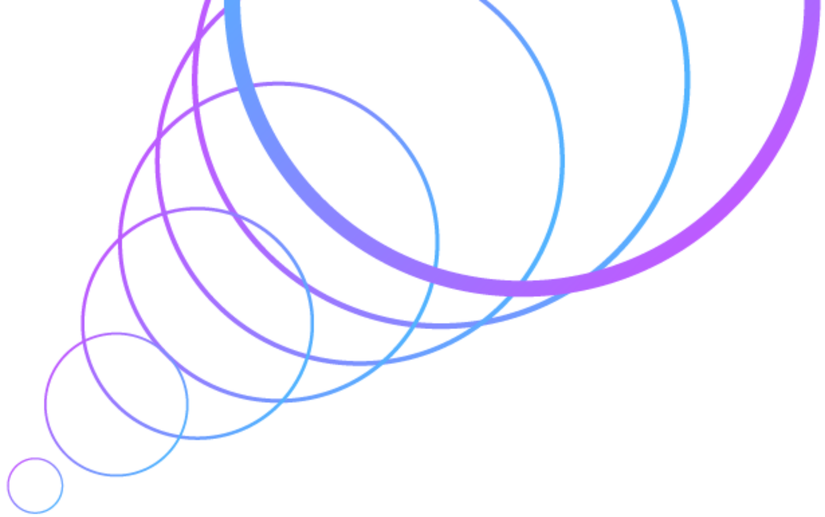
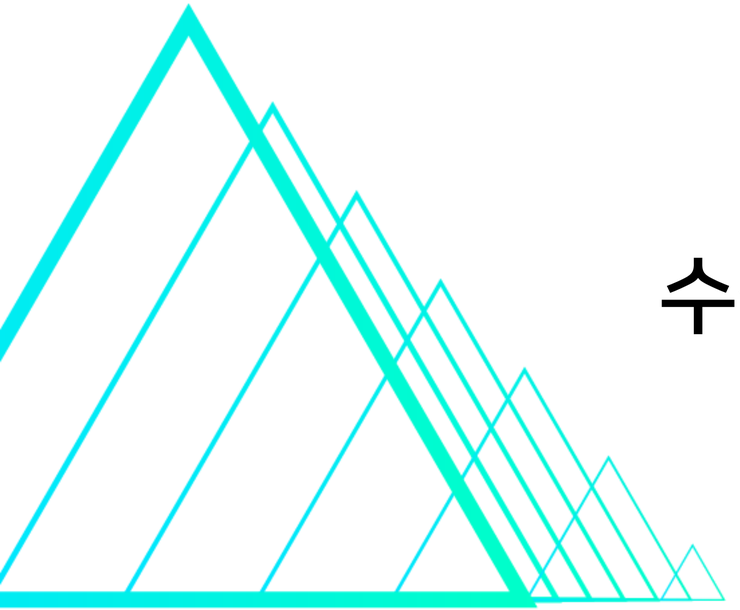
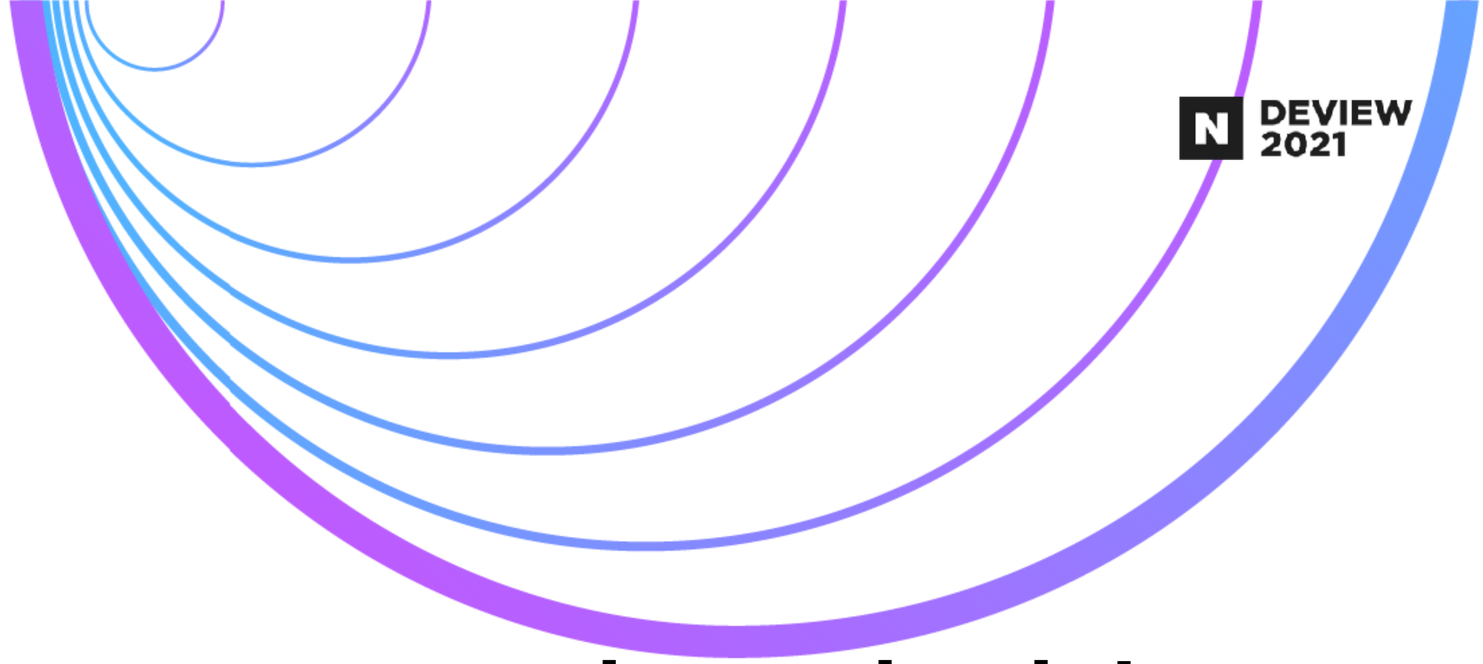




수십억 상품을 클러스터링 하는 방법

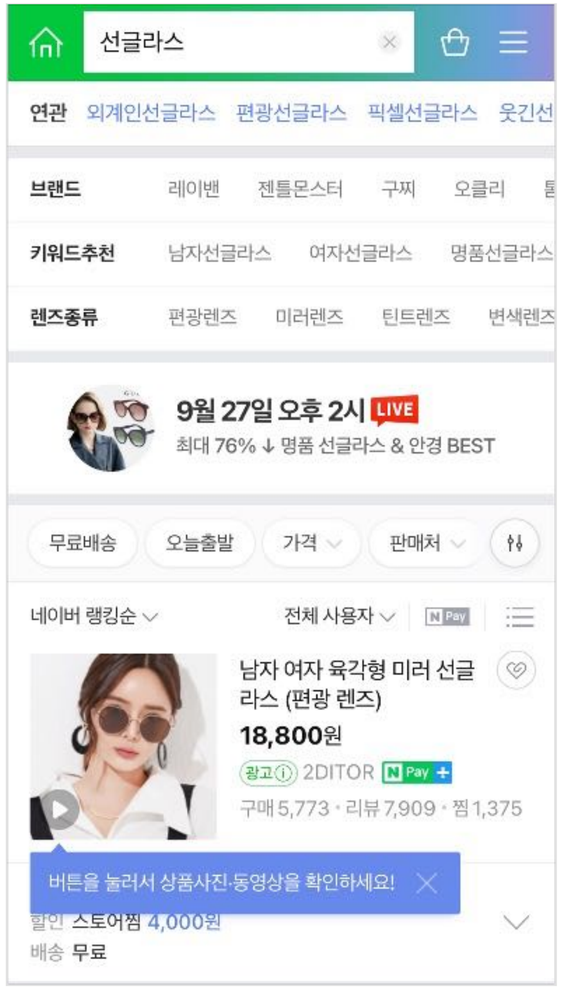
최승권 Catalog Matching, Shopping, NAVER Search CIC

CONTENTS



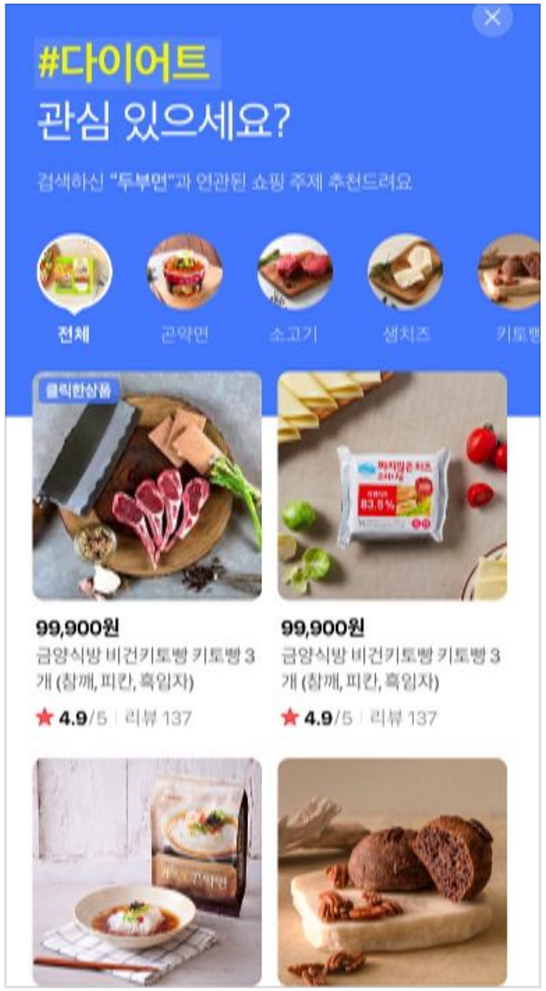
1. 네이버 Search CIC의 쇼핑 기술

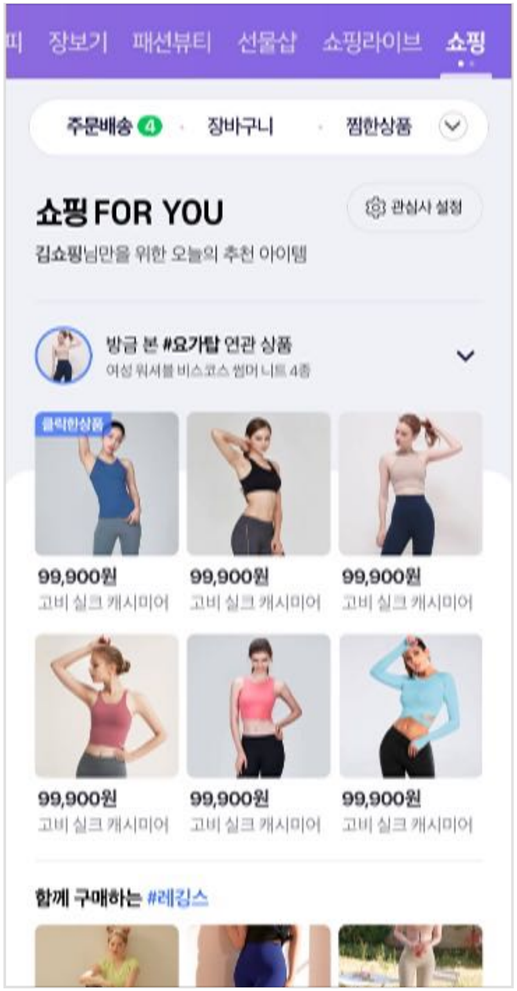
1.1 NAVER Search CIC의 e-Commerce 기술



상품 검색

이미지 검색





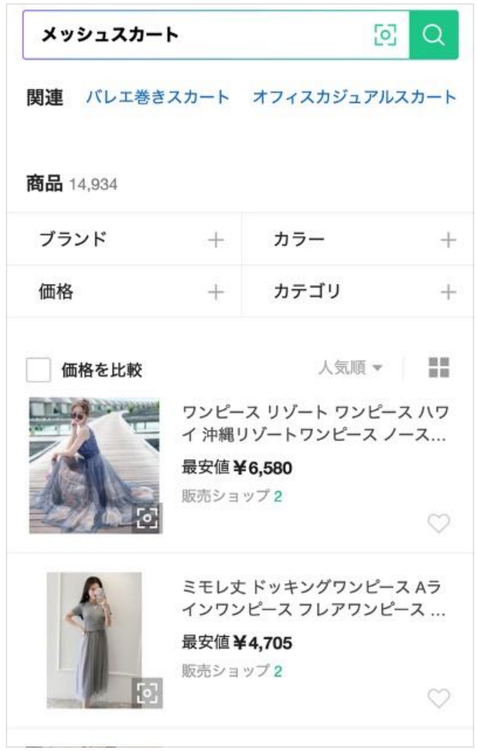
추천 (AiTEMS)

