

데이터 과학

L06: Similarity

Kookmin University

목차

❖ 비슷한 문서 찾기

- ❖ 문서를 표현하는 방법: Bag-of-Words
- ❖ Jaccard Similarity
- ❖ Cosine Similarity

❖ 비슷한 드라마 찾기

- ❖ Jaccard Similarity
- ❖ Cosine Similarity
- ❖ Centered Cosine Similarity (a.k.a. Pearson Correlation Coefficient)

서로 비슷한 문서는?

- 다음 세 문서(뉴스) 중 어떤 것이 서로 비슷할까?
 - 비슷한 정도를 어떻게 측정할 수 있을까?

[우크라 침공] 이근 전 대위 "우크라이나 무사히 도착"

(서울=연합뉴스) 김정진 기자 = 최근 우크라이나 의용군으로 참전하기 위해 출국했다고 밝힌 해군특수전전단(UDT/SEAL) 대위 출신 이근 씨가 우크라이나에 도착했다고 주장했다. 이 전 대위는 7일 자신의 인스타그램을 통해 "저의 팀은 무사히 우크라이나에 도착했다"며 "우리는 최전방에서 전투할 것"이라고 밝혔다. 자신을 둘러싼 여권 무효화 논란에 대해서는 "외교부는 시간 낭비하면서 우리 여권을 무효화 하는 것보다 어떻게 지원할 수 있는지나 ...

연합뉴스 (22.03.07)

[Y이슈] "러시아와 싸울 것"...이근, 우크라이나 의용군 참전(종합)

유튜브 채널 '가짜 사나이'로 이름을 알린 해군특수전전단(UDT/SEAL) 이근 전 대위가 팀을 꾸려 우크라이나 의용군으로 참전했다. 이 전 대위는 6일 자신의 인스타그램과 유튜브 채널 'ROKSEAL'을 통해 "우크라이나 대통령이 전 세계에 도움을 요청했을 때 'ROKSEAL'은 즉시 의용군 임무를 준비했다"며 "48시간 이내 계획을 수립하고 코디네이션, 장비를 준비했다"고 밝혔다. 그는 공식 절차를 통해 우크라이나로 출국하려 했으나 ...

YTN & YTN plus (22.03.07)

산불 번지는데 산불진화 헬기 왜 안오나 했더니...절반은 수리 중

지난 4일 발생한 경북·강원 지역 동해안 산불이 나흘째 이어지고 있는 가운데 산불 진화에 투입된 산림청 산림항공본부 소속 헬기가 노후화로 인한 정비 문제로 인해 절반 이상이 운항되지 못하고 있는 것으로 확인됐다. 산림청은 헬기가 산불 진화에 핵심 역할을 하고 있지만 보유한 헬기를 모두 가동하지 못하면서 산불 조기 진화에 어려움을 겪고 있다는 지적이 나온다. 7일 산림청에 따르면 산림항공본부가 보유한 헬기는 총 47대 ...

매일경제 (22.03.07)

Bag-Of-Words

- 문서를 정형화된 데이터로 표현할 수 있다면
→ 비슷한 정도(유사도)를 측정할 수 있다!

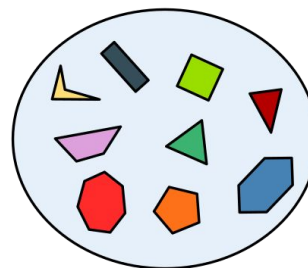
[우크라 침공] 이근 전 대위
"우크라이나 무사히 도착"

(서울=연합뉴스) 김정진 기자 = 최근 우크라이나 의용군으로 참전하기 위해 출국했다고 밝힌 해군특수전전단 (UDT/SEAL) 대위 출신 이근 씨가 우크라이나에 도착했다고 주장했다. 이 전 대위는 7일 자신의 인스타그램을 통해 "저의 팀은 무사히 우크라이나에 도착했다"며 "우리는 최전방에서 전투할 것"이라고 밝혔다. 자신을 둘러싼 여권 무효화 논란에 대해서는 "외교부는 시간 낭비하면서 우리 여권을 무효화 하는 것보다 어떻게 지원할 수 있는지나 ...



