

Introduction to Similarity Measures and the Limitations of Cosine Similarity

고차원 데이터 유사성 측정을 위한
코사인 유사도와 그의 한계점

Contents

1. Similarity Measure
2. Cosine Similarity
3. Limitations of Cosine Similarity
4. Introduction of DIEM
5. Research Topics



Similarity Measure _Vectors

- **Similarity Measure**

- 정의

- : Similarity Measure는 두 데이터 객체(벡터, 집합, 문자열 등) 간의 유사성을 수치적으로 표현하는 방법

- 활용 분야

- 패턴인식
 - 추천시스템
 - 문서분류
 - 클러스터링
 - 자연어 처리

- 벡터 간 유사성 측정

- 데이터의 방향, 크기, 분포 등을 기반으로 두 객체 간 관계를 정량적으로 평가
 - 고차원 데이터 분석에서 중요한 도구로 활용

Similarity Measure _Vectors

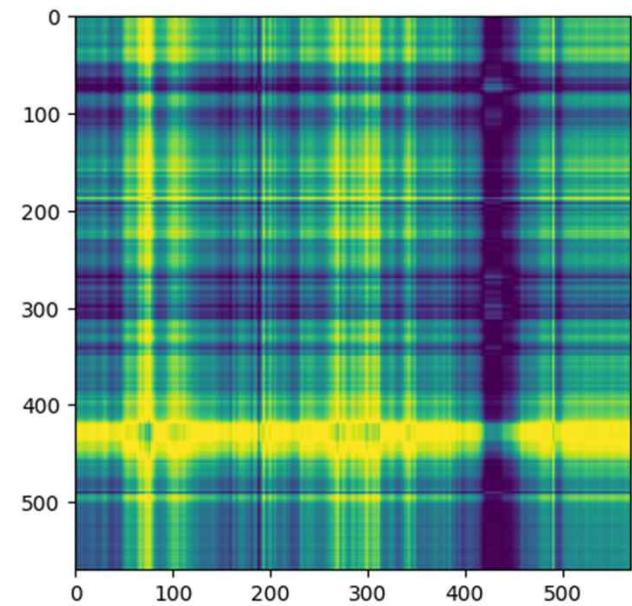
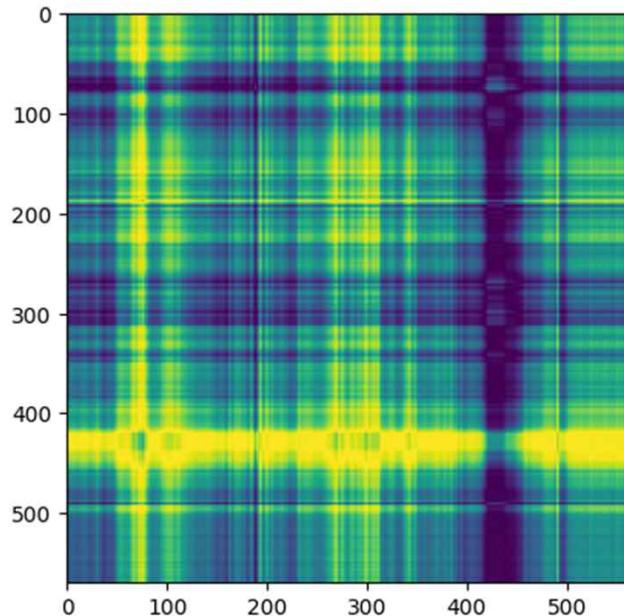
- **Similarity Measure**

- 정의

: Similarity Measure는 두 데이터 객체(벡터, 집합, 문자열 등) 간의 유사성을 수치적으로 표현하는 방법

- 활용 분야

- 패턴인식
 - 추천시스템
 - 문서분류
 - 클러스터링
 - 자연어 처리



Similarity Measure _Vectors

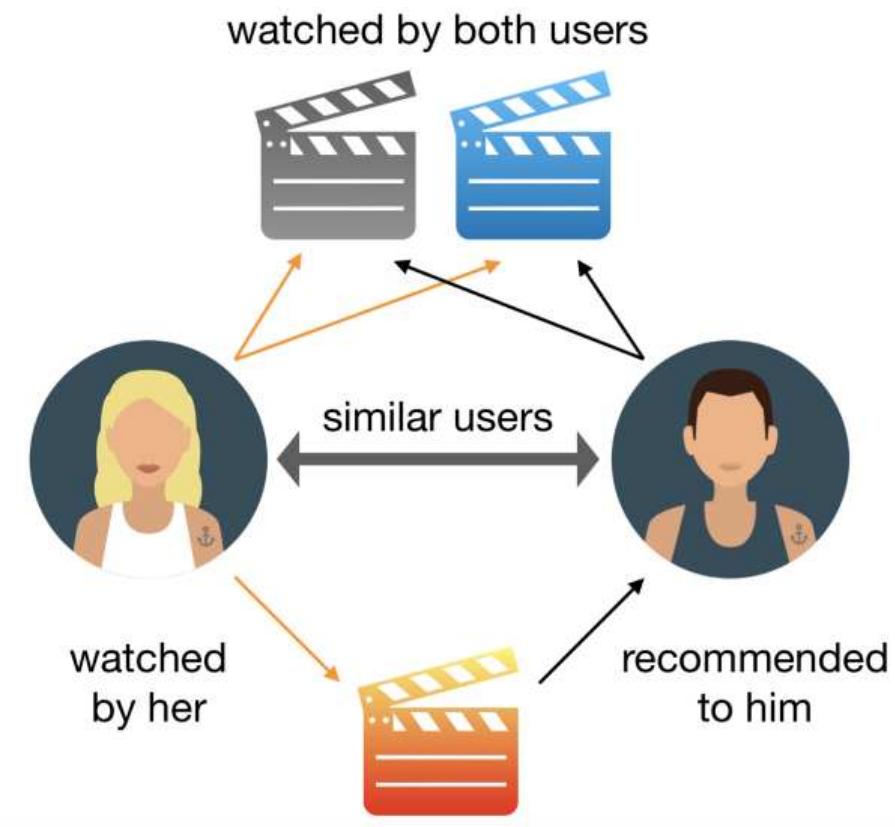
- **Similarity Measure**

- 정의

: Similarity Measure는 두 데이터 객체(벡터, 집합, 문자열 등) 간의 유사성을 수치적으로 표현하는 방법

- 활용 분야

- 패턴인식
- 추천시스템
- 문서분류
- 클러스터링
- 자연어 처리



reference : [1] A Guide to Similarity Measures, [2] Evolution of Semantic Similarity—A Survey, [3] A SURVEY ON SIMILARITY MEASURES IN TEXT.pdf