1.2 NAVER Search CIC의 글로벌 쇼핑 기술 지원

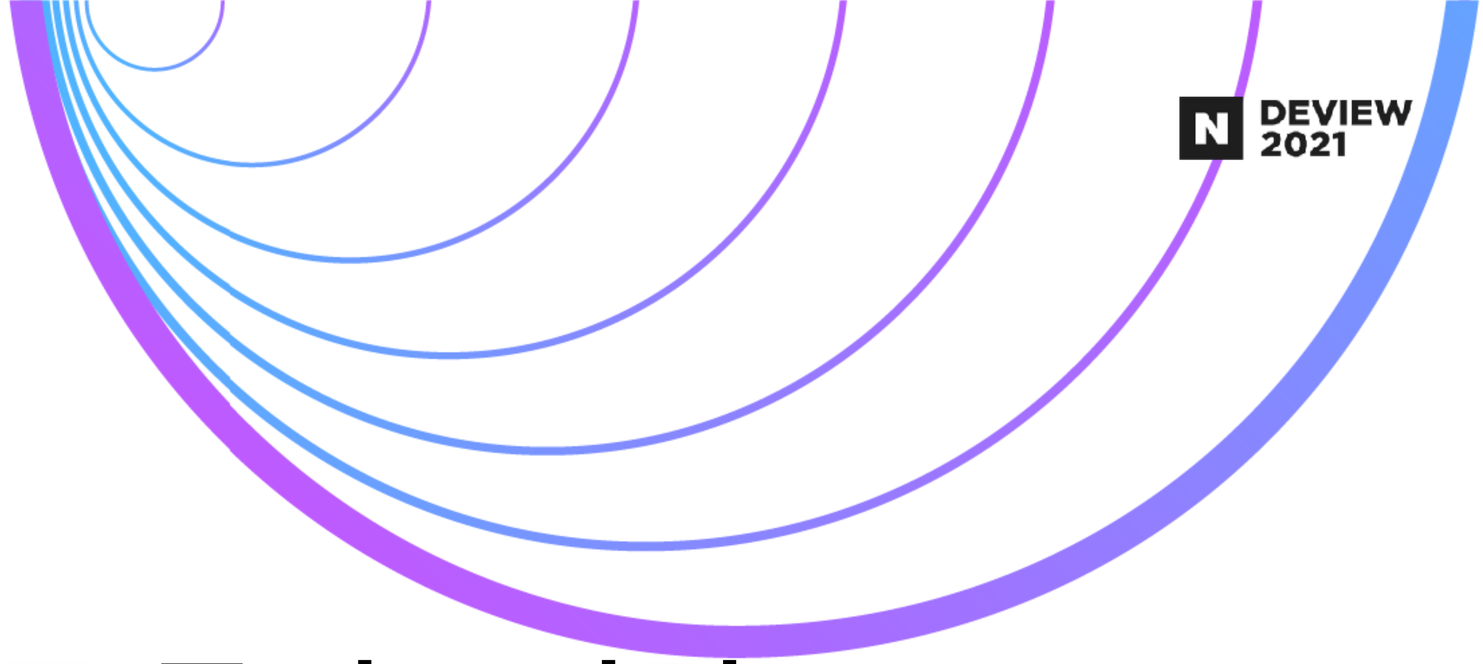
DEVIEW 2021

쇼핑

LINEショッピング

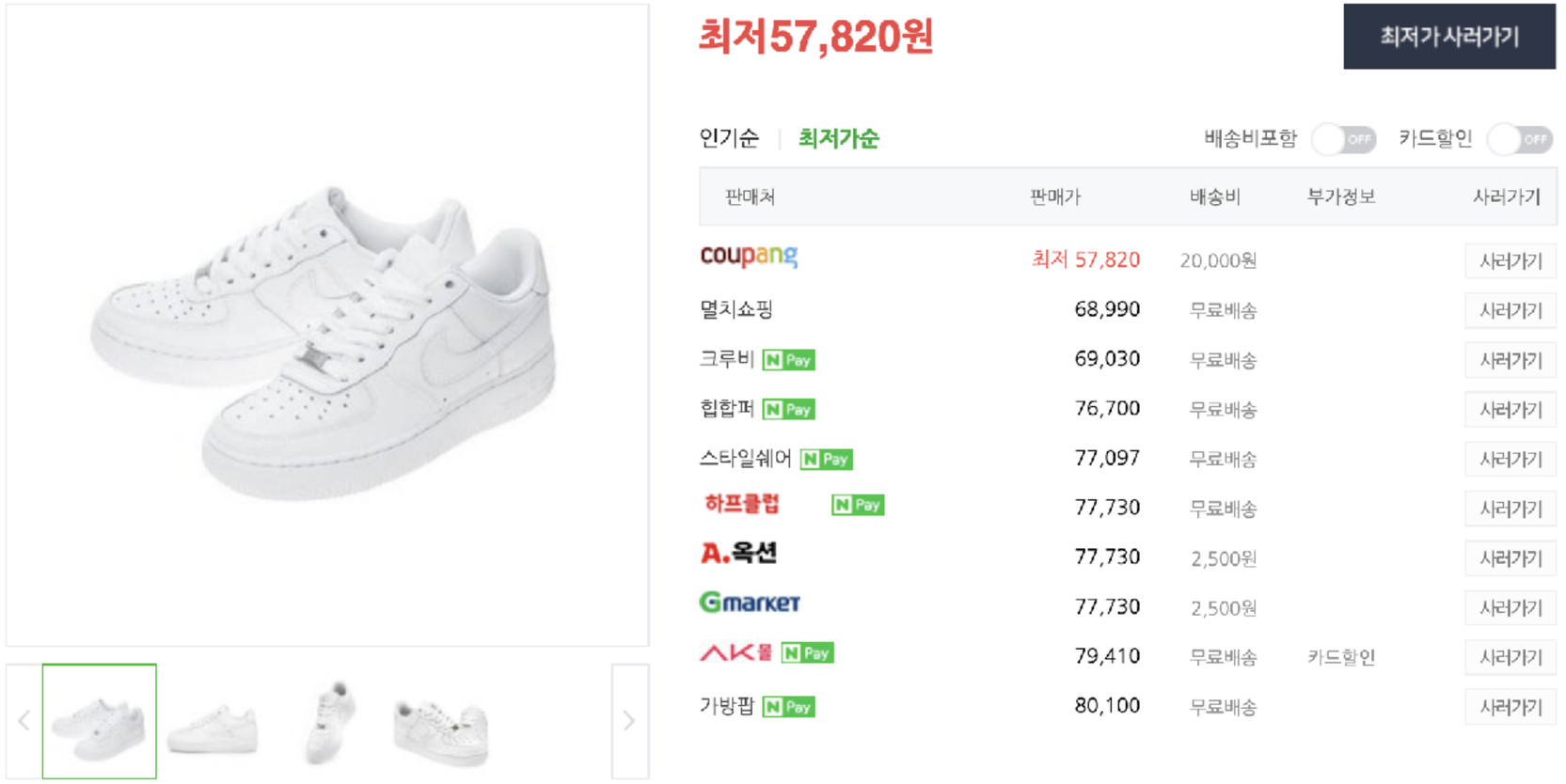


LINE 購物



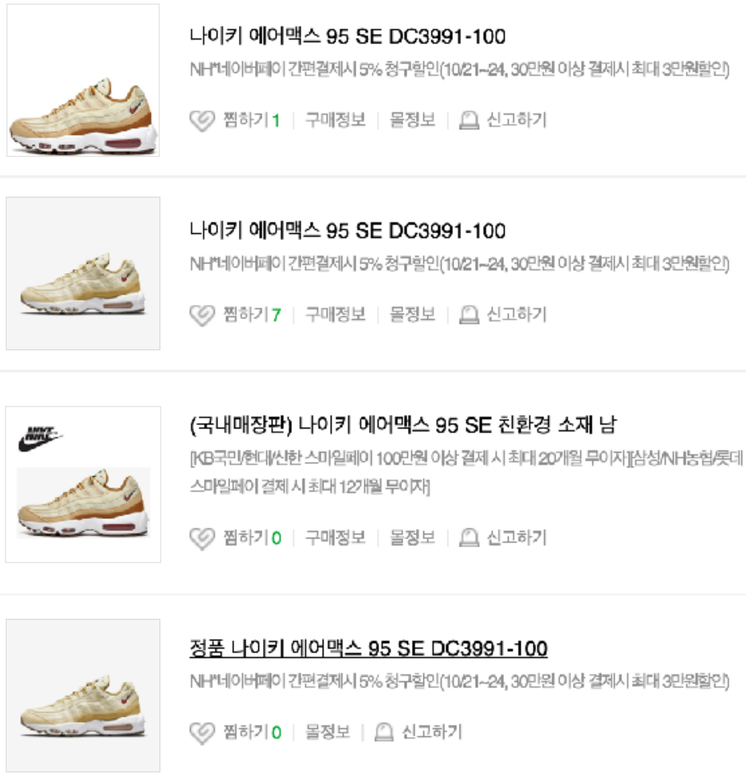
2. 상품 클러스터링

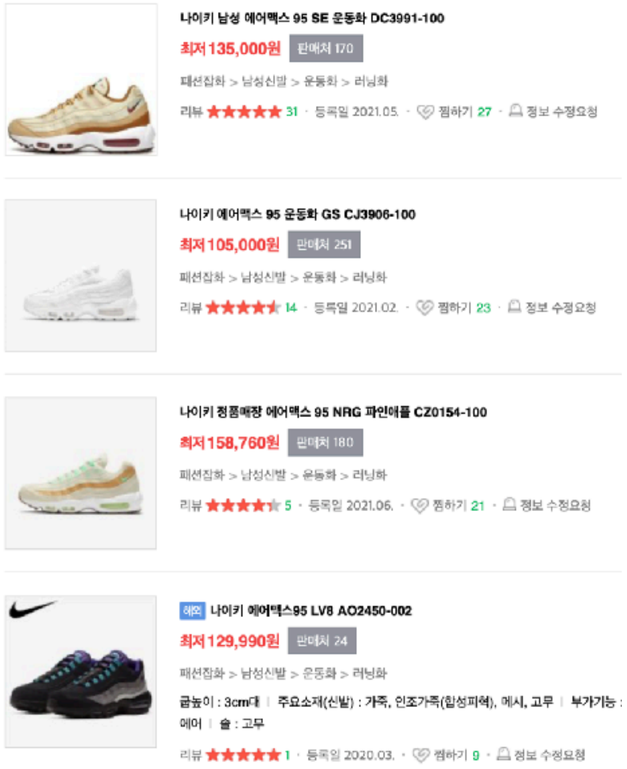
2.1 상품 클러스터링을 왜 해야 할까요?



상품 가격 비교

2.1 상품 클러스터링을 왜 해야 할까요?





중복 상품 제거

2.2 상품 클러스터링을 어렵게 하는 요인들

DEVIEW 2021

정보 부족

남자 운동화 조깅화 러닝화 워킹화 통풍 신발

정보 과잉

나이키 줌 플라이 에어 줌 페가수스 디비전 모음
남자 운동화 발편한 마라톤화 남성 조깅화

2.2 상품 클러스터링을 어렵게 하는 요인들

DEVIEW 2021

저품질 텍스트

[무료배송] [국내 매장판] 나이키 NIKE 에어 포스1 07 AIRFORCE 1 맨즈 우먼스 스포츠 운동화 런닝화 남여공용CW2288-111 DD8959-100트리플화이트 올흰올백 당일 발송

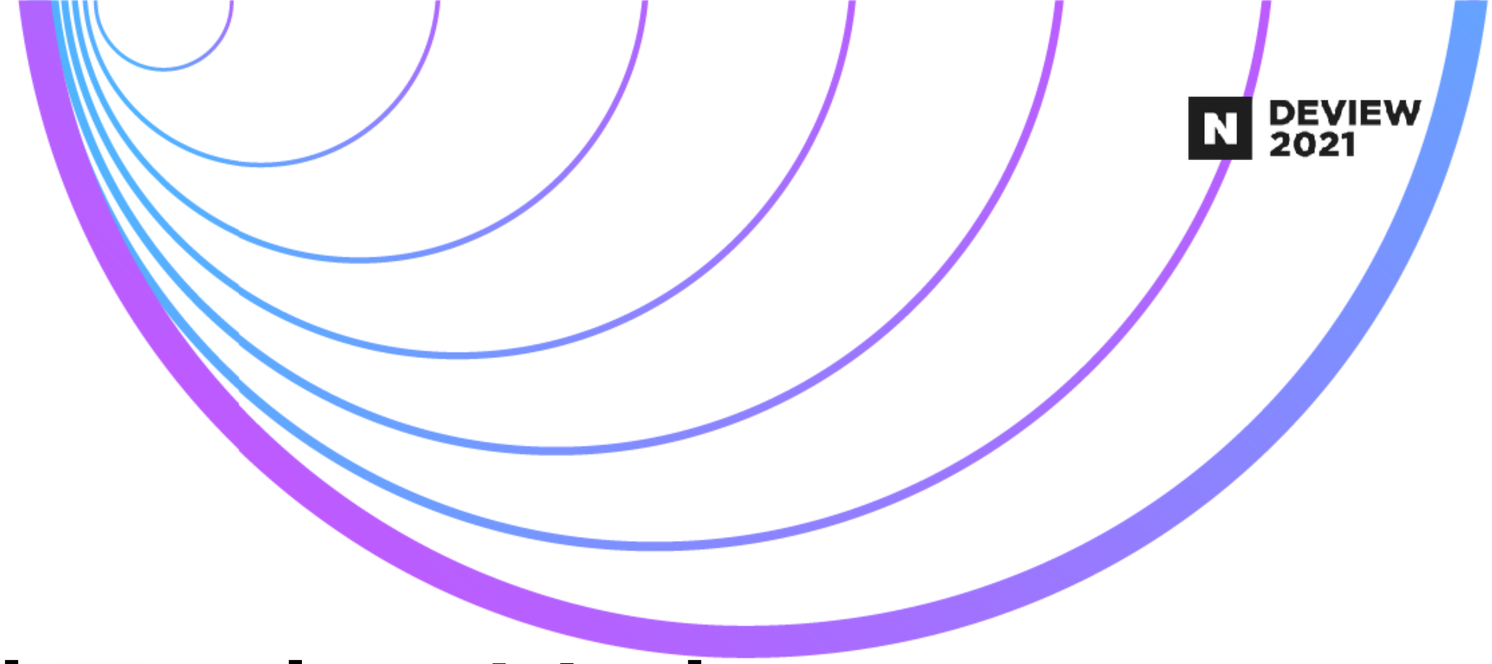
NO IMAGE

【新品】【本】かわいいね。ｋ．ｍ．ｐ．/ 絵・文・デザイン

이미지와 상품명이 다름

레드 블라우스 여성셔츠 상의 옥스포드 남방 LED-149

저품질 이미지



DEVIEW 2021

3. 상품 정보 분석

3.1 불규칙적인 상품 정보

상품 이름은 어디에 있나요?