2	3	1	6	9	2	3	4
---	---	---	---	---	---	---	---

벡터 데이터



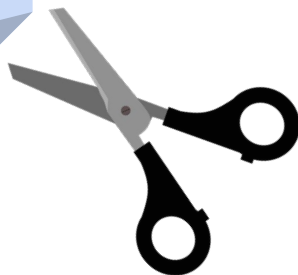
집합

Bag-Of-Words

- 문서를 단어들의 가방(Bag-Of-Words)으로 표현

[우크라 침공] 이근 전 대위
"우크라이나 무사히 도착"

(서울=연합뉴스) 김정진 기자 = 최근 우크라이나 의용군으로 참전하기 위해 출국했다고 밝힌 해군특수전전단(UDT/SEAL) 대위 출신 이근 씨가 우크라이나에 도착했다고 주장했다. 이 전 대위는 7일 자신의 인스타그램을 통해 "저의 팀은 무사히 우크라이나에 도착했다"며 "우리는 최전방에서 전투할 것"이라고 밝혔다. 자신을 둘러싼 여권 무효화 논란에 대해서는 "외교부는 시간 낭비하면서 우리 여권을 무효화 하는 것보다 어떻게 지원할 수 있는지나 ...



Bag-Of-Words

- 가정: 비슷한 문서는 단어들이 많이 겹칠 것이다!



단어가 많이 겹치네?
비슷하다!

겹치는 단어가 별로 없네...
안비슷하다...

Cosine Similarity

- 가정: 비슷한 문서는 단어들이 많이 겹칠 것이다!
→ 단어 가방을 **벡터로 표현**, 유사도 측정

은 어떻게 하지?



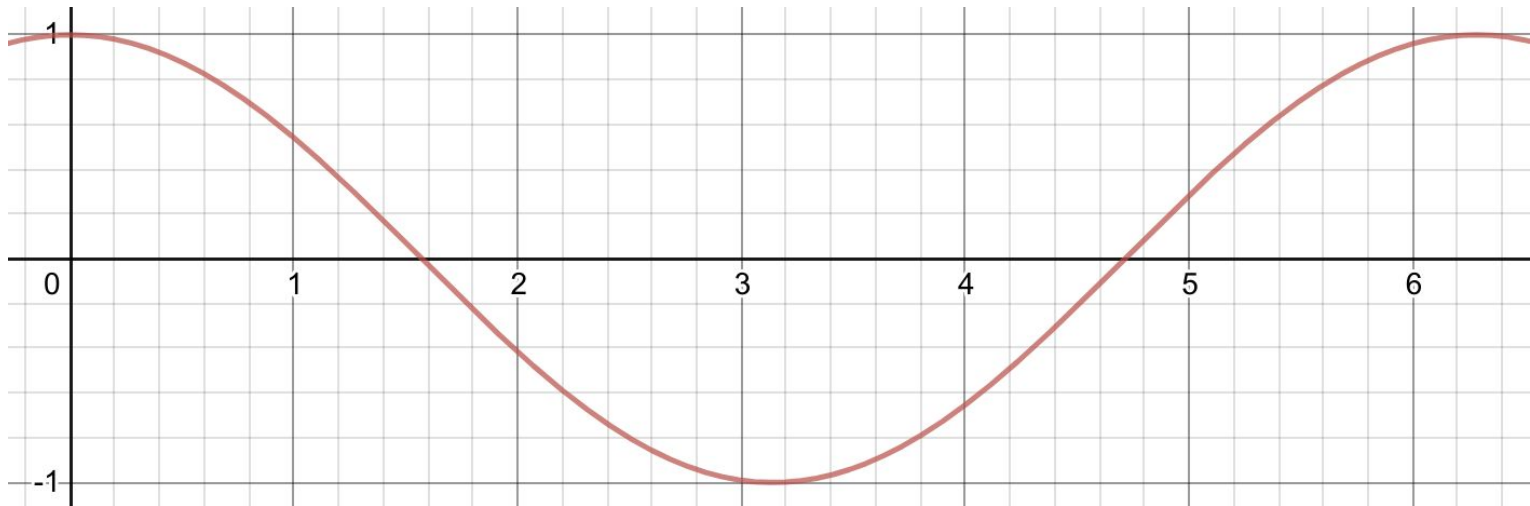
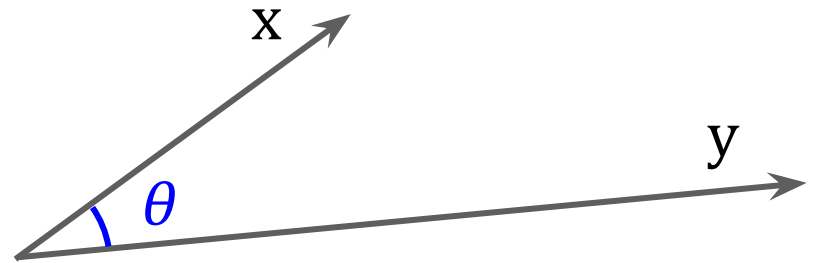
	우크라이나	대위	의용군	산불진화	왜	헬기	산불
[우크라 침공] 이군 전 대위 "우크라이나 무사히 도착"	5	5	2	0	6	0	0
[Y이슈] "러시아와 싸울 것"...이군, 우크라이나 의용군 참전(종합)	6	7	1	0	4	0	0
산불 번지는데 산불진화 헬기 왜 안오나 했더니...절반은 수리 중	1	0	0	6	5	5	7