Similarity Measure _Vectors

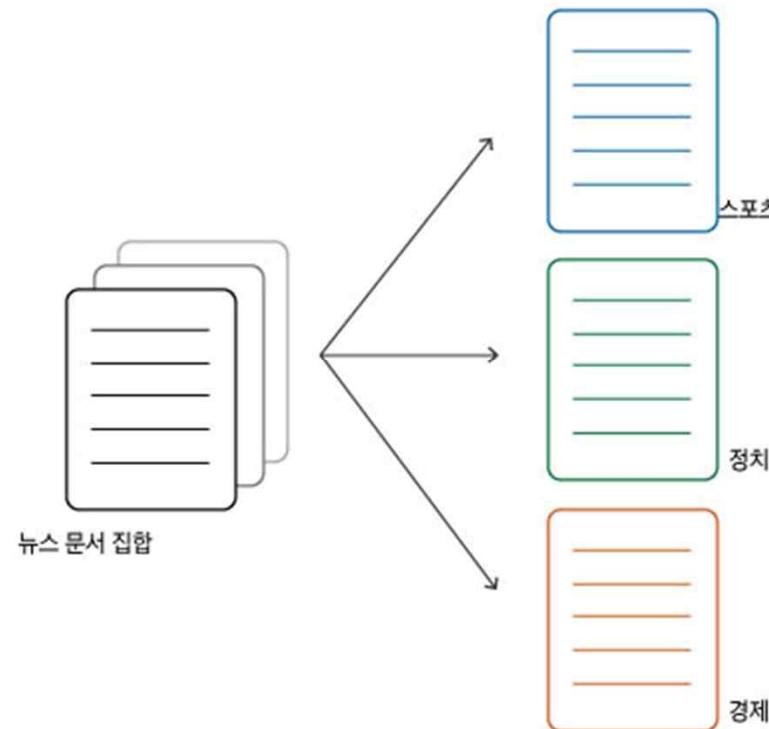
- **Similarity Measure**

- 정의

: Similarity Measure는 두 데이터 객체(벡터, 집합, 문자열 등) 간의 유사성을 수치적으로 표현하는 방법

- 활용 분야

- 패턴인식
 - 추천시스템
 - **문서분류**
 - 클러스터링
 - 자연어 처리



Similarity Measure _Vectors

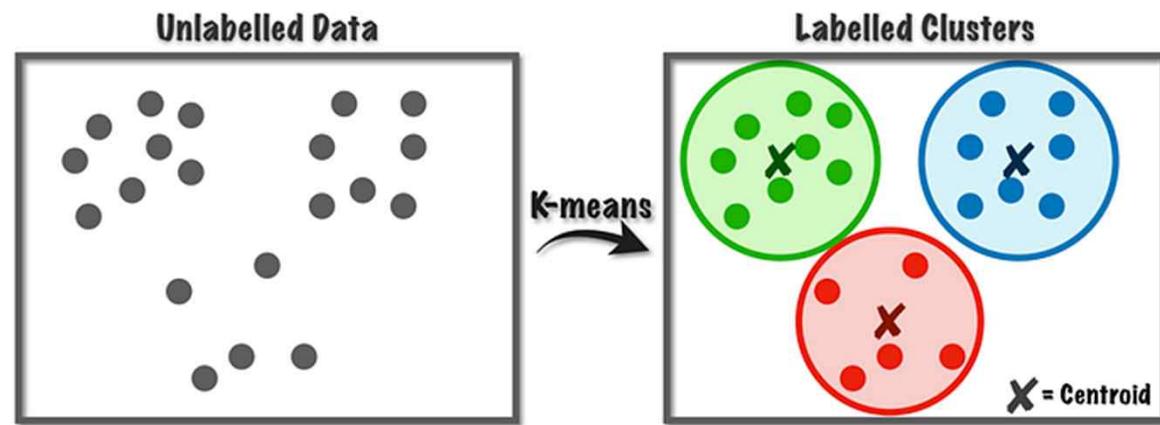
- **Similarity Measure**

- 정의

: Similarity Measure는 두 데이터 객체(벡터, 집합, 문자열 등) 간의 유사성을 수치적으로 표현하는 방법

- 활용 분야

- 패턴인식
 - 추천시스템
 - 문서분류
 - 클러스터링**
 - 자연어 처리



Similarity Measure _Vectors

- **Similarity Measure**

- 정의

: Similarity Measure는 두 데이터 객체(벡터, 집합, 문자열 등) 간의 유사성을 수치적으로 표현하는 방법

- 활용 분야

- 패턴인식
- 추천시스템
- 문서분류
- 클러스터링
- 자연어 처리

```
[analogy] king:man = queen:?
woman: 5.16015625
veto: 4.9296875
ounce: 4.69140625
earthquake: 4.6328125
successor: 4.609375
```

```
[analogy] take:took = go:?
went: 4.55078125
points: 4.25
began: 4.09375
comes: 3.98046875
oct.: 3.90625
```

```
[analogy] car:cars = child:?
children: 5.21875
average: 4.7265625
yield: 4.20703125
cattle: 4.1875
priced: 4.1796875
```

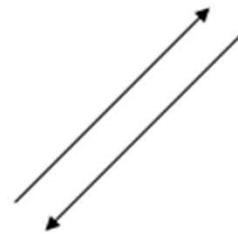
```
[analogy] good:better = bad:?
more: 6.6484375
less: 6.0625
rather: 5.21875
slower: 4.734375
greater: 4.671875
```

Cosine Similarity

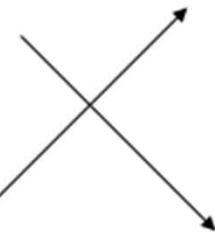
- **Cosine Similarity**

- 두 벡터 간 유사성을 수치적으로 표현하는 방법
- 크기보다 방향에 집중하여, 데이터의 패턴을 이해하는 데 유용
- 특히 고차원 데이터에서도 직관적으로 유사성을 계산 가능

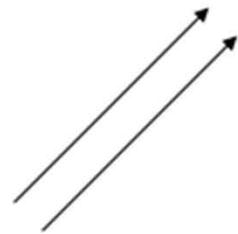
$$\text{similarity} = \cos(\Theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$



코사인 유사도 : -1



코사인 유사도 : 0



코사인 유사도 : 1

Similarity Measure _Vectors

- **Similarity Measure**

- 정의

- : Similarity Measure는 두 데이터 객체(벡터, 집합, 문자열 등) 간의 유사성을 수치적으로 표현하는 방법

- 활용 분야

- 패턴인식
 - 추천시스템
 - 문서분류
 - 클러스터링
 - 자연어 처리

- 벡터 간 유사성 측정

- 데이터의 방향, 크기, 분포 등을 기반으로 두 객체 간 관계를 정량적으로 평가
 - 고차원 데이터 분석에서 중요한 도구로 활용

Cosine Similarity

- **Cosine Similarity**

- 두 벡터 간 유사성을 수치적으로 표현하는 방법
- 크기보다 방향에 집중하여, 데이터의 패턴을 이해하는 데 유용
- 특히 고차원 데이터에서도 직관적으로 유사성을 계산 가능

- 예) 이미지 분석



image 1



image 2



CNN



image 1 representation vector

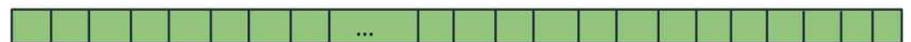


image 2 representation vector

$$\text{cosine similarity} = S_C(A, B) := \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}},$$

reference : [1] A Guide to Similarity Measures, [2] Evolution of Semantic Similarity—A Survey, [3] A SURVEY ON SIMILARITY MEASURES IN TEXT, [Book] 딥러닝을 이용한 자연어 처리 입문

Cosine Similarity

- **Cosine Similarity**

- 두 벡터 간 유사성을 수치적으로 표현하는 방법
- 크기보다 방향에 집중하여, 데이터의 패턴을 이해하는 데 유용
- 특히 고차원 데이터에서도 직관적으로 유사성을 계산 가능

- 예) 문서 분류

-	바나나	사과	저는	좋아요
문서1	0	1	1	1
문서2	1	0	1	1
문서3	2	0	2	2

```
print(cos_sim(doc1, doc2)) #문서1과 문서2의 코사인 유사도
print(cos_sim(doc1, doc3)) #문서1과 문서3의 코사인 유사도
print(cos_sim(doc2, doc3)) #문서2과 문서3의 코사인 유사도
```

0.67
0.67
1.00

Cosine Similarity

- **Cosine Similarity**

- 두 벡터 간 유사성을 수치적으로 표현하는 방법
- 크기보다 방향에 집중하여, 데이터의 패턴을 이해하는 데 유용
- 특히 고차원 데이터에서도 직관적으로 유사성을 계산 가능

- 예) 추천 시스템 (영화)

```
get_recommendations('The Dark Knight Rises')
```

-	...	original_title	overview	...	title	video
0	...	Toy Story	Led by Woody, Andy's toys live happily in his ... 중략	Toy Story	False
1	...	Jumanji	When siblings Judy and Peter discover an encha ... 중략	Jumanji	False

12481	The Dark Knight
150	Batman Forever
1328	Batman Returns
15511	Batman: Under the Red Hood
585	Batman
9230	Batman Beyond: Return of the Joker
18035	Batman: Year One
19792	Batman: The Dark Knight Returns, Part 1
3095	Batman: Mask of the Phantasm
10122	Batman Begins
Name: title, dtype: object	

reference : [1] A Guide to Similarity Measures, [2] Evolution of Semantic Similarity—A Survey, [3] A SURVEY ON SIMILARITY MEASURES IN TEXT, [Book] 딥러닝을 이용한 자연어 처리 입문

Limitations of Cosine similarity

- **Matrix Factorization and Solution Multiplicity**

- 추천 시스템의 MF 모델 :

사용자 – 아이템 관계를 저차원 잠재 공간으로 축소하여 학습

- 핵심 문제 :

동일한 결과를 나타내는 여러 해가 존재 가능

모델이 선택하는 해에 따라 코사인 유사도 값이 고유하지 않음

$$X = \begin{bmatrix} 5 & 3 & ? & 1 & ? \\ 4 & ? & 2 & ? & 3 \\ ? & 2 & 4 & 1 & ? \\ 1 & ? & ? & 5 & 3 \end{bmatrix} \quad A = \begin{bmatrix} 0.2 & 0.8 \\ 0.6 & 0.3 \\ 0.4 & 0.7 \\ 0.5 & 0.1 \\ 0.9 & 0.5 \end{bmatrix} \quad B^T = \begin{bmatrix} 0.3 & 0.5 & 0.8 & 0.2 & 0.7 \\ 0.7 & 0.1 & 0.6 & 0.4 & 0.9 \end{bmatrix}$$

$$XA = \begin{bmatrix} 3.9 & 5.3 \\ 3.8 & 3.9 \\ 2.6 & 3.5 \\ 4.9 & 3.8 \end{bmatrix} \quad XAB^T = \begin{bmatrix} 3.9 & 5.3 \\ 3.8 & 3.9 \\ 2.6 & 3.5 \\ 4.9 & 3.8 \end{bmatrix} \begin{bmatrix} 0.3 & 0.5 & 0.8 & 0.2 & 0.7 \\ 0.7 & 0.1 & 0.6 & 0.4 & 0.9 \end{bmatrix}$$

Limitations of Cosine similarity

- Matrix Factorization and Solution Multiplicity (Detail Part)

To do $X \approx XAB^T$, $\min_{A,B} \|X - XAB^T\|_F^2 + \lambda \|AB^T\|_F^2$: (Eq. 5)

If \hat{A}, \hat{B} are solutions of Equation (5), then $\hat{A}R$ and $\hat{B}R$ are also solutions for an arbitrary rotation matrix R .