[무료배송] [국내 매장판] 나이키 NIKE 에어포스1 07 AIRFORCE 1 맨즈 우먼스 스포츠 운동화 런닝화 남여공용 CW2288-111 DD8959-100트리플화이트 올흰올백 당일발송

3.1 불규칙적인 상품 정보

불용어

동의어 (브랜드)

[무료배송] [국내 매장판] 나이키 NIKE 에어포스1 07
AIRFORCE 1 맨즈 우먼스 스포츠 운동화 런닝화 남여공용
CW2288-111 DD8959-100트리플화이트 올흰올백 당일발송

불필요한 정보

상품 코드

나이키 에어포스1 07 CW2288-111

3.2 클러스터링에 부적합한 이미지를 가진 상품 검출

어두운 이미지

넓은 마진

콜라주

이미지 없음

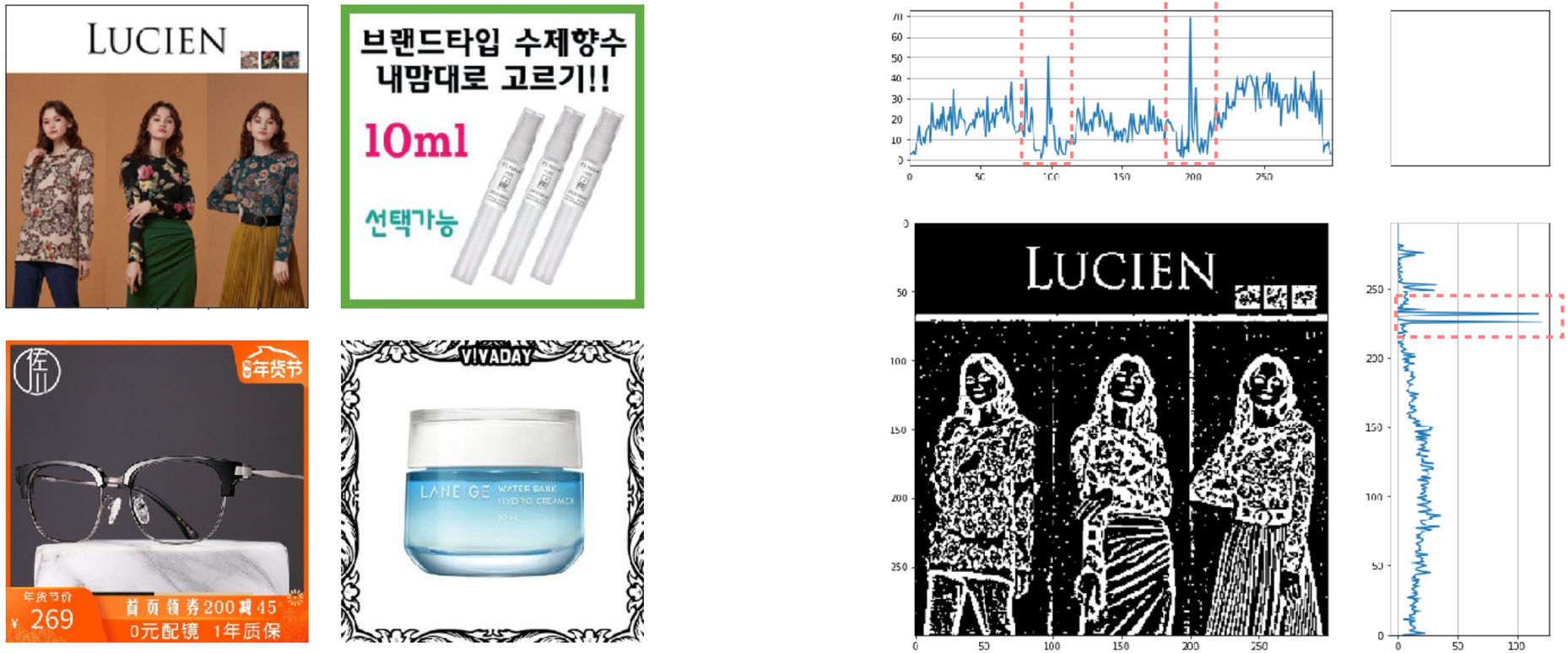
저해상도 이미지

다중 텍스트

멀티 오브젝트

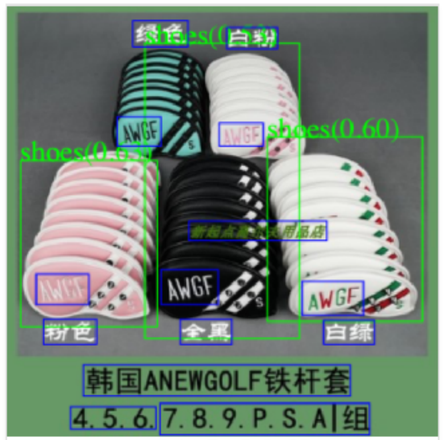
테두리

3.2 클러스터링에 부적합한 이미지를 가진 상품 검출



Collage / Border-line 검출 예시

3.2 클러스터링에 부적합한 이미지를 가진 상품 검출



스포츠/레저>골프>골프필드용품

object count: 3
text: 0.15
collage: 0.23
nuki: 0.12
border line: 0.73

생활/건강>관상어용품>수족관/어항

object count: 0
text: 0.45
collage: 0.28
nuki: 0.29
border line: 1.00

생활/건강>생활용품>세제/세정제

object count: 1
text: 0.20
collage: 0.41
nuki: 0.38
border line: 0.82

스포츠/레저>낚시>바다낚시

object count: 0
text: 0.07
collage: 0.61
nuki: 0.27
border line: 0.94

상품의 품질이 나쁨을 의미하지는 않음

상품 정보를 정제하는 것 만으로는 부족합니다

3.3 상품의 가격에 영향을 미치는 '구매 조건' 추출

[신세계몰](프리미엄 로사) 석류 원액 1개(0.5L) 8,800원

개수 표현 방식 차이

용량 표기의 차이

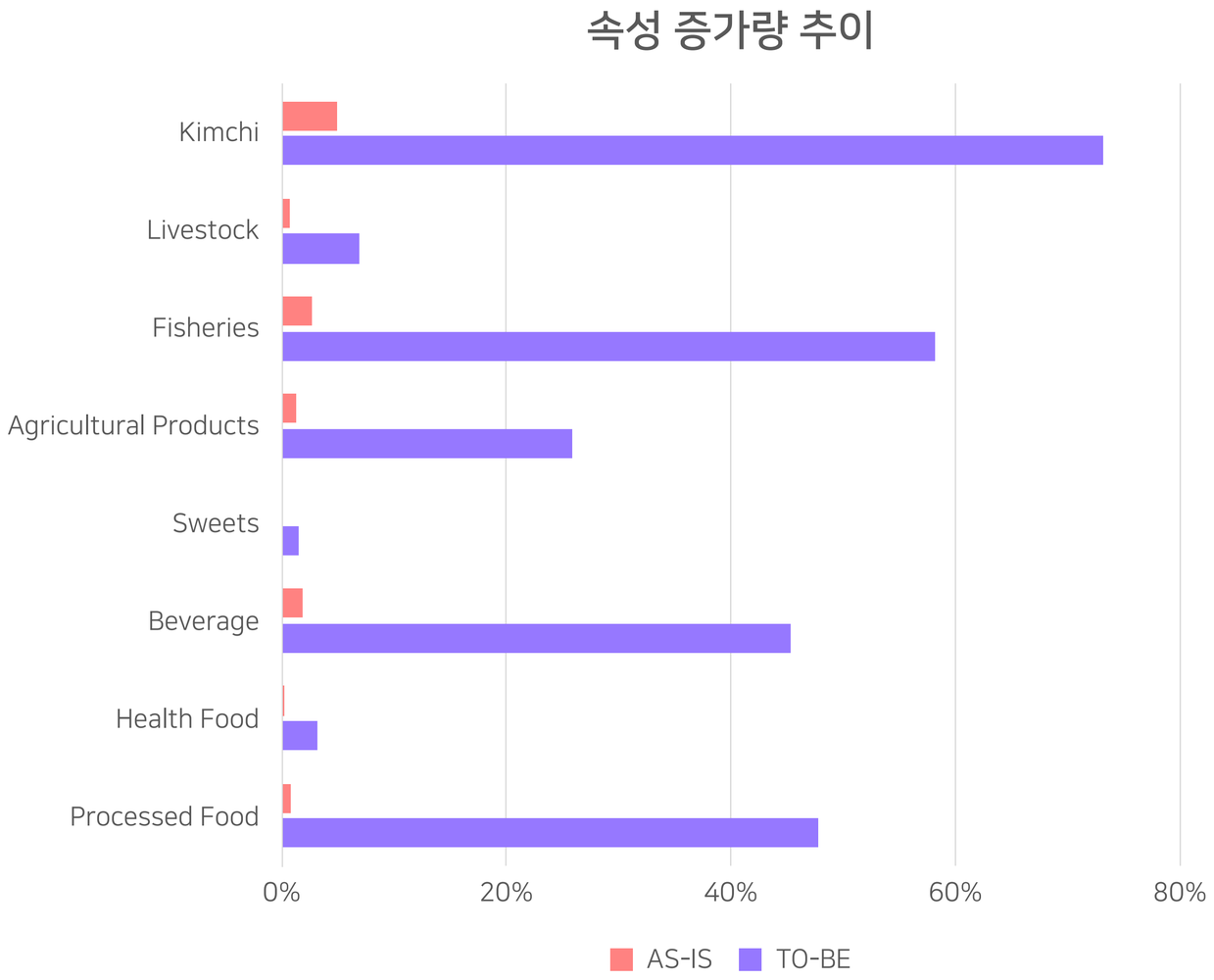
[프리미엄로사] 석류 원액 100% 3병X500ml 7,500원

개수 대비 싼 가격

3.3 상품의 가격에 영향을 미치는 '구매 조건'

DEVIEW 2021

용량	수량	타입
g	개입	Black pepper
kg	ea	Green pepper
ml	개	Red pepper
L	컵	Whole pepper
s	캔	Ground pepper
m	pcs	honeydew melon
l	병	Cantaloupe melon
xl	팩	Papaya
xxl	봉지	Button Mushrooms
cc	봉	Cremini Mushrooms
lb	번들	Portobello Mushrooms
mm	box	Oyster Mushrooms
...	...	...

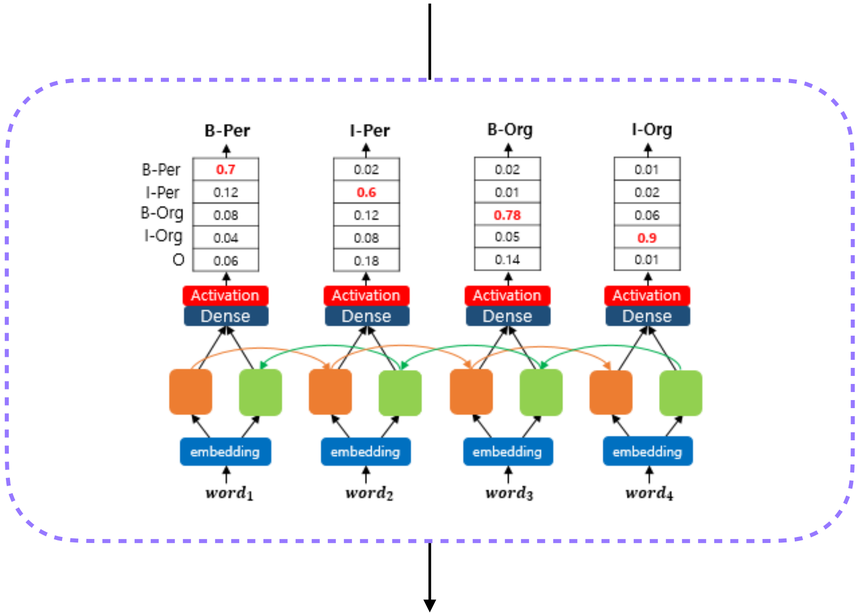


3.3 상품의 가격에 영향을 미치는 '구매 조건' 추출

DEVIEW 2021

Bidirectional LSTM + CRF

[Sunshine beautiful] Musk melon gift set (3kg/2pcs)



[Sunshine beautiful] Musk melon gift set (3kg/2pcs)

Seller Type Vol Qty

id 26230202464
categoryName Food>Agricultural Products>Fruits>Melon
productName Gyeongbuk Goryeong Papaya Melon 2kg (3~6 pieces) First shipment this year
price KRW 17310

attribute(tobe) QTY:6 , VOL:2kg , Variety: Papaya ,

id 25692850099
categoryName Food>Agricultural Products>Fruits>Melon
productName Seongju Cantaloupe Melon Cantaloupe Melon 5kg
price KRW 53020

attribute(tobe) VOL:5kg , Variety:Kantalof ,

id 26224812508
categoryName Food>Agricultural Products>Fruits>Melon
productName [Sunshine beautiful] Musk meron paulownia gift set (more than 3kg 2pcs
price KRW 58900

attribute(tobe) QTY:2 , VOL:3kg , Variety: Musk ,

567 종류의 자동 속성 추출

3.4 상품 정보 추출 - 브랜드, 상품 코드

DEVIEW 2021

Transformers XLM-R Token Classification

Product Name : 三栄水栓製作所 SANEI 多角穴ザルボ T22-25X15

Collection Pcode : t22-25x15

Extract Pcode : T22-25X15

Predictions

0.5 : T22-25X15

0.6 : T22-25X15

0.7 : T22-25X15

0.8 : T22-25X15

0.9 : T22-25X15

Ratio	三	栄	水	栓	製	作	所		S	A	N	E	I		多	角	穴	ザ	ル	ボ		T	2	2	-	2	5	X	1	5
jp_pcode_final	三	栄	水	栓	製	作	所		S	A	N	E	I		多	角	穴	ザ	ル	ボ		T	2	2	-	2	5	X	1	5

jp_pcode_final :

{'entity_group': 'B', 'score': 0.9511885046958923, 'word': 'T', 'start': 21, 'end': 22}

{'entity_group': 'B', 'score': 0.9543043375015259, 'word': '22', 'start': 22, 'end': 24}