단어 등장 횟수

Cosine Similarity

벡터로 봤을 때, 같은 방향을 가리키면 유사도가 높다!

$$\cos(\theta) = \frac{x \cdot y}{|x||y|}$$



Cosine Similarity

- 단어가 많이 겹치면, 코사인 유사도가 높다.

$$U(x, y) = \cos(\theta) = \frac{x \cdot y}{|x||y|}$$

- $U(a, b) = 0.95$
- $U(a, c) = 0.32$
- $U(b, c) = 0.22$



		우크라이나	대위	의용군	산불진화	왜	헬기	산불
a	[우크라 침공] 이군 전 대위 "우크라이나 무사히 도착"	5	5	2	0	6	0	0
b	[Y이슈] "러시아와 싸울 것"...이군, 우크라이나 의용군 참전(종합)	6	7	1	0	4	0	0
c	산불 번지는데 산불진화 헬기 왜 안오나 했더니...절반은 수리 중	1	0	0	6	5	5	7

Jaccard Similarity

- **Jaccard Similarity:** 집합간 유사도 측정
- BOW를 집합으로 보고 자카드 유사도 활용

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

- $U(a, b) = 1/4 = 1.0$
- $U(a, c) = 2/7 = 0.28$
- $U(b, c) = 2/7 = 0.28$



		우크라이나	대위	의용군	산불진화	왜	헬기	산불
a	[우크라 침공] 이군 전 대위 "우크라이나 무사히 도착"	1	1	1	0	1	0	0
b	[Y이슈] "러시아와 싸울 것"...이군, 우크라이나 의용군 참전(종합)	1	1	1	0	1	0	0
c	산불 번지는데 산불진화 헬기 왜 안오나 했더니...절반은 수리 중	1	0	0	1	1	1	1

An Example of Bag-Of-Words

- 다음 두 문장의 코사인유사도는?

It's bad, not good at all.

It's good, not bad at all.

두 문장은 반대의 의미지만, 유사도는 1.0... 왜..?

Drawbacks of BOW

- **단어의 순서**를 고려하지 않는다.
- **자주 등장하는 단어**가 유사도에 큰 영향을 미친다.
 - 예: a, the, of, you ...
- **Sparsity**: 문서에 등장하지 않은 단어가 훨씬 많다
 - 벡터로 표현하면? [1,0,0,0,0,0,1,0,0,0,0,0,0,0,0 ...]
- **새로운 단어**가 등장한다면...?
 - 새로운 단어는 BOW에서 아무런 의미가 없다.

n-gram

- 연달아 등장하는 단어 n개를 하나의 묶음으로!
 - 단어 순서를 고려한다
 - 자주 등장하는 단어의 '힘'이 약해진다
 - 더욱 심해지는 Sparsity...
 - 새로운 단어는 여전히 골칫거리

It's bad, not good at all.