Proof: Let \hat{A}, \hat{B} be solutions of Equation (5). Suppose that $R \in \mathbb{R}^{k \times k}$ is a rotation matrix such that $R^T R = R R^T = I$. It suffices to show the following two properties:

$$\|X - X\hat{A}R(\hat{B}R)^T\|_F^2 = \|X - X\hat{A}\hat{B}^T\|_F^2 \quad \text{and} \quad \|\hat{A}R(\hat{B}R)^T\|_F^2 = \|\hat{A}\hat{B}^T\|_F^2.$$

1) For the first term:

$$X - X\hat{A}R(\hat{B}R)^T = X - X\hat{A}RR^T\hat{B}^T = X - X\hat{A}\hat{B}^T,$$

2) For the second term:

$$\hat{A}R(\hat{B}R)^T = \hat{A}RR^T\hat{B}^T = \hat{A}\hat{B}^T.$$

Thus, the solution remains invariant under rotation. \square

Similarly, the cosine similarity remains invariant under rotation. Let u and v denote the row vectors from XA and B , corresponding to specific indices i_u and i_v , respectively.

Now consider the vectors uR and vR obtained from the same rows of XA^R and B^R . The cosine similarity remains unchanged:

$$\cos_{\text{sim}}(u, v) = \frac{u \cdot v}{\|u\| * \|v\|}, \quad \cos_{\text{sim}}(uR, vR) = \frac{uR \cdot vR}{\|uR\| * \|vR\|} = \frac{u \cdot v}{\|u\| * \|v\|}.$$

This is because $\|uR\| = \|u\|$ and $\|vR\| = \|v\|$, as R is a rotation matrix. Therefore, cosine similarity remains invariant under rotation.

Limitations of Cosine similarity

- Matrix Factorization and Solution Multiplicity (Detail Part)

To do, $X \approx XAB^\top$, $\min_{A,B} ||X - XAB^\top||_F^2 + \lambda ||AB^\top||_F^2$

If $\hat{A}\hat{B}^\top$ is a solution of the first objective, then so is $ADD^{-1}\hat{B}^\top$, where $D \in \mathbb{R}^{k \times k}$ is an arbitrary diagonal matrix.

We can hence define a new solution (as a function of D) as follows:

$$\begin{aligned}\hat{A}^{(D)} &:= \hat{A}D \\ \hat{B}^{(D)} &:= \hat{B}D^{-1}\end{aligned}$$

Case 1 : $D = \text{dMat}(..., \frac{1}{1+\lambda/\sigma_i^2}, ...)^{\frac{1}{2}}$

$$\text{cosSim}(\hat{B}_{(1)}^{(D)}, \hat{B}_{(1)}^{(D)}) = VV^\top = I$$

$$\text{cosSim}(X\hat{A}_{(1)}^{(D)}, \hat{B}_{(1)}^{(D)}) = \Omega_A \cdot X \cdot \hat{A}_{(1)} \hat{B}_{(1)}^\top$$

$$\text{cosSim}(X\hat{A}_{(1)}^{(D)}, X\hat{A}_{(1)}^{(D)}) = \Omega_A \cdot X \cdot X^\top \cdot \Omega_A$$

Case 2 : $D = \text{dMat}(..., \frac{1}{1+\lambda/\sigma_i^2}, ...)^{-\frac{1}{2}}$

$$\text{cosSim}(\hat{B}_{(1)}^{(D)}, \hat{B}_{(1)}^{(D)}) = \Omega_B \cdot V \cdot \text{dMat}(..., \frac{1}{1+\lambda/\sigma_i^2}, ...)^2 \cdot V^\top \cdot \Omega_B$$

$$\text{cosSim}(X\hat{A}_{(1)}^{(D)}, \hat{B}_{(1)}^{(D)}) = \Omega_A \cdot X \cdot \hat{A}_{(1)} \cdot \hat{B}_{(1)}^\top \cdot \Omega_B$$

$$\text{cosSim}(X\hat{A}_{(1)}^{(D)}, X\hat{A}_{(1)}^{(D)}) = \Omega_A \cdot X \cdot X^\top \cdot \Omega_A$$

$$\hat{A}_{(1)} = \hat{B}_{(1)} := V_k \cdot \text{dMat}(..., \frac{1}{1+\lambda/\sigma_i^2}, ...)^{\frac{1}{2}}_k$$

Second limitation of Cosine similarity

- **Its Limitations in High Dimensions (Cures of dimensionality)**

- 차원 의존성 문제 :

- 차원이 증가함에 따라, 코사인 유사도 값이 특정 범위로 수렴

- 데이터 간 차이 구별이 어려움

- 크기 정보 소실:

- 벡터의 크기를 무시하고 방향성만 평가

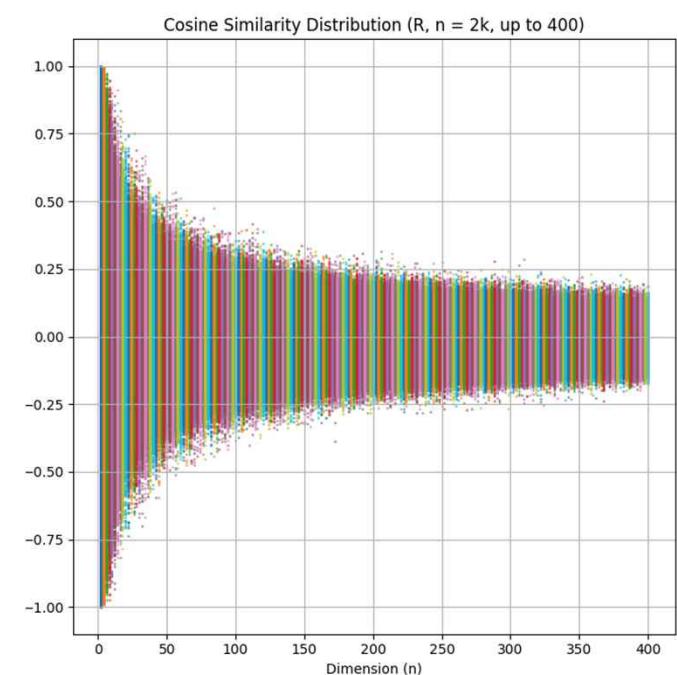
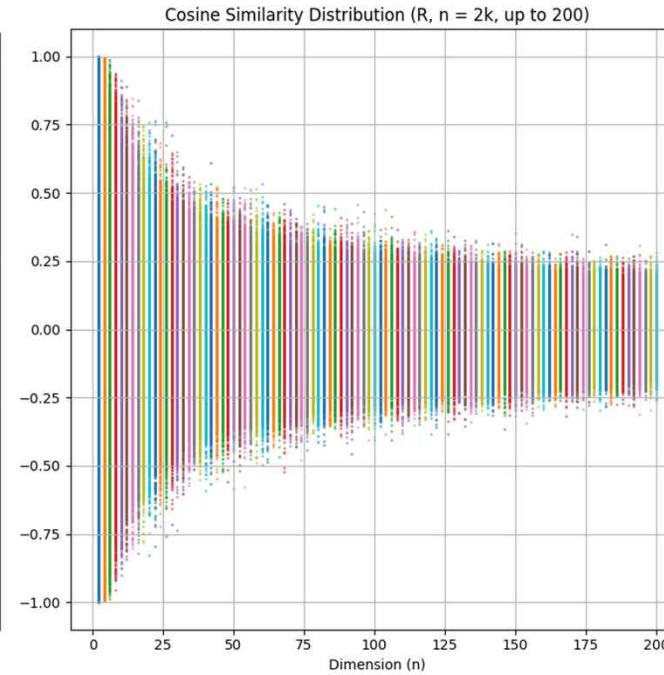
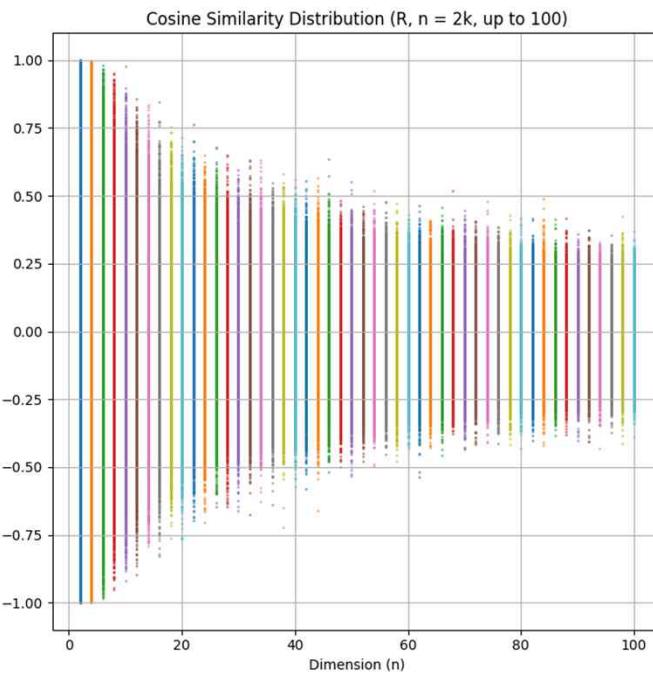
- 크기가 중요한 데이터에는 적합하지 않음