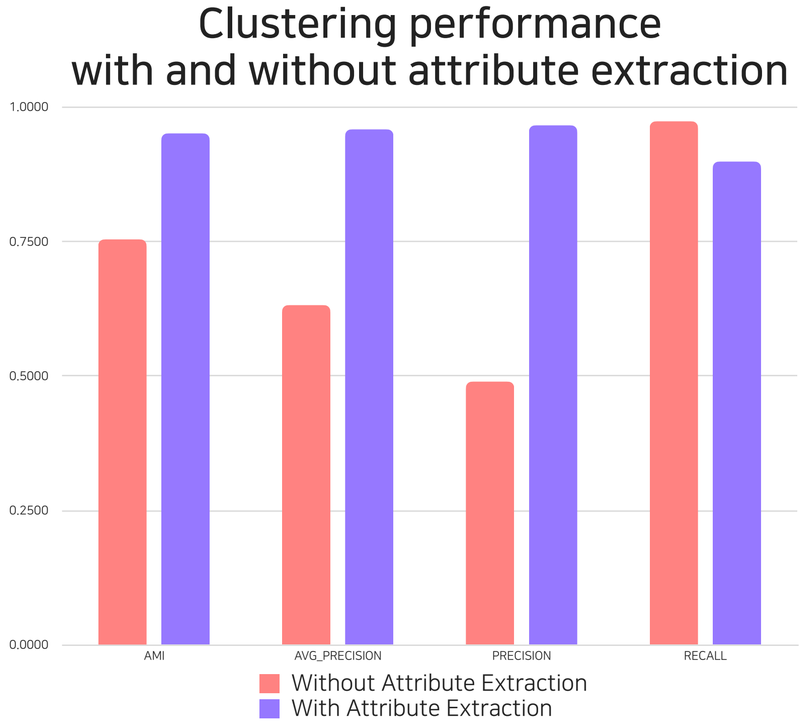
{'entity_group': 'B', 'score': 0.9561530947685242, 'word': '-25', 'start': 24, 'end': 27}

{'entity_group': 'B', 'score': 0.9560521245002747, 'word': 'X', 'start': 27, end': 28}

{'entity_group': 'B', 'score': 0.9582672119140625, 'word': '15', 'start': 28,'end': 30}

ratio : 모델 학습 시 학습데이터에 PCODE가 없는(= BIO 태깅이 안된)데이터의 비중

score > 0.9 score > 0.8 score > 0.7 score > 0.6 score > 0.5 score < 0.4



* AMI (Adjusted Mutual Information)

$$= \frac{MI(U, V) - E\{MI(U, V)\}}{max\{H(U), H(V)\} - E\{MI(U, V)\}}$$

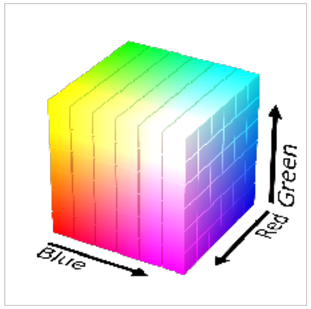
Ajinkya More, "Attribute Extraction from Product Titles in eCommerce." 2016.
Guineng Zheng, et al., "Open Tag: Open Attribute Value Extraction from Product Profiles" 2018.

3.5 상품 색상 추출

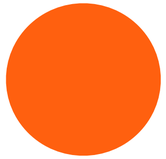
IMAGE

DETECTION & SEGMENTAION

COLOR QUANTIZATION



RESULT



3.5 상품 색상 추출

패션의류>여성의류>블라우스/셔츠

Object, Texture	Color
Shirts (0.84) ,None	→

패션의류>여성의류>티셔츠

Object, Texture	Color
Shirts (0.93) ,None	→
Shirts (0.75) ,None	→

패션의류>여성의류>점퍼

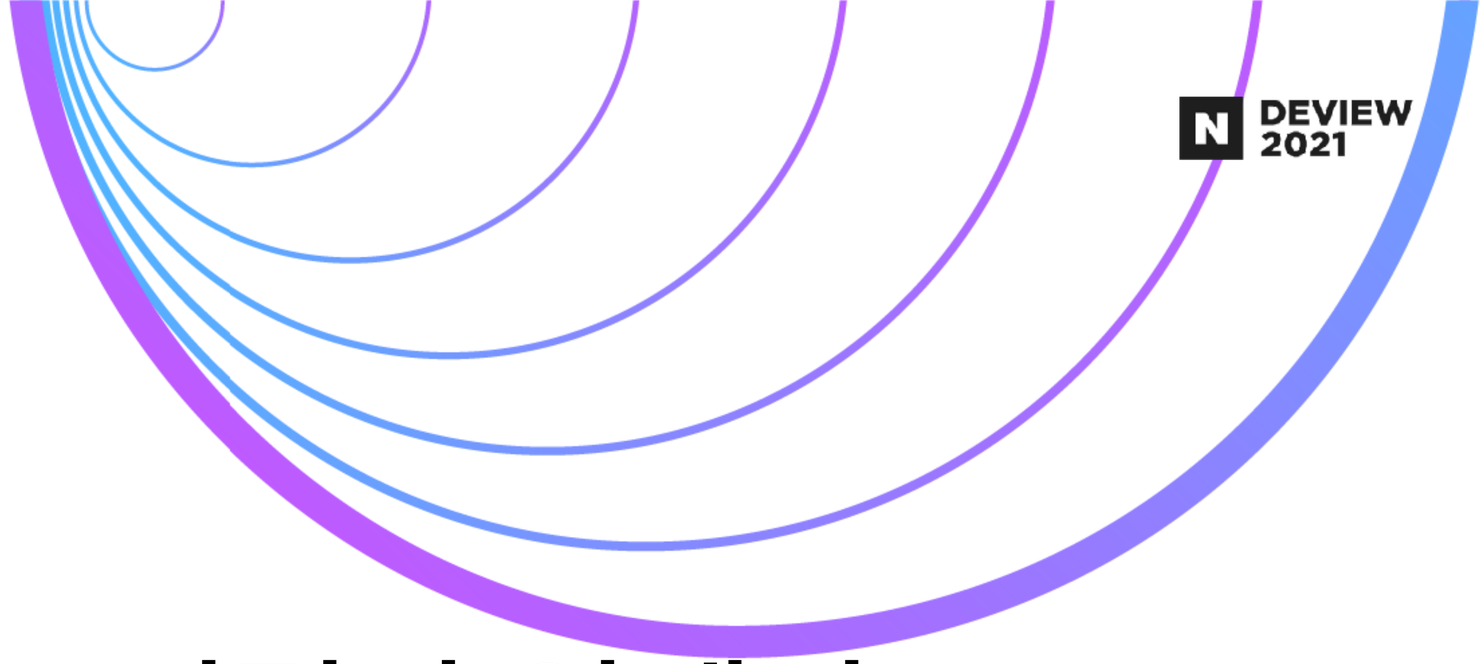
Object, Texture	Color
Coats (0.75) ,Stripe	→

패션의류>여성의류>조끼

Object, Texture	Color
Coats (0.91) ,None	→

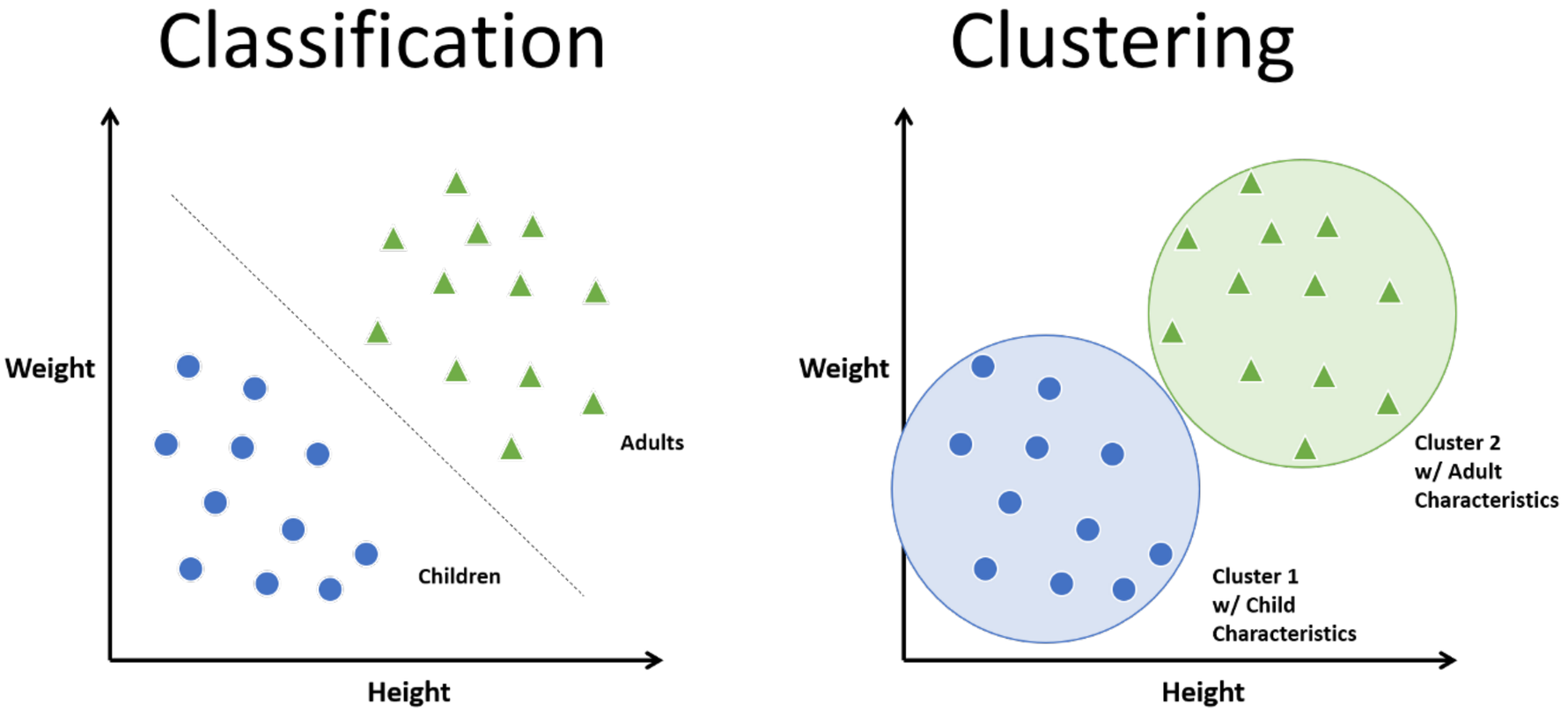
대 카테고리	중 카테고리	상품	색상 추출상품	비율
패션의류	여성의류	39,696	26,530	66.83%
	남성의류	20,303	15,129	74.52%
	Subtotal	59,999	41,659	69.43%
패션잡화	양말	2,344	1,361	58.06%
	여성신발	18,468	15,771	85.40%
	남성신발	10,231	9,276	90.67%
	여성가방	9,975	7,549	75.68%
	남성가방	4,714	4,122	87.44%
	이행용가방/스몰	931	660	70.89%
	지갑	2,767	2,134	77.12%
	모자	4,724	2,761	58.45%
	패션소품	2,312	1,322	57.18%
	시계	3,529	2,569	72.80%
	Subtotal	59,995	47,525	79.21%
출산/육아	임부복	978	624	63.80%
	신생아의류	3,030	1,959	64.65%
	유아동의류	37,112	26,416	71.18%
	유아동잡화	18,784	12,611	75.23%
	수영복/용품	2,112	1,258	59.56%
	Subtotal	59,996	42,868	71.46%
스포츠/레저	등산	21,214	17,937	84.55%
	골프	15,324	12,550	81.90%
	수영	23,461	14,731	62.79%
	Subtotal	59,999	45,218	75.36%
Total		239,989	177,270	73.87%

네이버 쇼핑 의류 카테고리 54색으로 확장 예정



4. 클러스터링과 임베딩

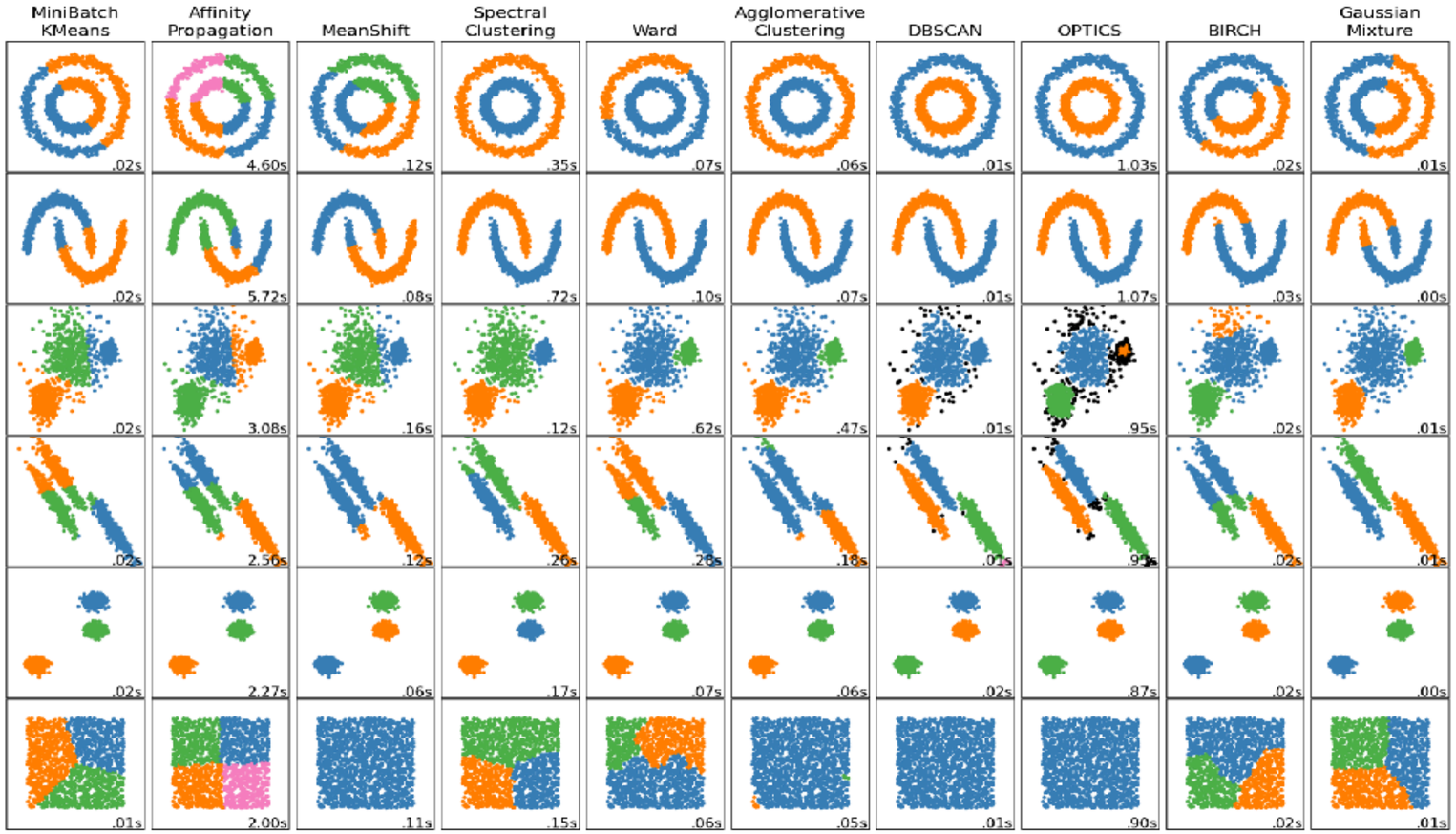
4.1 Clustering이란?