1-gram = BOW

it

bad

not

good

at

all

2-gram = bigram

it bad

bad not

not good

good at

at all

3-gram = trigram

it bad not

bad not good

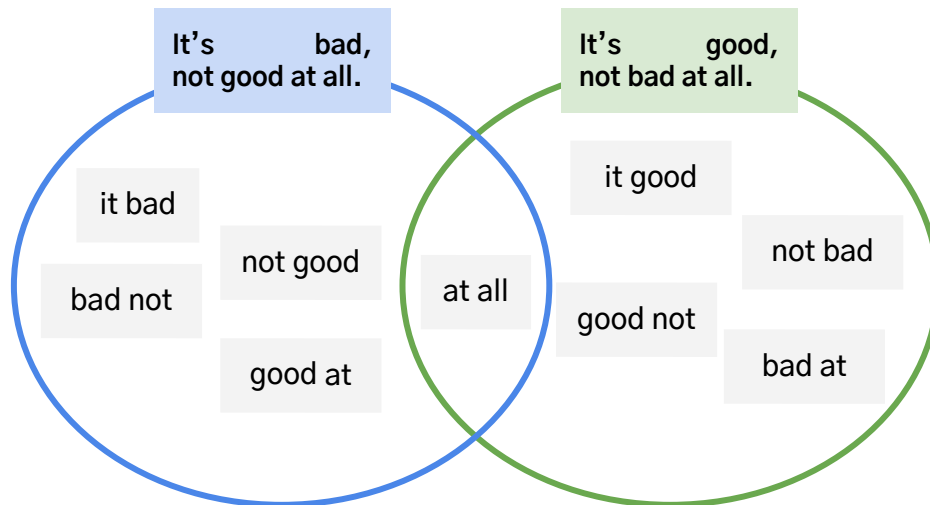
not good at

good at all

Similarities with n-gram

- n-gram을 벡터 혹은 집합으로 표현, 유사도 적용

	it bad	bad not	not good	good at	at all	it good	good not	not bad	bad at
It's bad, not good at all.	1	1	1	1	1	0	0	0	0
It's good, not bad at all.	0	0	0	0	1	1	1	1	1



코사인유사도 = 0.20
자카드 유사도 = $1/9 = 0.11$

TF-IDF

- 어떤 단어가 중요할까?

이 문서에서 자주 등장하는 단어
다른 문서에는 잘 등장하지 않는 단어

[우크라 침공] 이근 전 대위
"우크라이나 무사히 도착"

(서울=연합뉴스) 김정진 기자 = 최근
우크라이나 의용군으로 참전하기 위해
출국했다고 밝힌 해군특수전전단
(UDT/SEAL) 대위 출신 이근 씨가
우크라이나에 도착했다고 주장했다. 이
전 대위는 7일 자신의 인스타그램을 통해
"저의 팀은 무사히 우크라이나에
도착했다"며 "우리는 최전방에서 전투할
것"이라고 밝혔다. 자신을 둘러싼 여권
무효화 논란에 대해서는 "외교부는 시간
낭비하면서 우리 여권을 무효화 하는
것보다 어떻게 지원할 수 있는지나 ...

TF-IDF

f_{ij} = 문서 j 에서 단어 i 가 등장한 빈도수
 n_i = 단어 i 가 등장한 문서 수
 N = 전체 문서 수

- TF-IDF: 주어진 문서에서 단어마다 중요도 매기기
 - **TF (Term Frequency)**: 이 단어가 **이 문서에서** 얼마나 자주 등장했는가?

$$TF_{ij} = \frac{f_{ij}}{\max_k f_{kj}}$$

- **DF (Document Frequency)**: 이 단어가 **모든 문서에서** 얼마나 자주 등장했는가?
- **IDF (Inverse DF)**: DF의 역수

$$IDF_i = \log \frac{N}{n_i}$$

$$\text{TF-IDF score: } w_{ij} = TF_{ij} \times IDF_i$$

목차

❖ 비슷한 문서 찾기

- ❖ 문서를 표현하는 방법: Bag-of-Words
- ❖ Jaccard Similarity
- ❖ Cosine Similarity

❖ 비슷한 드라마 찾기

- ❖ Jaccard Similarity
- ❖ Cosine Similarity
- ❖ Centered Cosine Similarity (a.k.a. Pearson Correlation Coefficient)

추천시스템

내가 재밌게 본 드라마와 **비슷한** 드라마는?