Second limitation of Cosine similarity

- **Its Limitations in High Dimensions**

- 차원이 증가함에 따라 코사인 유사도 값이 특정 범위로 수렴

R vector space, (-1,1)

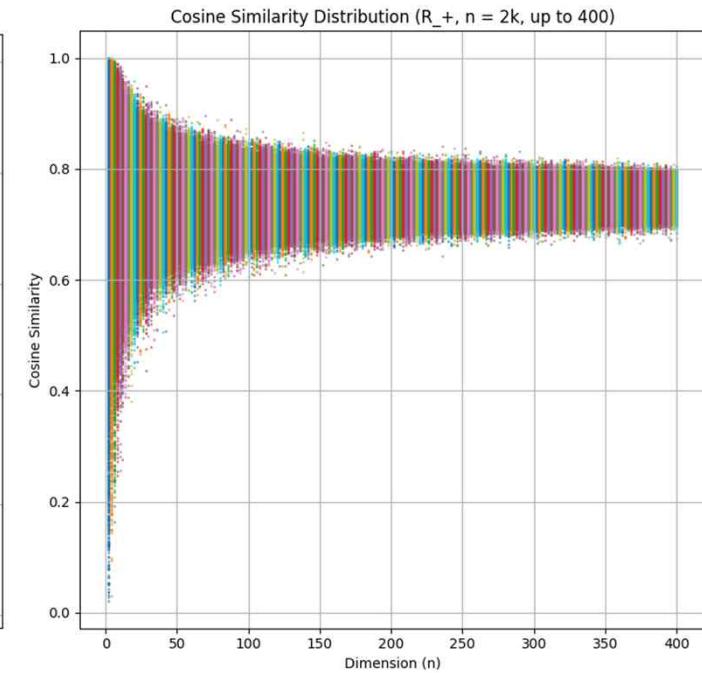
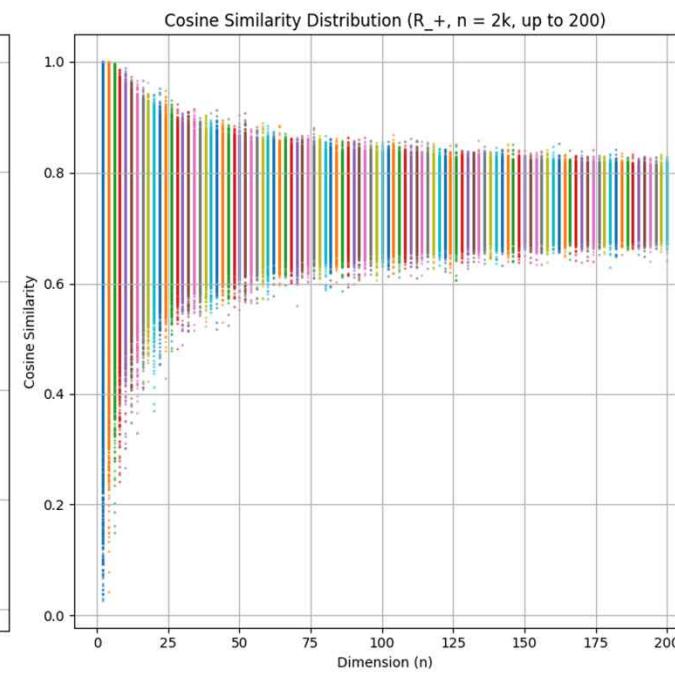
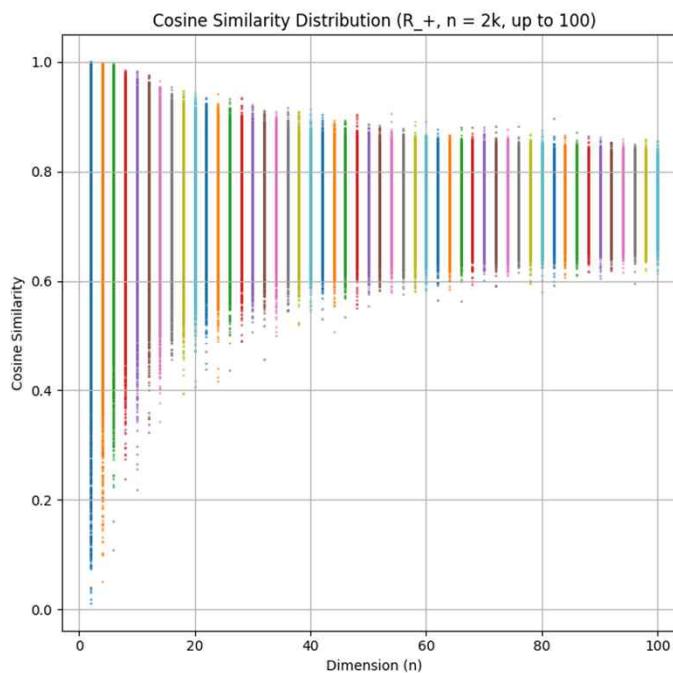


Second limitation of Cosine similarity

- **Its Limitations in High Dimensions**

- 차원이 증가함에 따라 코사인 유사도 값이 특정 범위로 수렴

R^+ vector space, (0,1)

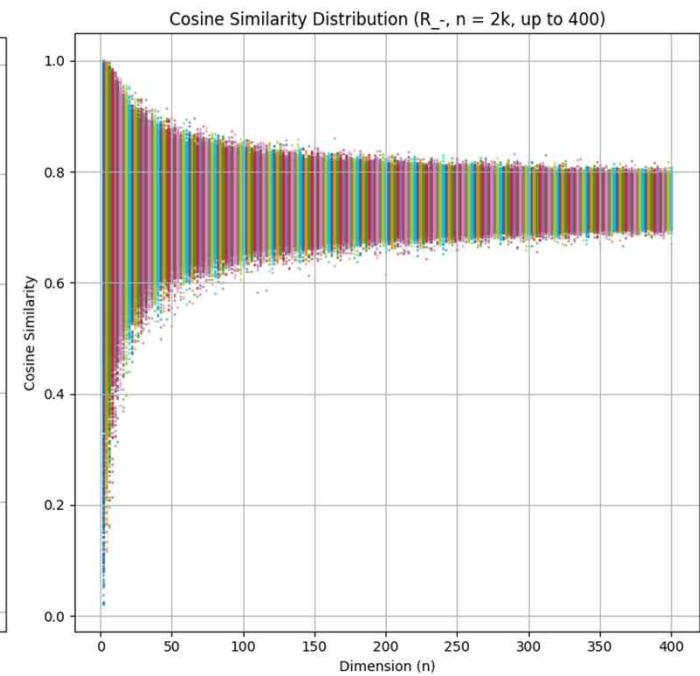
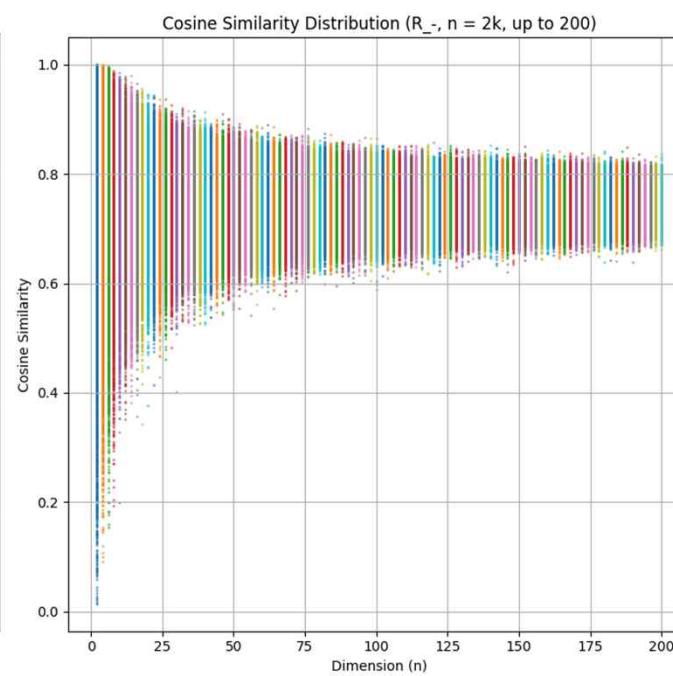
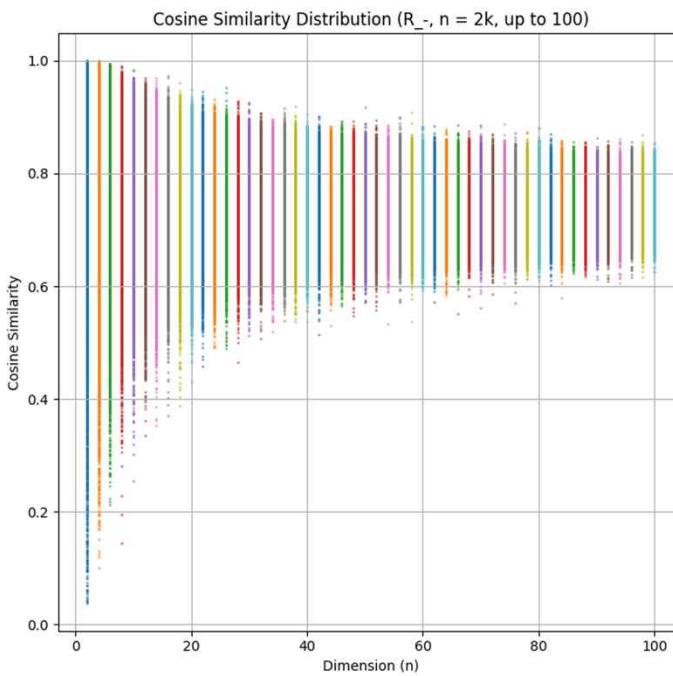