https://kevin-c-lee26.medium.com/machine-learning-101-classification-vs-clustering-e11b12c71243

4.2 Clustering methods

DEVIEW 2021



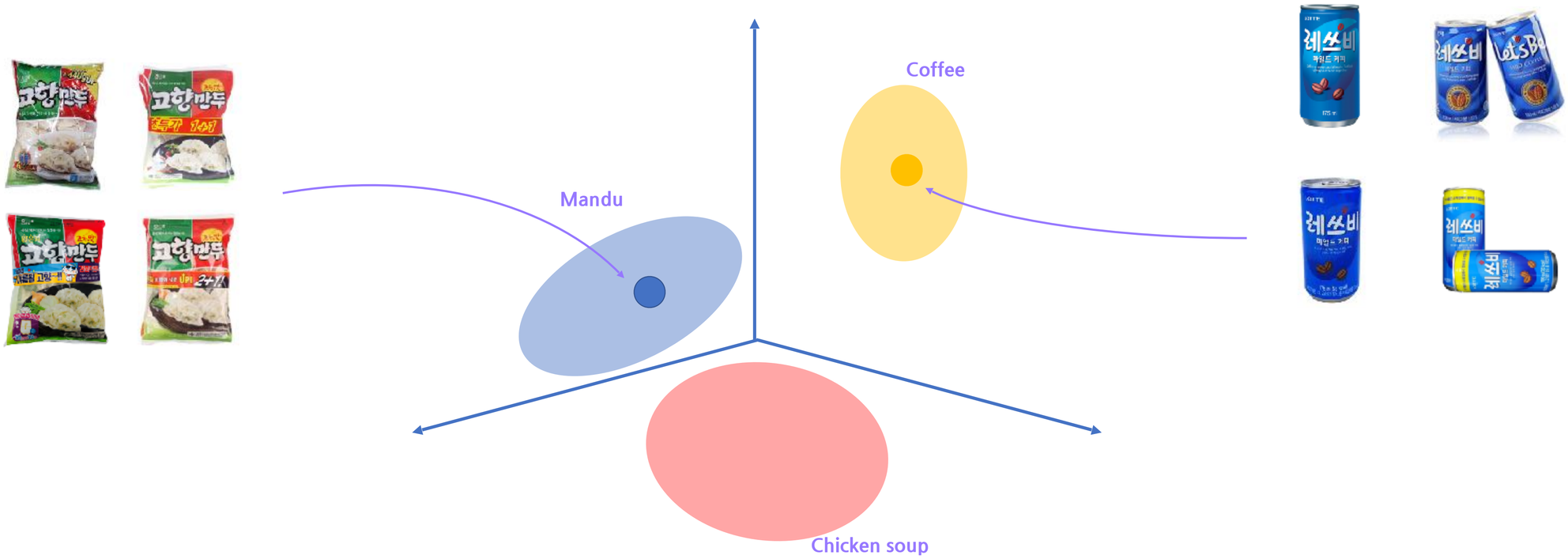
A comparison of the clustering algorithms in scikit-learn

https://scikit-learn.org/stable/modules/clustering.html

4.3 상품 정보를 벡터 공간에 임베딩

DEVIEW 2021

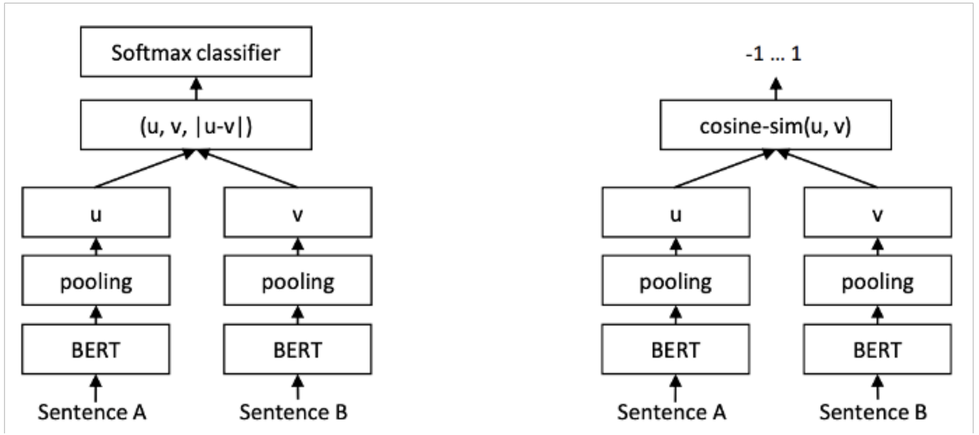
Everything but the kitchen sink

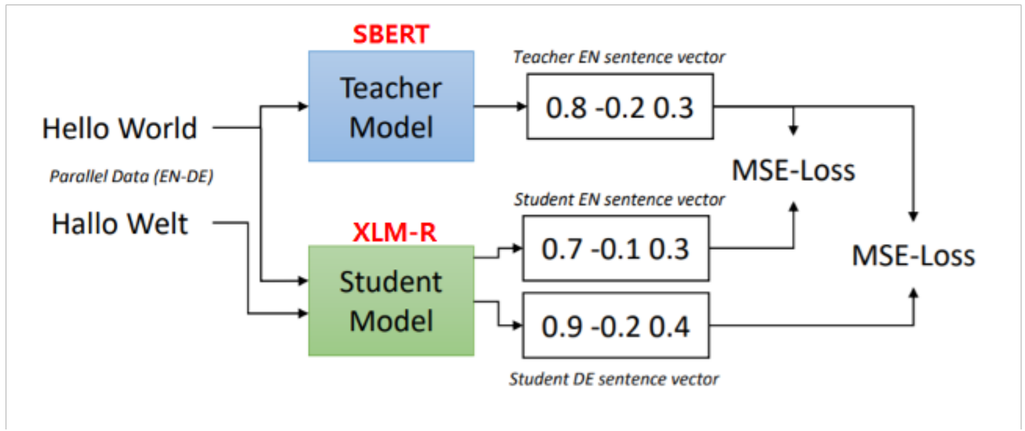


4.4 Sentence-BERT

DEVIEW 2021

Sentence-BERT & XLM-R (cross-lingual)





Multilingual Universal Sentence Encoder

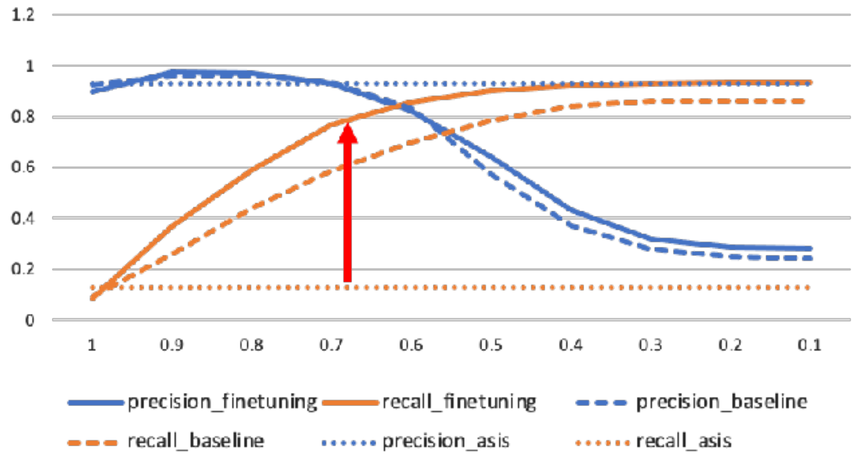
Nils Reimers and Iryna Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." 2019.
Nils Reimers and Iryna Gurevych, "Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation." 2020.

4.4 Sentence-BERT

DEVIEW 2021

Embedding Clustering Results

Sentence-BERT & XLM-R (cross-lingual)



Cosine Similarity	Product Name
1.0000001	아르떼 후드 트렌치 코트 DLCPWU060
0.91460687	아르떼 W몰 아르떼 후드 트렌치 코트 DLCPWU060
0.9446484	W몰 아르떼 후드 트렌치 코트 DLCPWU060_1
0.8338849	[삼성카드 최대 7% 할인][신세계몰]아르떼 W몰 아르떼
0.8654208	[롯데아이몰][아르떼] W몰 후드 트렌치 코트 DLCPWU(
0.87161934	[삼성카드 최대 7% 할인][패션플러스]W몰 아르떼 후드 (DLCPWU060_1)
0.12797707	남자 와이드 루즈핏 후드 캐쥬얼 코트 BF 드랜드 10160(
0.10409761	[삼성카드 최대 7% 할인][패션플러스]DB_ 올리브데올레 (OW0WH619) (166908236)

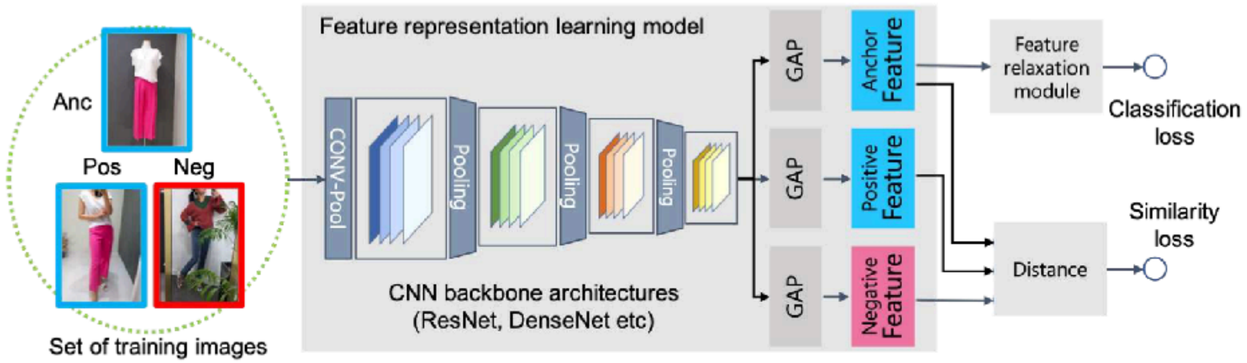
Recall Improved by 47%

Nils Reimers and Iryna Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.", In Proc. of EMNLP 2019.
Nils Reimers and Iryna Gurevych, "Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation." 2020.

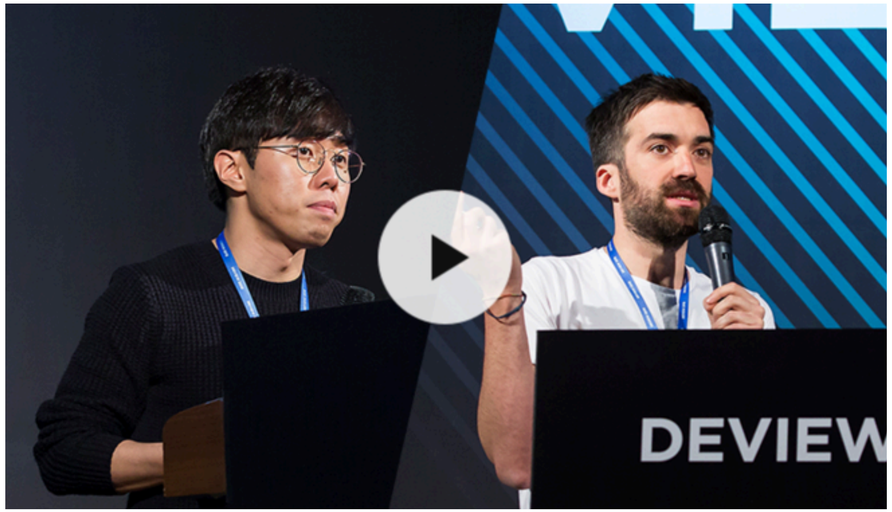
4.5 Deep feature extraction from image

DEVIEW 2021

Product image feature embedding



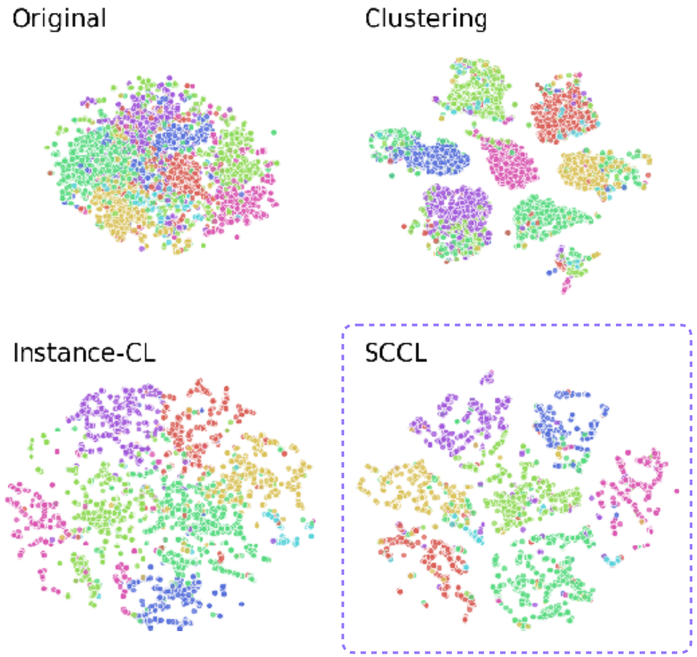
Fashion image retrieval (FIR)

