \times k}
\underbrace{\begin{bmatrix} - & \mathbf{v}_1 & - \\ & \vdots & \\ - & \mathbf{v}_n & - \end{bmatrix}^T}_{n \times k}
$$

잠재의미분석

```
d1 : Chicago Chocolate. Retro candies made with love.
d2 : Chocolate sweets and candies. Collection with mini love hearts.
d3 : Retro sweets from Chicago for chocolate lovers.
```

$$
\mathbf{X} = \begin{array}{c} \\ \text{chicago} \\ \text{chocolate} \\ \text{retro} \\ \text{candy} \\ \text{made} \\ \text{love} \\ \text{sweet} \\ \text{collection} \\ \text{mini} \\ \text{heart} \end{array}
\begin{array}{c} \begin{array}{ccc} d_1 & d_2 & d_3 \end{array} \\
\begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix} \end{array}
$$

$$
\mathbf{U}_2 = \begin{array}{c} \\ \text{chicago} \\ \text{chocolate} \\ \text{retro} \\ \text{candy} \\ \text{made} \\ \text{love} \\ \text{sweet} \\ \text{collection} \\ \text{mini} \\ \text{heart} \end{array}
\begin{array}{c} \begin{array}{cc} \text{concept 1} & \text{concept 2} \end{array} \\
\begin{bmatrix} -0.318 & \mathbf{0.424} \\ \mathbf{-0.486} & 0.018 \\ -0.318 & \mathbf{0.424} \\ -0.333 & -0.148 \\ -0.166 & 0.257 \\ \mathbf{-0.488} & 0.018 \\ -0.320 & -0.239 \\ -0.168 & -0.406 \\ -0.168 & -0.406 \\ -0.168 & -0.406 \end{bmatrix} \end{array}
$$

$$
\Sigma_2 = \begin{bmatrix} 3.562 & 0 \\ 0 & 1.966 \end{bmatrix}
$$

$$
\mathbf{V}_2 = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \end{array}
\begin{array}{c} \begin{array}{cc} \text{concept 1} & \text{concept 2} \end{array} \\
\begin{bmatrix} -0.592 & 0.505 \\ -0.598 & -0.798 \\ -0.541 & 0.329 \end{bmatrix} \end{array}
$$

$$
\mathbf{q}_{\text{chicago}} = [\,1\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\,]
$$

$$
\mathbf{q}_{\text{candy}} = [\,0\ \ 0\ \ 0\ \ 1\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\,]
$$

Query	d_1	d_2	d_3
Chicago	0.891	-0.510	0.806
Candy	0.183	0.969	0.338

잠재의미분석

$$X = U\Sigma V^T$$

$$= \underbrace{\begin{bmatrix} | & & | \\ u_1 & \cdots & u_r \\ | & & | \end{bmatrix}}_{m \times r} \underbrace{\begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r \end{bmatrix}}_{r \times r} \underbrace{\begin{bmatrix} - & v_1 & - \\ & \vdots & \\ - & v_n & - \end{bmatrix}}_{n \times r}^T$$

$$V = X^T U \Sigma^{-1}$$

$$p = q^T U \Sigma^{-1}$$

$$\text{score}(q, d_i) = \cos(p, v_i) = \frac{p \cdot v_i}{\|p\| \, \|v_i\|}$$