Second limitation of Cosine similarity

- Its Limitations in High Dimensions

- 차원이 증가함에 따라 코사인 유사도 값이 특정 범위로 수렴

R^- vector space, (-1,0)



DIEM (Dimension Insensitive Euclidean Metric)

- Introduction to the DIEM Metric

- DIEM : Dimension Insensitive Euclidean Metric

기존 코사인 유사도와 유클리드 거리가 가진 차원 의존성 문제를 해결

- DIEM 주요 특징 :

차원이 증가하더라도, 데이터 간 관계를 왜곡하지 않음

분산 기준으로 정규화 하여 차원의 영향을 최소화

→ 고차원 데이터에서 기존 메트릭의 한계를 극복

$$\begin{matrix} & \begin{matrix} 1 & 2 & \dots & n \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ \vdots \\ m \end{matrix} & \left[\begin{matrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ a_{31} & a_{32} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{matrix} \right] \end{matrix}$$

$$DIEM = \frac{\nu_M - \nu_m}{\sigma(n)^2} \left(\sqrt{\sum_{i=1}^n (a_i - b_i)^2} - E[d(n)] \right)$$

DIEM (Dimension Insensitive Euclidean Metric)

- Introduction to the DIEM Metric (Detail Part)

- (a_i, b_i) : 벡터 a 와 b 의 i -번째 요소.
- $\sum_{i=1}^n (a_i - b_i)^2$: 벡터 a 와 b 간의 유클리드 거리를 계산하는 식.
- $E[d(n)]$: 차원 n 에서 유클리드 거리의 기대값.
- $\sigma(n)^2$: 차원 n 에서 유클리드 거리 분포의 분산.
- v_M, v_m : v_M 은 데이터 값의 최대값, v_m 은 데이터 값의 최소값.
- n : 벡터의 차원.

$$\begin{matrix} & \begin{matrix} 1 & 2 & \dots & n \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ \vdots \\ m \end{matrix} & \left[\begin{matrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ a_{31} & a_{32} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{matrix} \right] \end{matrix}$$

$$DIEM = \frac{v_M - v_m}{\sigma(n)^2} \left(\sqrt{\sum_{i=1}^n (a_i - b_i)^2} - E[d(n)] \right)$$

DIEM (Dimension Insensitive Euclidean Metric)

- Introduction to the DIEM Metric (Detail Part.2)

$E[d(n)]$: 차원 n 에서 유클리드 거리의 기대값.

$$\begin{aligned}
 E[d] &= E\left[\sqrt{\sum_{i=1}^n (a_i - b_i)^2}\right] \leq \sqrt{E\left[\sum_{i=1}^n (a_i - b_i)^2\right]} \\
 &= \sqrt{n \cdot E[(a - b)^2]} \\
 &= \sqrt{n} \left(\iint_{v_m v_m}^{v_M v_M} \frac{(a - b)^2}{(v_M - v_m)^2} dadb \right)^{\frac{1}{2}} \quad (10)
 \end{aligned}$$

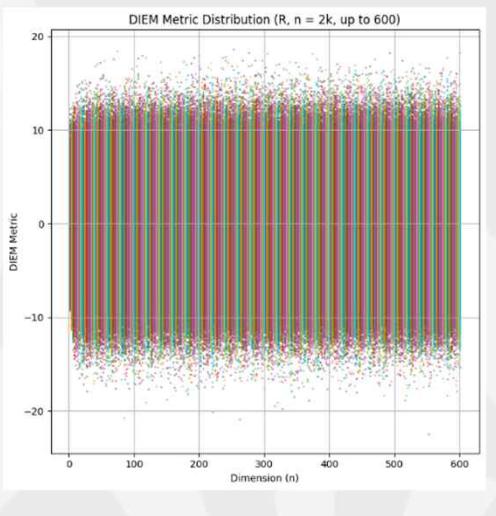
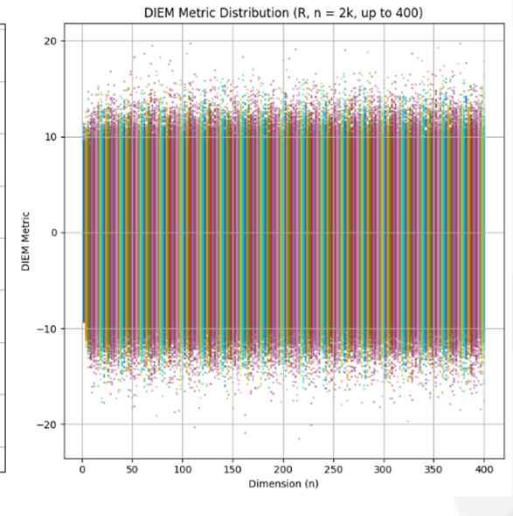
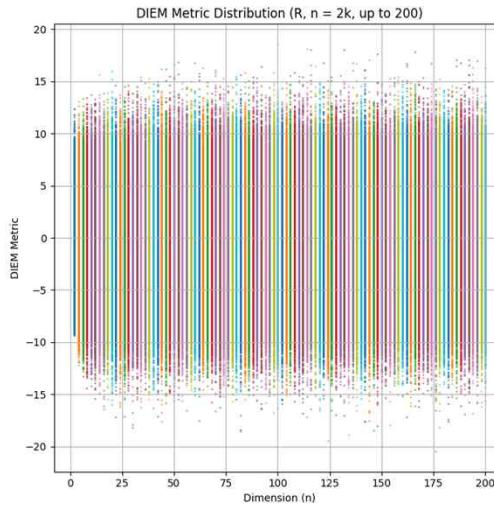
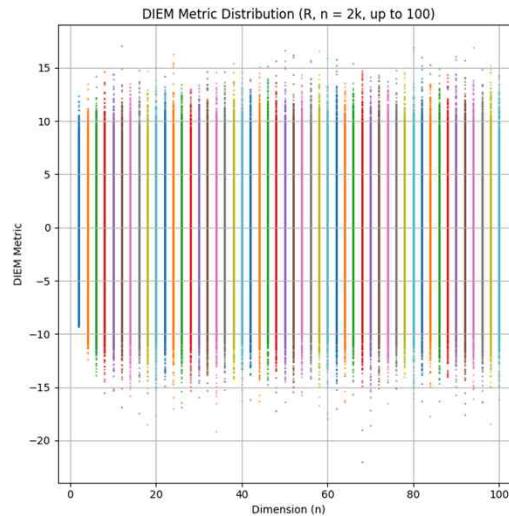
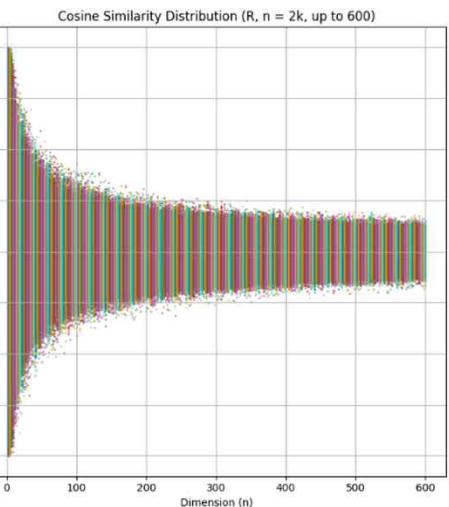
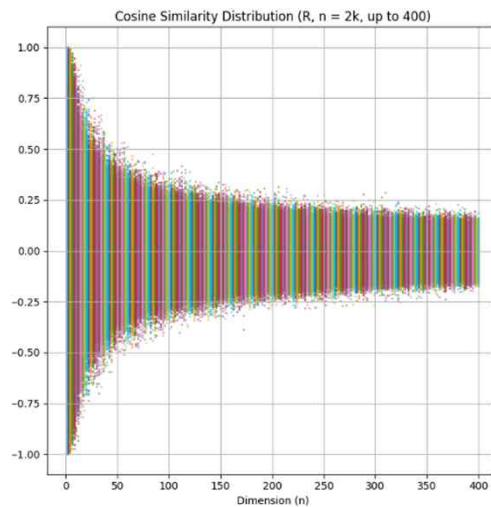
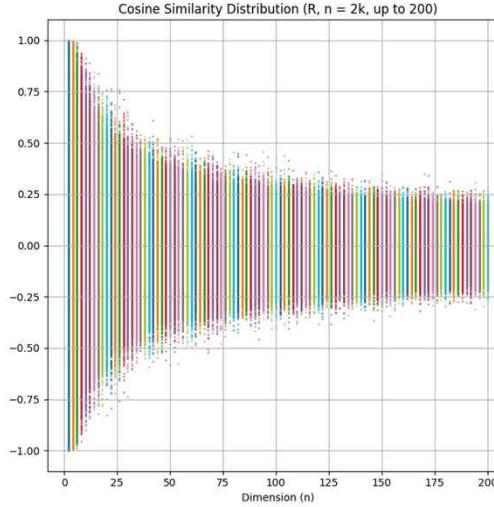
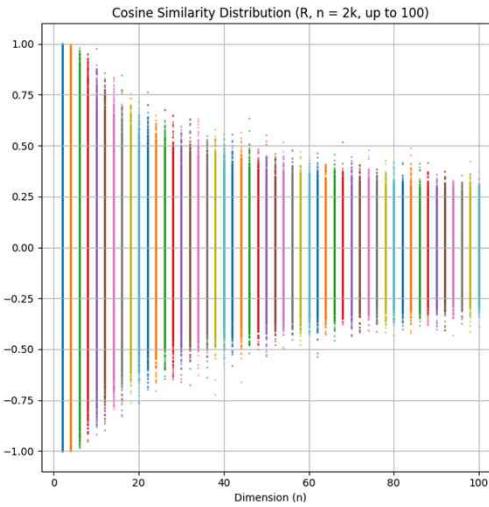
The integral in Equation 10 can easily be solved by direct integration. We leave readers the pleasure to do so. The final results can be expressed in the following form:

$$\begin{aligned}
 E[d] &\leq \sqrt{n} \sqrt{\frac{2}{3}(v_M^2 + v_M v_m + v_m^2) - \frac{1}{2}(v_M + v_m)^2} \\
 &= \sqrt{\frac{n}{6}}(v_M - v_m) \quad (11)
 \end{aligned}$$

Equation 11 provides an analytical upper bound to the expected Euclidean distance between any two random vectors $\mathbf{a}, \mathbf{b} \sim U(v_m, v_M)$. Figure 8 provides a graphical representation of

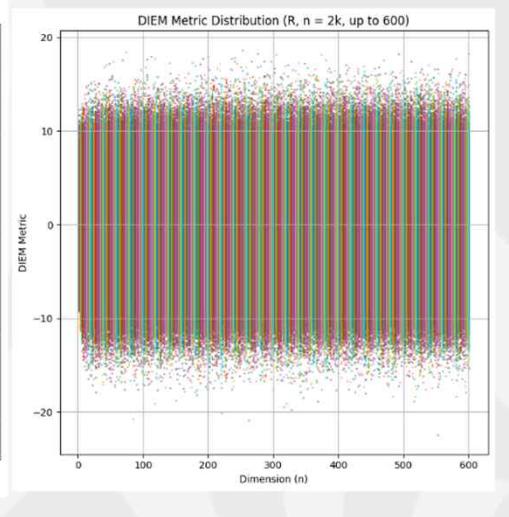
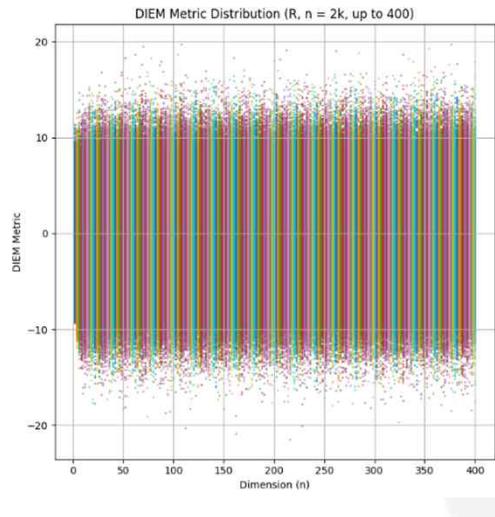
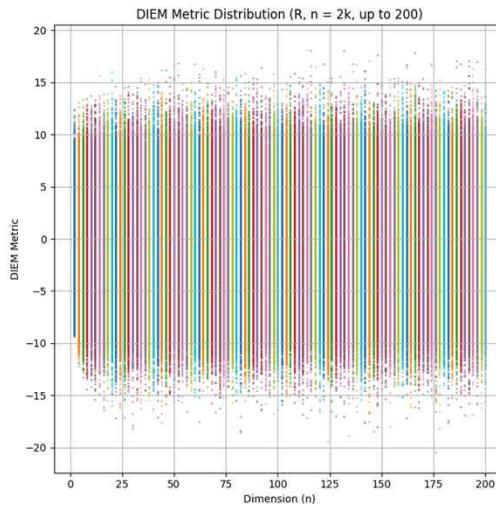
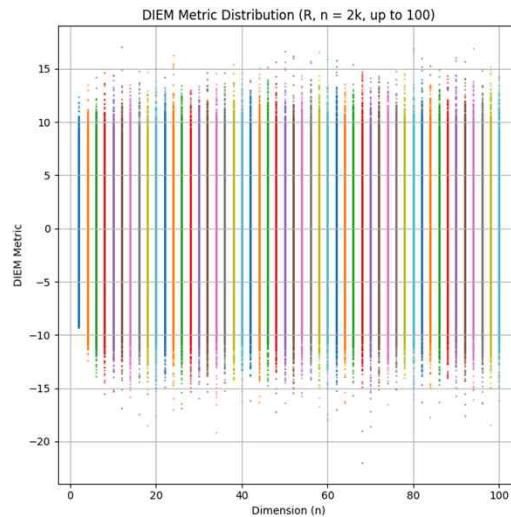
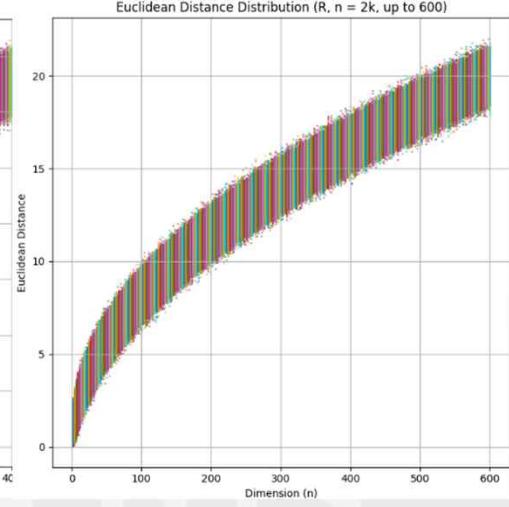
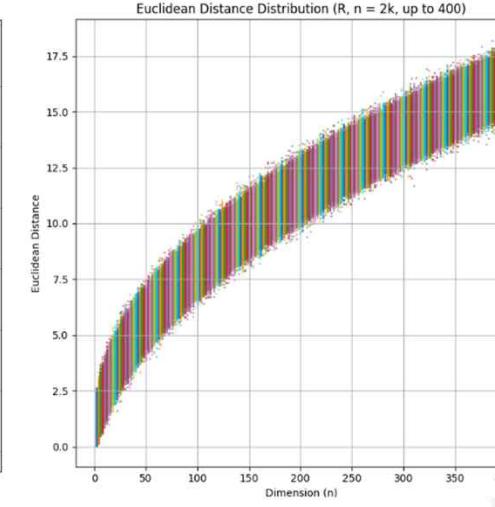
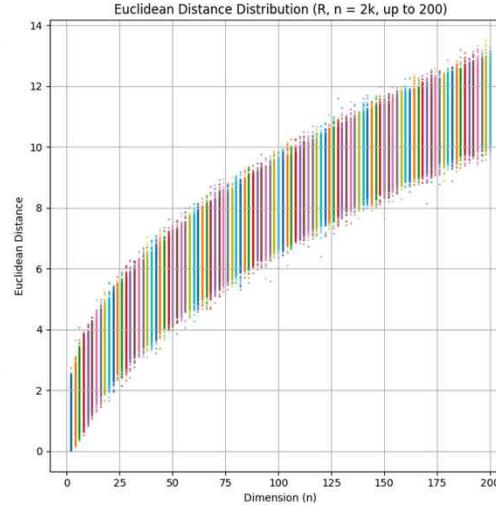
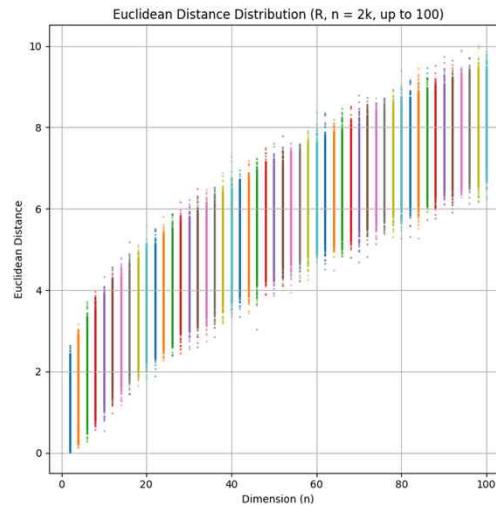
DIEM (Dimension Insensitive Euclidean Metric)