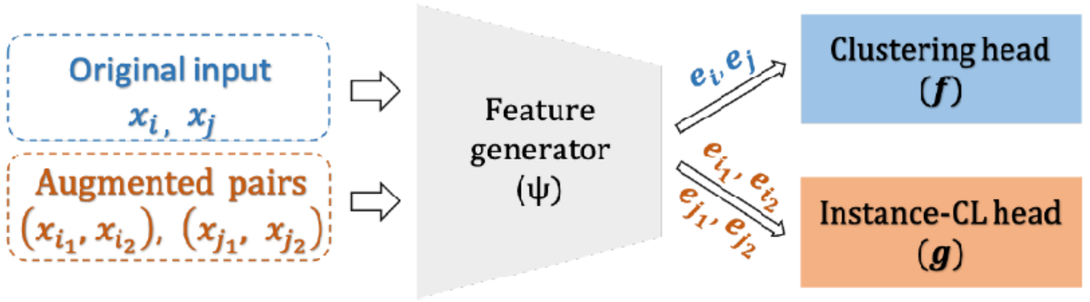


Deview 2018 - Fashion Visual Search

Albert Gordo et al., "End-to-end Learning of Deep Visual Representations for Image Retrieval." 2017.
Sanghyuk Park et al., “Study on Fashion Image Retrieval Methods for Efficient Fashion Visual Search.” 2019

4.6 Multimodal (text + image)





$$\mathcal{L} = \mathcal{L}_{\text{Instance-CL}} + \mathcal{L}_{\text{Cluster}} = \sum_{j=1}^{M} \ell_j^C / M + \eta \sum_{i=1}^{2M} \ell_i^I / 2M . \quad (7)$$

Batch size

$$\tilde{z}_j = g(\psi(\tilde{x}_j)), j = i^1, i^2.$$

i1 과 i2 는 최대한 가깝게

$$\ell_{i^1}^I = -\log \frac{\exp(\text{sim}(\tilde{z}_{i^1}, \tilde{z}_{i^2})/\tau)}{\sum_{j=1}^{2M} \mathbb{1}_{j \neq i^1} \cdot \exp(\text{sim}(\tilde{z}_{i^1}, \tilde{z}_j)/\tau)} \quad (1)$$

i1 과 다른 augmentation 과의 거리는 최대한 멀리

P 와 Q 의 발산 최적화
→ 각 클러스터의 Centroid 를 근사하고, auxiliary 보조를 활용하여 반복적으로 e 수정

$$\ell_j^C = \text{KL}[p_j \| q_j] = \sum_{k=1}^{K} p_{jk} \log \frac{p_{jk}}{q_{jk}} . \quad (5)$$

$$f_k = \sum_{j=1}^{M} q_{jk}, k = 1, \ldots, K$$

Batch 내에서의 클러스터 k의 빈도수

$$p_{jk} = \frac{q_{jk}^2 / f_k}{\sum_{k'} q_{jk'}^2 / f_{k'}} . \quad (4)$$

auxiliary probability
제곱을 사용하여 희귀 확률 값(q) 극대화한 후 클러스터 빈도로 정규화
→ 높은 confidence 값 학습, 클러스터 불균형으로 인해 발생하는 bias 완화

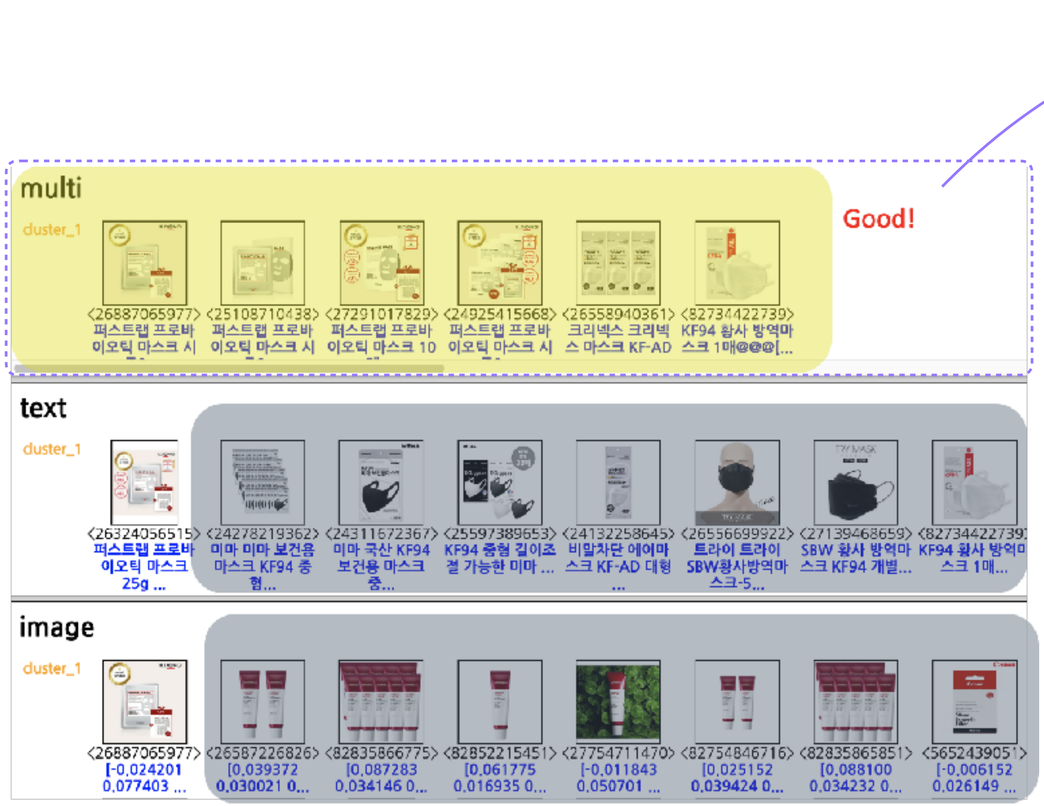
$$q_{jk} = \frac{(1 + \|e_j - \mu_k\|_2^2/\alpha)^{-\frac{\alpha+1}{2}}}{\sum_{k'=1}^{K} (1 + \|e_j - \mu_{k'}\|_2^2/\alpha)^{-\frac{\alpha+1}{2}}} . \quad (3)$$

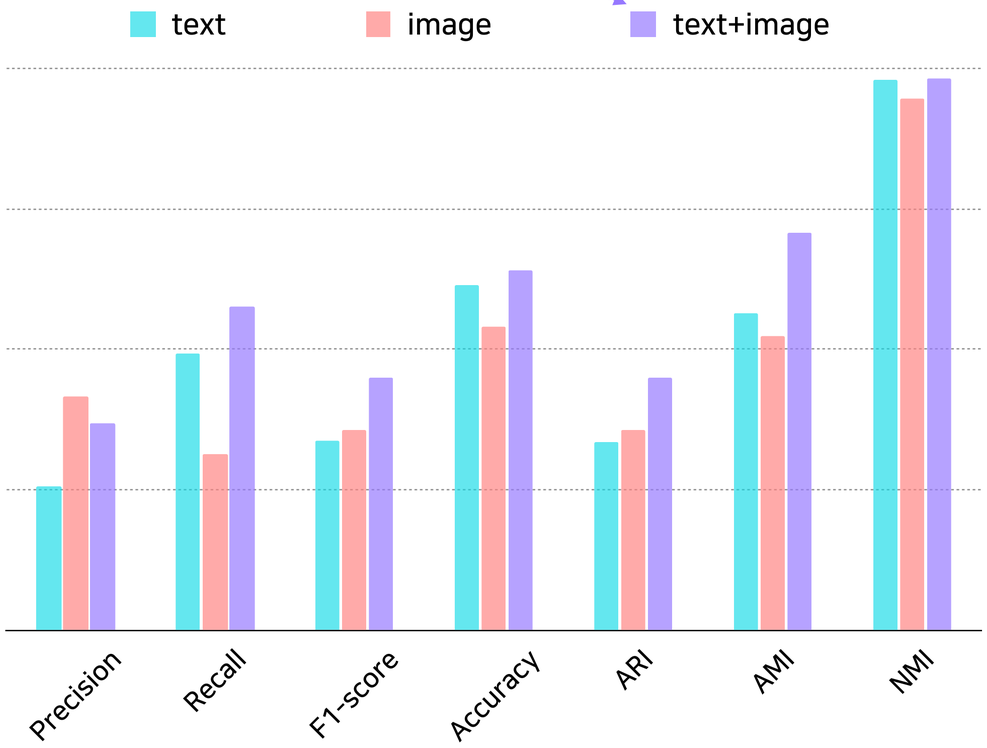
xj 가 클러스터 K에 할당될 확률
선제 클러스터의 centroid 와의 거리함에 따라 얼마나 가장 가까운지 확인

Supporting Clustering with Contrastive Learning

Dejiao Zhang et al., "Supporting Clustering with Contrastive Learning." 2021.

4.6 Multimodal (text + image)





Supporting Clustering with Contrastive Learning

Dejiao Zhang et al., "Supporting Clustering with Contrastive Learning." 2021.

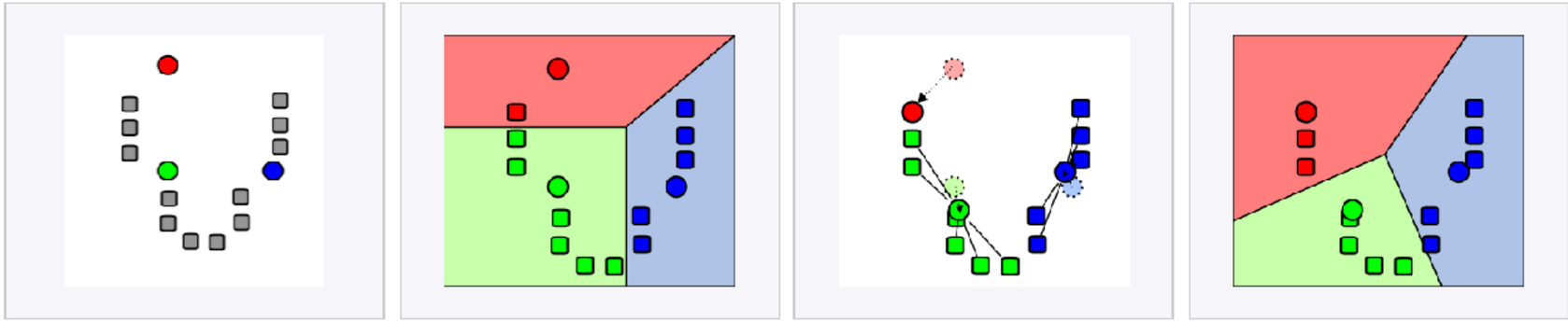
5. 대표적인 클러스터링 기법

5.1 k-Means Clustering

n개의 d-차원 데이터 오브젝트 ($\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$) 집합이 주어졌을 때,
k-평균 알고리즘은 n개의 데이터 오브젝트들을 각 집합 내 오브젝트 간 응집도를 최대로 하는
$k(\leq n)$개의 집합 $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ 으로 분할한다. 다시 말해, $\boldsymbol{\mu}_i$가 집합 S_i의 중심점이라 할 때

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \| \mathbf{x} - \boldsymbol{\mu}_i \|^2$$

각 집합별 중심점~집합 내 오브젝트간 거리의 제곱합을 최소로 하는 집합 **S**를 찾는 것이 이 알고리즘의 목표다.



1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown

2. *k* clusters are created by associating every observation with the nearest mean. The

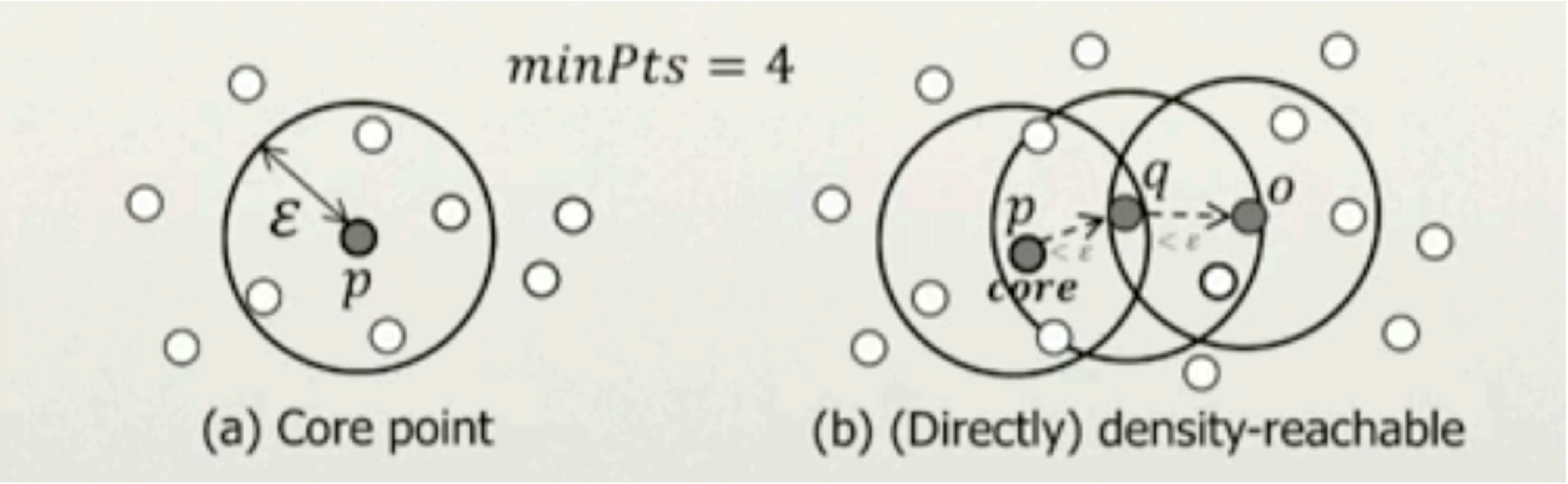
3. The centroid of each of the *k* clusters becomes the new mean.