NETFLIX

시그널과 비슷한 콘텐츠



이상한 변호사 우영우와 비슷한 콘텐츠



유사도 Similarity

두 드라마의 비슷함의 정도 (유사도)를 어떻게 측정할 수 있을까?

1. 장르나 키워드가 비슷하면 비슷하다.
2. 사람들의 평가가 비슷하면 비슷하다..!



장르·키워드 유사도

어떤 드라마가 서로 비슷할까?



한국 드라마
스릴러
TV 드라마
긴장감 넘치는
흥미진진



범죄
스페인 작품
스릴러
긴장감 넘치는
흥미진진



웹툰 원작
한국 드라마
사회 문제
다크

자카드 유사도 Jaccard Similarity

- 자카드 유사도 Jaccard Similarity: 두 집합이 얼마나 비슷한지를 측정

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

- $J(a, b) = 3/7 = 0.43$
- $J(a, c) = 1/7 = 0.14$
- $J(b, c) = 0/10 = 0$



한국 드라마
스릴러
TV 드라마
긴장감 넘치는
흥미진진



범죄
스페인 작품
스릴러
긴장감 넘치는
흥미진진



웹툰 원작
한국 드라마
사회 문제
다크

평가 유사도

종이의집, DP 중에 오징어게임과 더 비슷한 드라마는?



4.5

2.5

4.0

2.0

1.5



4.0

4.5

2.0



2.5

5.0

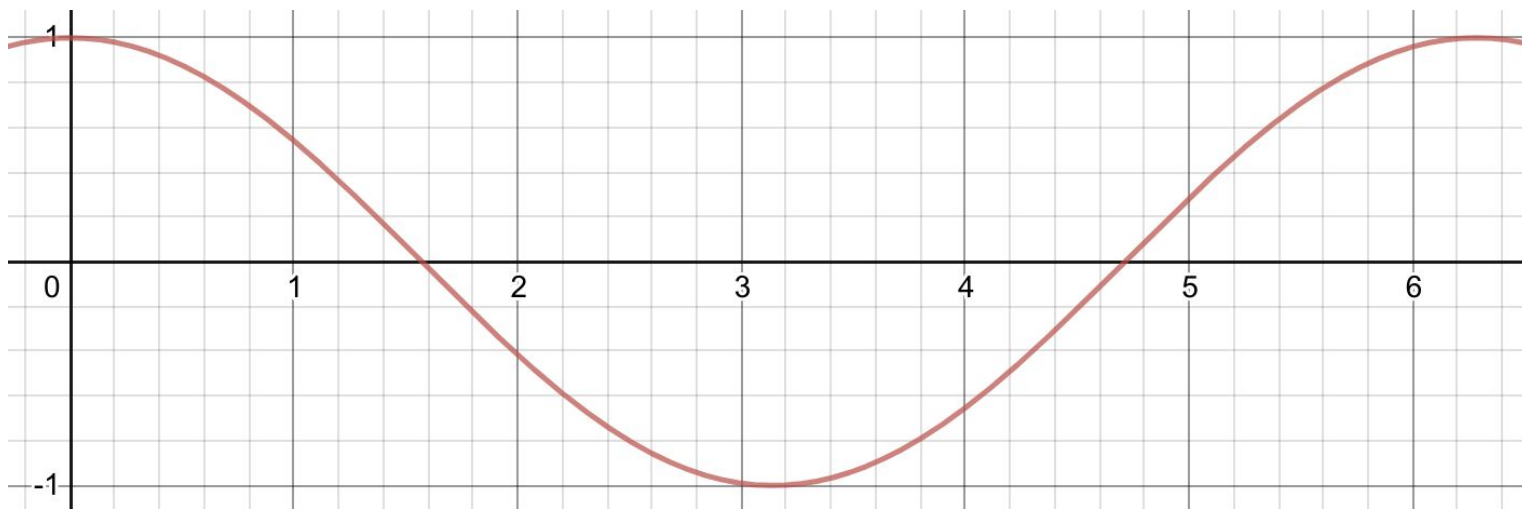
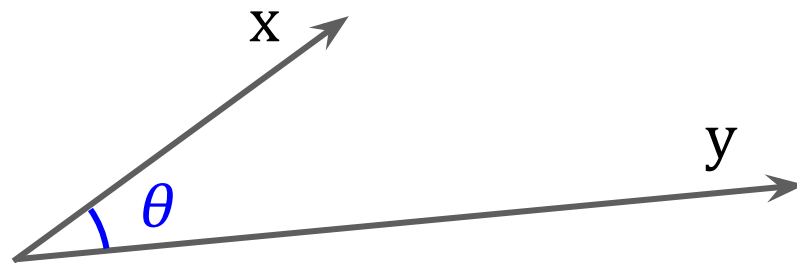
2.5