- Variance Graph of DIEM and Cosine Similarity (with respect to Dimension n)



DIEM (Dimension Insensitive Euclidean Metric)

- Variance Graph of DIEM and Euclidean Metric (with respect to Dimension n)



Experiment

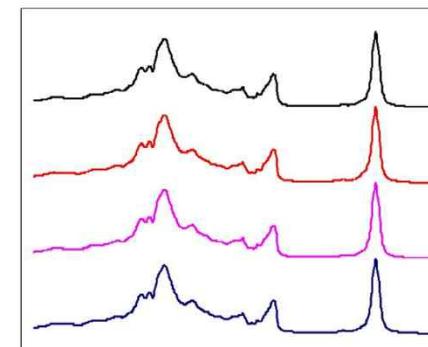
- **Experiment Description**
 - 실험 데이터 : 올리브 오일 FTIR 스펙트럼
 - FTIR : 오일의 화학적 특성 (570 가지)
 - 라벨 : 각 샘플의 생산 국가
 - 실험 목표 : DIEM, Cos-Sim, L2-norm 비교
 - 동일 라벨 간 유사도 측정
 - 라벨별 결과의 분포 분석

Dataset: OliveOil

Train Size	Test Size	Length	Number of Classes	Number of Dimensions	Type
30	30	570	4	1	SPECTRO

Data Source:	Link Here
Donated By:	K. Kemsley, A. Bagnall
Description:	Food spectrographs are used in chemometrics to classify food types, a task that has obvious applications in food safety and quality assurance. Each class of this data set is an extra virgin olive oil from alternative countries. Further information can be found in the original paper <i>FTIR spectroscopy and multivariate analysis can distinguish the geographic origin of extra virgin olive oils</i> . The data was first used in the time series classification literature in <i>Transformation Based Ensembles for Time Series Classification</i>

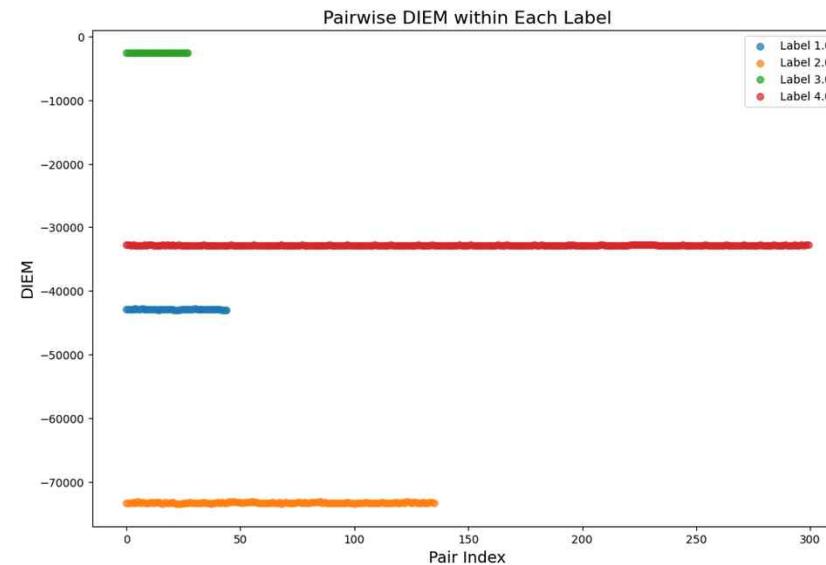
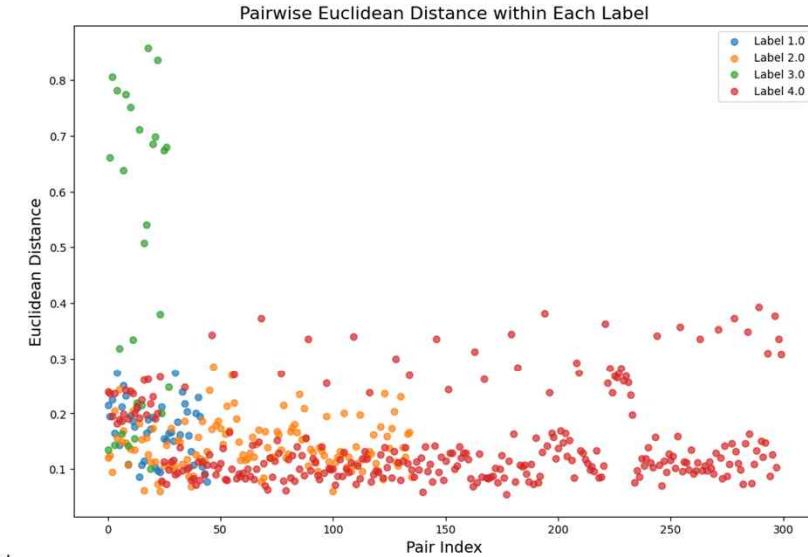
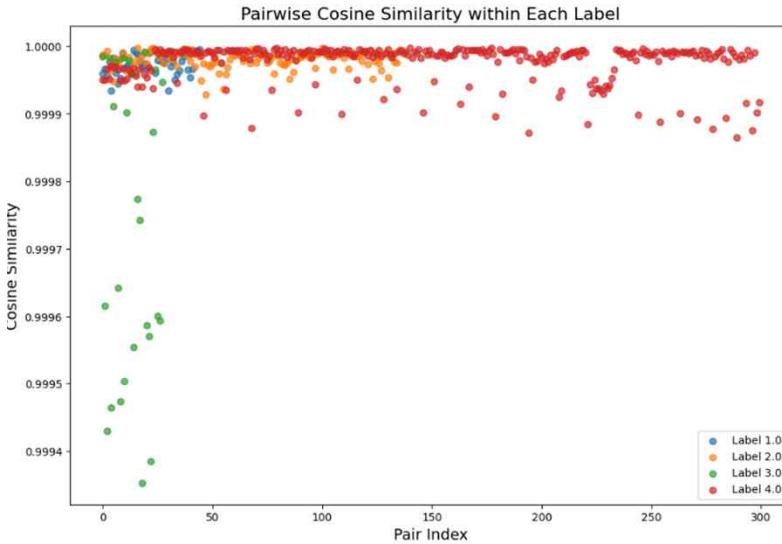
[Download this dataset](#)



Reference : <https://www.timeseriesclassification.com/dataset.php>

Experiment

- Measuring Similarity Among Same-Label Data in Olive-oil Dataset



Experiment

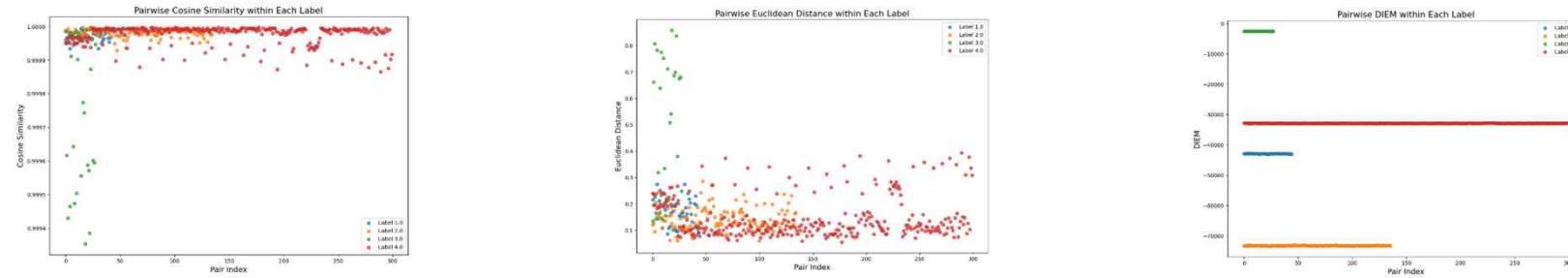
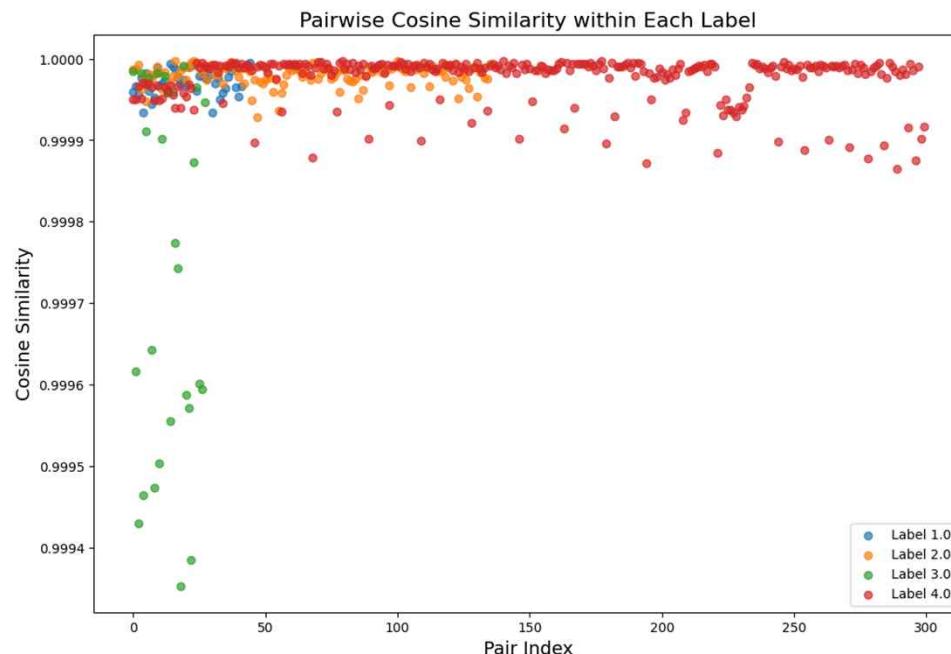