4. Steps 2 and 3 are repeated until convergence has been reached.

https://ko.wikipedia.org/wiki/K-%ED%8F%89%EA%B7%A0_%EC%95%8C%EA%B3%A0%EB%A6%AC%EC%A6%98

5.2 DBSCAN Clustering

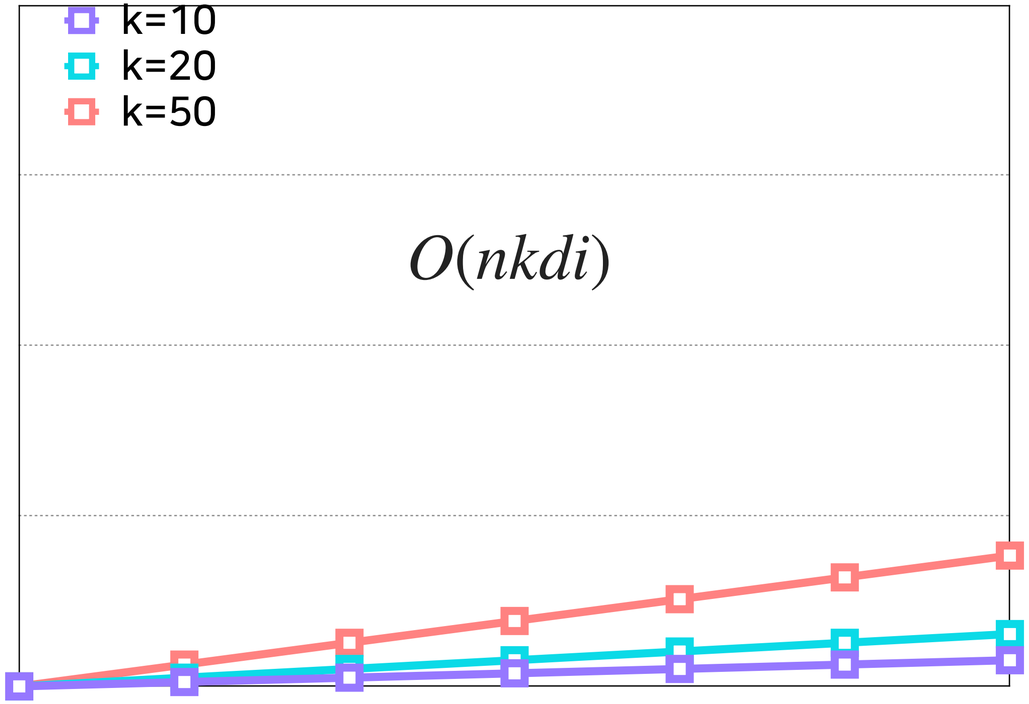
- Two algorithm parameters: (ε, $minPts$)
- Captures arbitrary shape of clusters and does not require the number of clusters in advance
- Finds dense regions and expands them in order to form clusters



https://www.slideshare.net/NaverEngineering/clustering-120426179

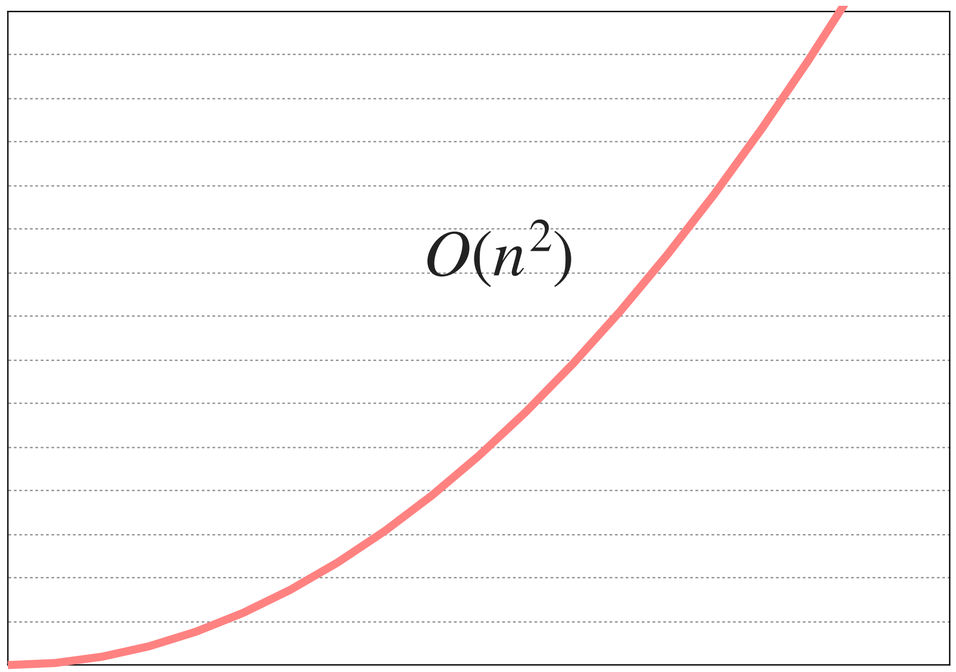
5.3 클러스터링 알고리즘의 시간 복잡도

k-Means



Linear Polynomial

DBSCAN



Quadratic Polynomial

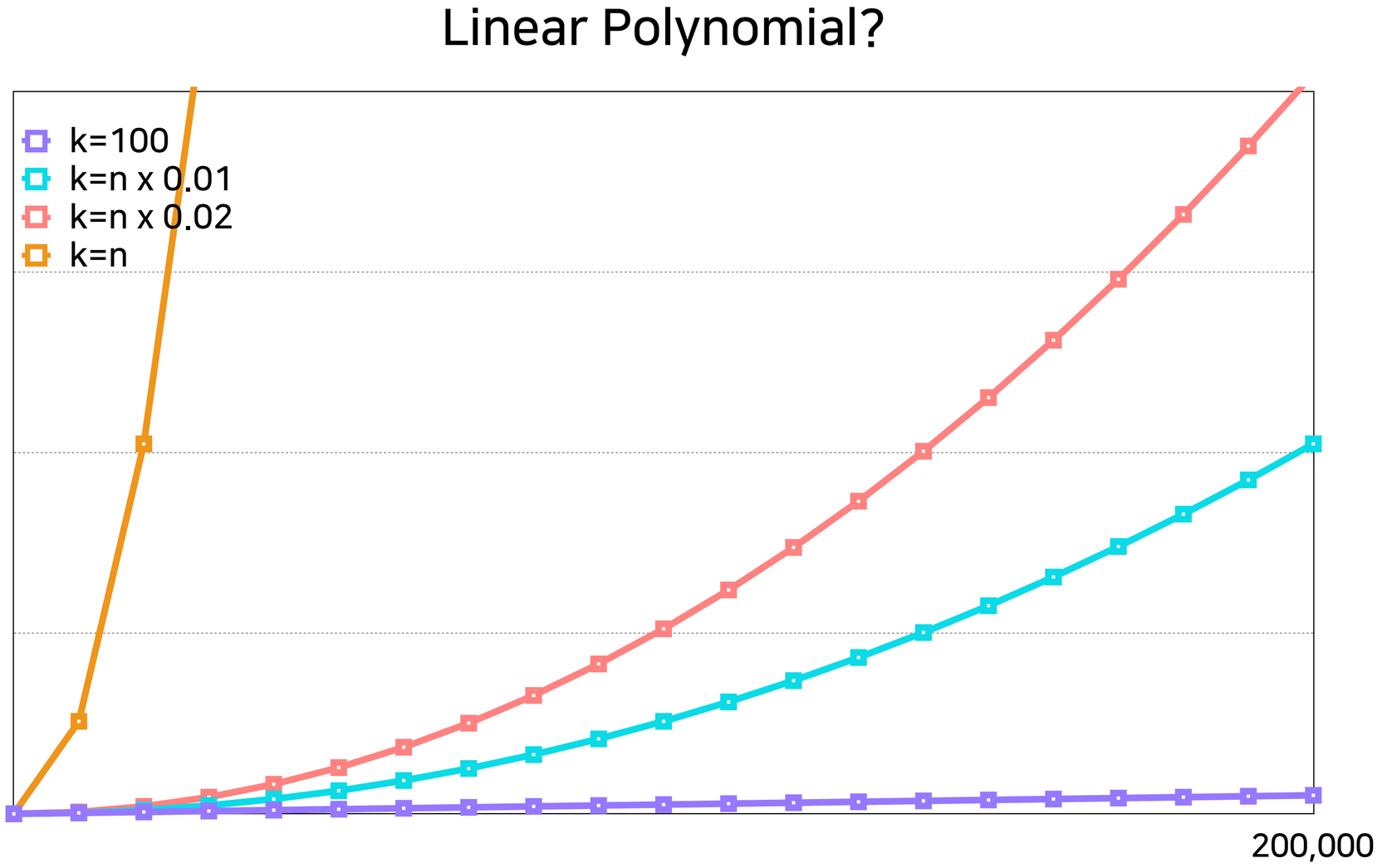
5.3 클러스터링 알고리즘의 시간 복잡도

최근 하루 상품 증가량 4백만개

클러스터 하나 당 평균 약 20개의 상품

우리 데이터는 너무나 빠르게 증가한다
그리고 클러스터의 수는 상품 수에 비례한다

5.3 클러스터링 알고리즘의 시간 복잡도

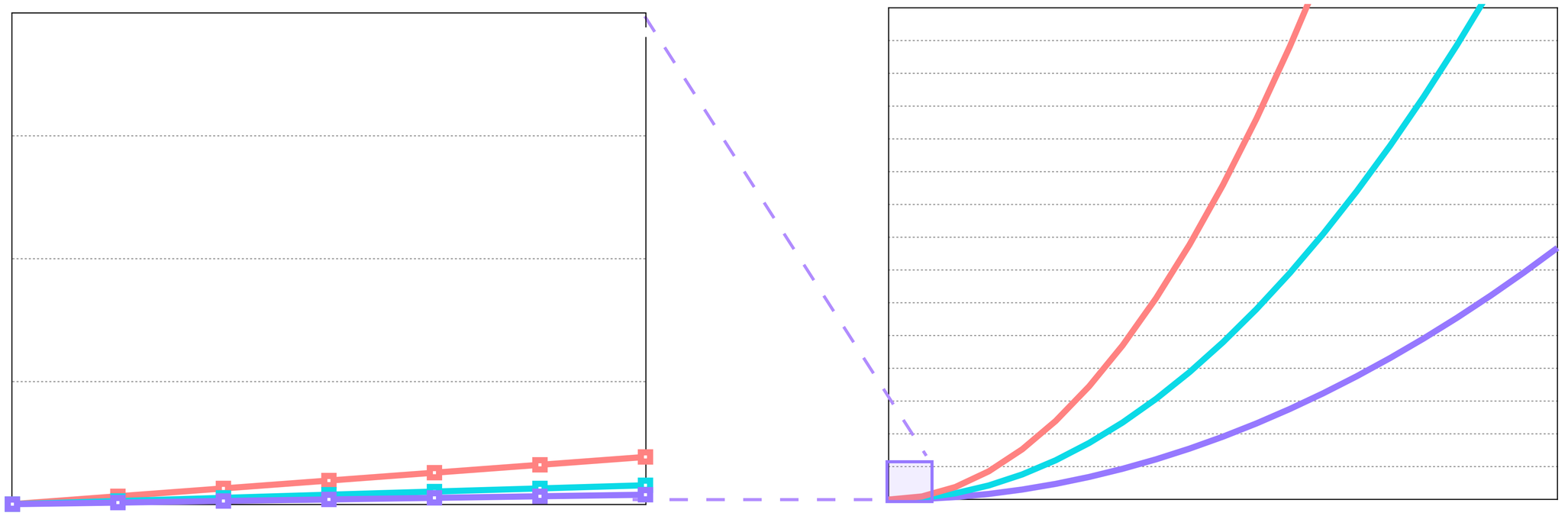


$$k = n \times \alpha$$
$$d = 128$$
$$i = 100$$

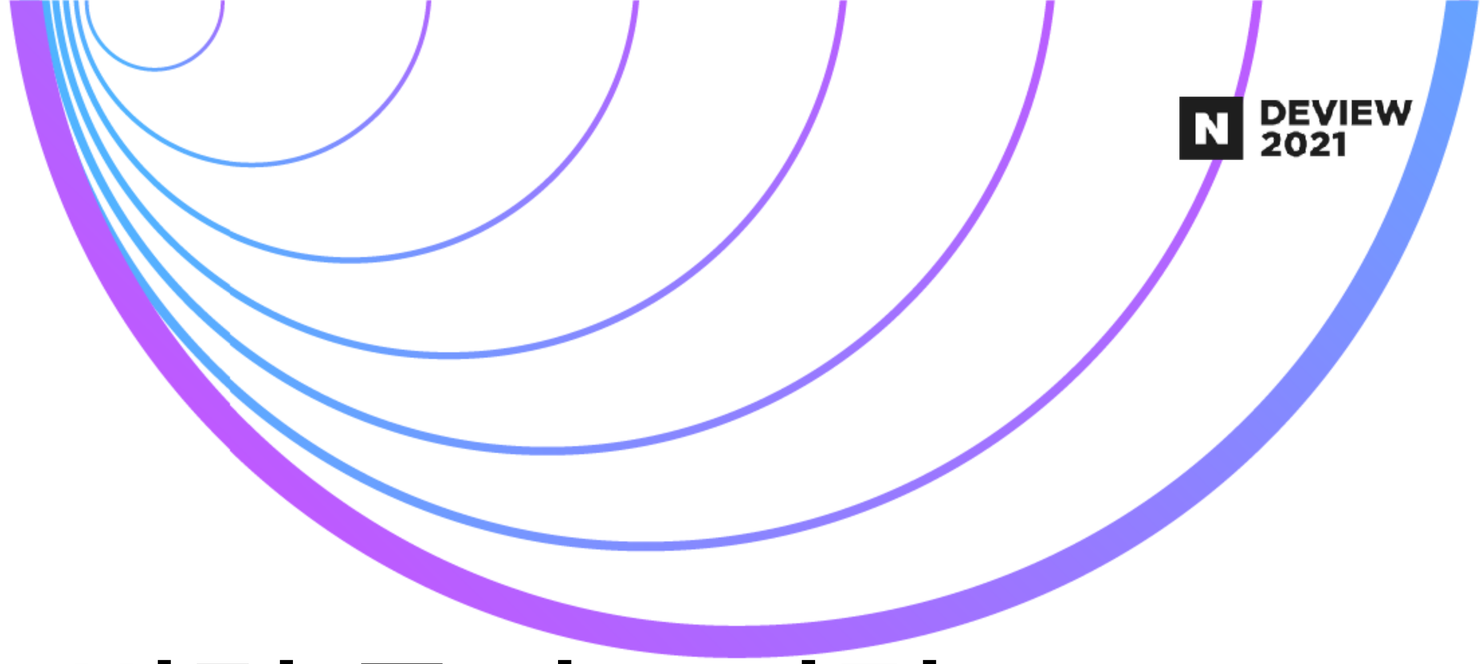
5.3 클러스터링 알고리즘의 시간 복잡도

Linear Polynomial?