4.5

코사인 유사도 Cosine Similarity

벡터로 봤을 때, 같은 방향을 가리키면 유사도가 높다!

$$\cos(\theta) = \frac{x \cdot y}{|x||y|}$$












코사인 유사도 Cosine Similarity

코사인 유사도로 드라마 끼리의 유사도를 측정해보면?

$$U(x, y) = \cos(\theta) = \frac{x \cdot y}{|x||y|}$$

- $U(a, b) = 0.47$
- $U(a, c) = 0.67$
- $U(b, c) = 0.38$

	 A	 B	 C	 D	 E	 F
a 	4	2		4	2	1
b 	4		4			2
c 	2	5		2		4










코사인 유사도 Cosine Similarity

왜 $U(a, c)$ 의 유사도가 $U(a, b)$ 의 유사도보다 높게 계산되는가..?!

평가 안한 것을 0으로 생각하니까..!

$$U(x, y) = \cos(\theta) = \frac{x \cdot y}{|x||y|}$$

- $U(a, b) = 0.47$
- $U(a, c) = 0.67$
- $U(b, c) = 0.38$










	 A	 B	 C	 D	 E	 F
a 	4	2	0	4	2	1
b 	4	0	4	0	0	2
c 	2	5	0	2	0	4

Centered Cosine Similarity

- 평가하지 않은 경우, **평균 점수**를 부여

- $U(a, b) = 0.96$
- $U(a, c) = 0.67$
- $U(b, c) = 0.38$

이제 $U(a, b)$ 가 $U(a, c)$ 보다 크긴 한데,
왜 $U(a, c)$ 와 $U(b, c)$ 가 양수일까...?










	 A	 B	 C	 D	 E	 F	평균
a 	4	2	13/5	4	2	1	13/5
b 	4	10/3	4	10/3	10/3	2	10/3
c 	2	5	13/4	2	13/4	4	13/4

Centered Cosine Similarity

- 평가하지 않은 경우, **평균 점수**를 부여
- 모든 평점을 평균 점수만큼 빼줌

- $U(a, b) = 0.70$
- $U(a, c) = -0.82$
- $U(b, c) = -0.43$

이제 $U(a,b)$ 가 $U(a,c)$ 보다 크긴 한데,
왜 $U(a,c)$ 와 $U(b,c)$ 가 양수일까...?
⇒ **모든 평점이 긍정적이라서...!**

	 A	 B	 C	 D	 E	 F	평균
a 	7/5	-5/3	0	7/5	-5/3	-8/5	13/5
b 	2/3	0	2/3	0	0	-4/3	10/3
c 	-5/4	7/4	0	-5/4	0	3/4	13/4

추천하기

- 내가 재밌게 본 드라마 a와 유사도가 가장 높은 드라마 n개 추천하기
- 장르·키워드 유사도 $J(a, x)$ 와 평가 유사도 $U(a, x)$ 를 모두 활용

$$Score(a, x) = \alpha J(a, x) + (1 - \alpha)U(a, x)$$

- $Score(a, b) = 0.5 J(a, b) + 0.5 U(a, b) = 0.69$
- $Score(a, c) = 0.5 J(a, c) + 0.5 U(a, c) = 0.17$ (α 가 0.5일 때)

내가 재밌게 본 드라마



Questions?