Figure 1: 코사인 유사도의 라벨별 분포

- 대부분의 값이 0.999 이상으로 집중
 - 모든 라벨에서 비슷한 분포
- 라벨 간 구분이 어려움



Experiment

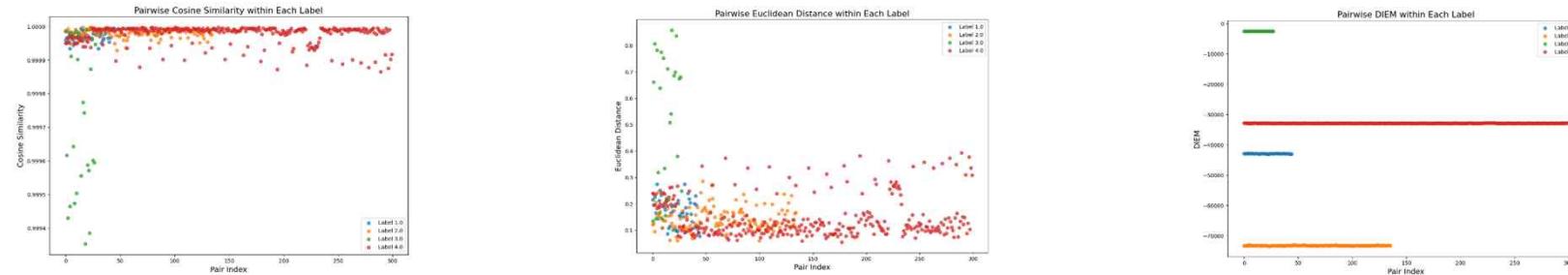
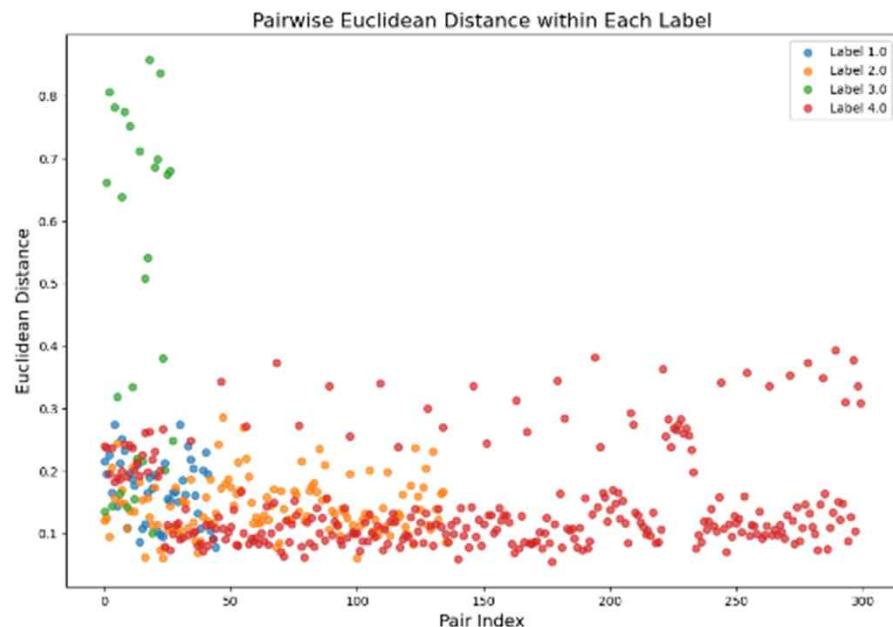


Figure 2: 유클리드 유사도의 라벨별 분포

- 대부분의 값이 0.999 이상으로 집중
 - 모든 라벨에서 비슷한 분포
- 라벨 간 구분이 어려움



Experiment

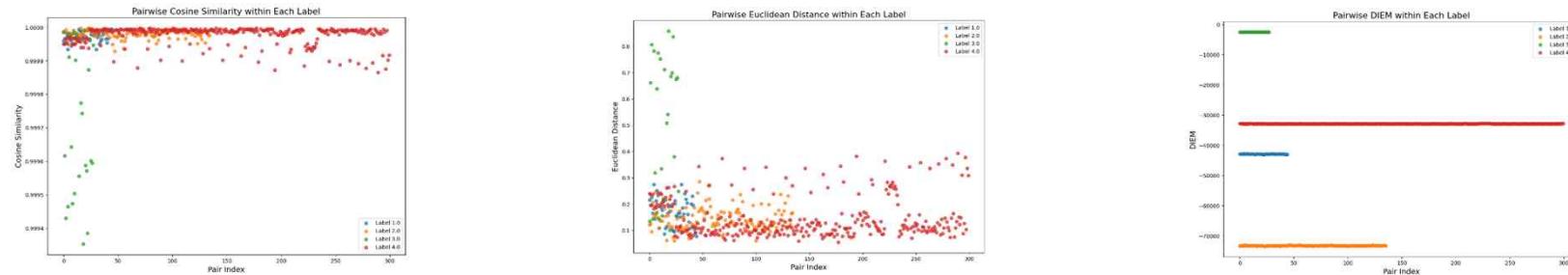
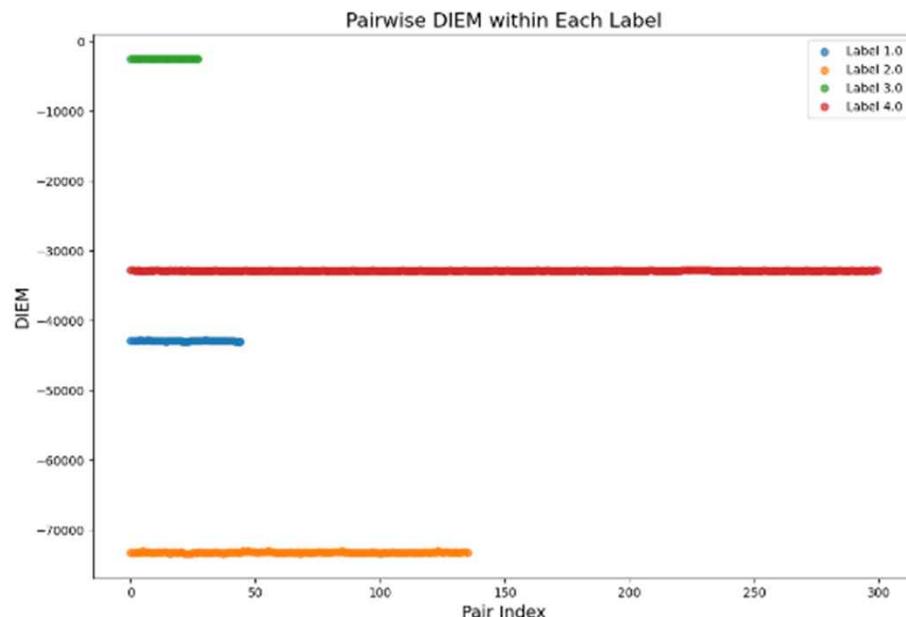


Figure 3: 코사인 유사도의 라벨별 분포

- 각 라벨의 DIEM 값이 명확히 구분
 - 다른 라벨 간 값이 겹치지 않음
- 라벨 간 차이를 효과적으로 반영



Research Topics

- Research Topics

1. L_p norm 분포 특성의 수학적 탐구

- L_p Norm의 p -값과 차원 증가에 따른 분포의 기대값과 분산 변화를 수학적으로 분석.
- 특정 p -값에서 분산이 일정하게 유지되는 조건을 도출하고, 이를 기반으로 다양한 데이터 유형에 적합한 유사성 메트릭 설계 가능성 탐색.

2. 차원 증가에도 분포가 유지되는 메트릭 개발

- 차원이 증가해도 분포가 안정적으로 유지되거나 기대값과 분산이 제어 가능한 새로운 메트릭 설계.
- 기존 메트릭(예: 유clidean 거리, 코사인 유사도)의 정규화와 중심화 기법을 확장하거나 확률 기반 메트릭 설계로 차원의 저주 극복.

목표: 고차원 데이터에서도 유사성과 차이를 신뢰성 있게 측정할 수 있는 새로운 메트릭의 이론적 기틀 마련 및 응용 확대.