Quadratic polynomial

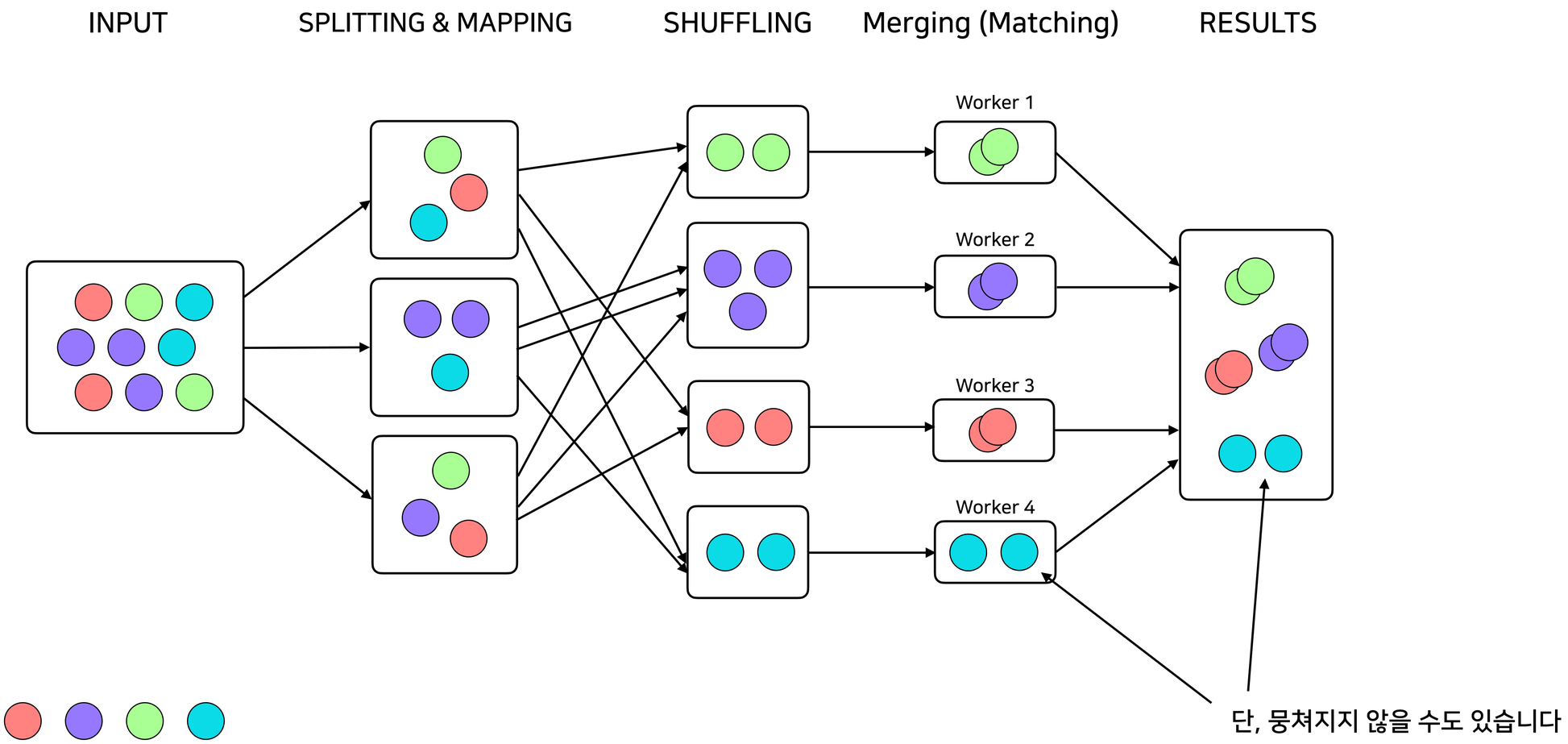


1 Billion scale



6. 대규모 병렬 클러스터링

6.1 병렬 클러스터링 시스템 구조



6.2. Weak entity

DEVIEW 2021

예를 들어, RX9 를 key로 정렬

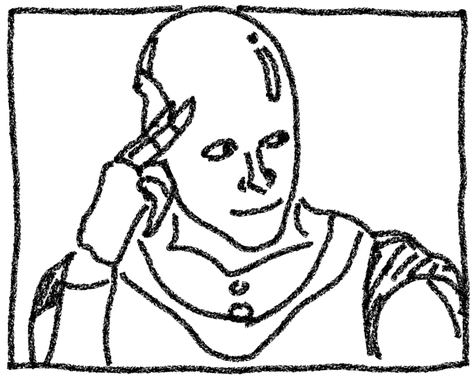
라이스타 로봇청소기 RX9

LISTAR RX9

AEG RX9 Robot Hoover

같은 이름, 다른 상품

6.3 정밀도를 높이는 상품 매칭 방법



매우 강력한 존재

이름이 RX9 이니?	네	네	네
너도 로봇이니?	네	네	네
30만원 넘니?	네	네	네
하얀색이니?	네	네	아니요
라이스타꺼니?	네	네	아니요
반말 기분 나쁘니?	아니요	아니요	아니요
NUM('아니요')	1	1	3 (탈락)

단순히 이미지나 상품 타이틀의 유사도만을 보는 것은 아니다

6.3 정밀도를 높이는 상품 매칭 방법

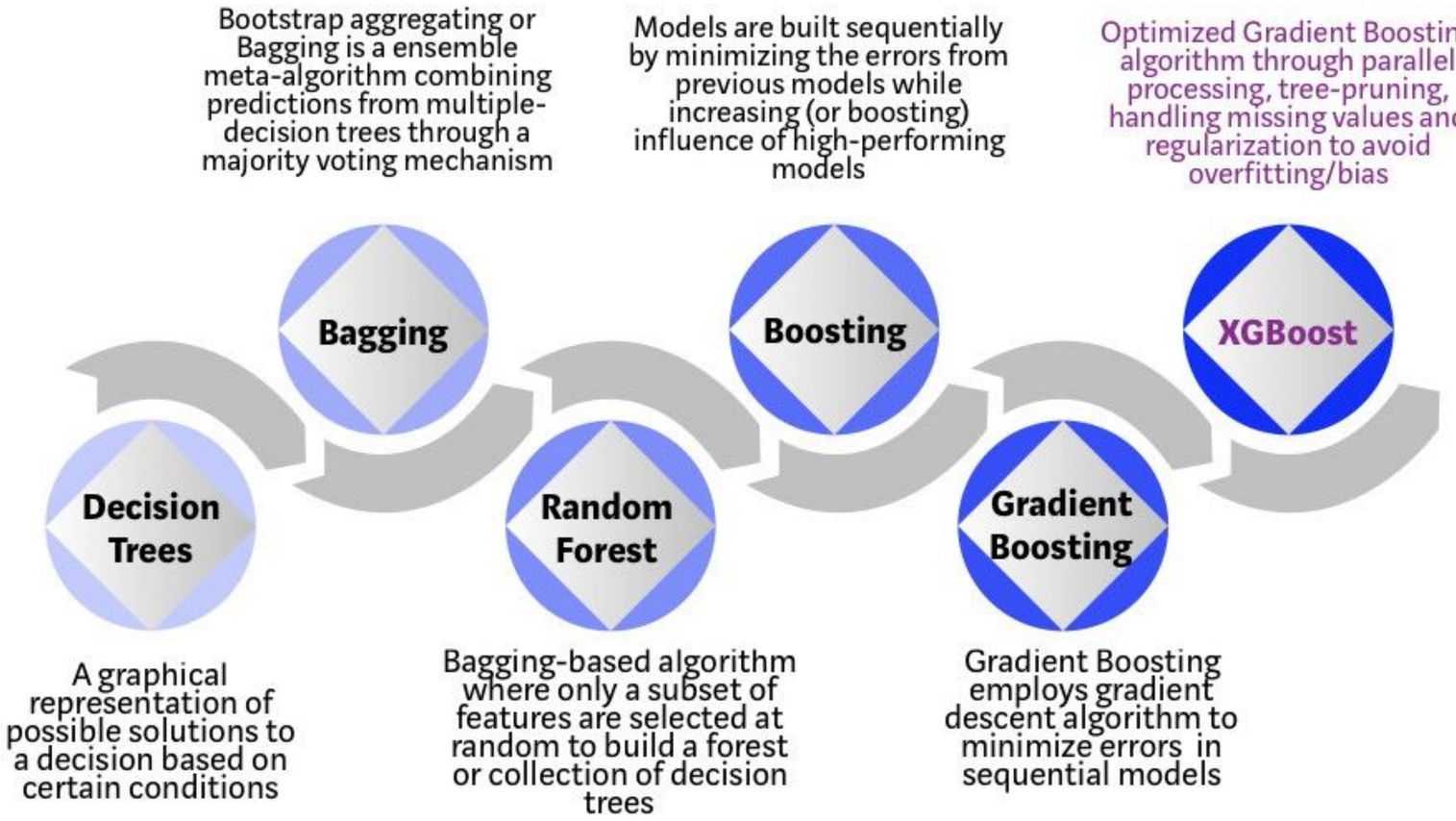
DEVIEW 2021

Name	RX9	같음	RX9
Category ID	30010	같음	30010
Price	₩300,010	₩ +-xxx	₩300,010
Color distance	“0xFFFFFF”	13.4 (*delta E)	“0x0F0100”
Brand	“LISTAR”	다름	“AEG”
Title embedding similarity	[0.1, 0.7, 0.5, ⋯, 0.6]	0.2	[0.1, 0.2, 0.0, ⋯, 0.1]
Image Similarity	[0.2, 0.2, 0.1, ⋯, 0.9]	0.9	[0.8, 0.2, 0.5, ⋯, 0.0]
…	…	…	…

어떤 기준으로 두 상품이 같다고 판단 할 수 있을까?

*delta E: http://zschuessler.github.io/DeltaE/learn/

6.3 정밀도를 높이는 상품 매칭 방법

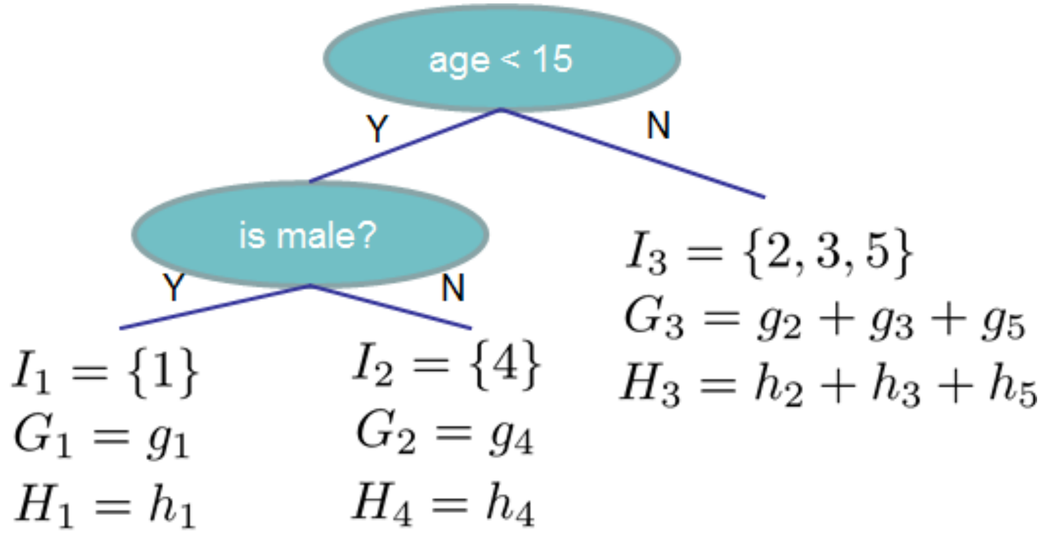


Evolution of XGBoost Algorithm from Decision Trees

https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

6.3 정밀도를 높이는 상품 매칭 방법

Instance index		gradient statistics
1		g1, h1
2		g2, h2
3		g3, h3
4		g4, h4
5		g5, h5



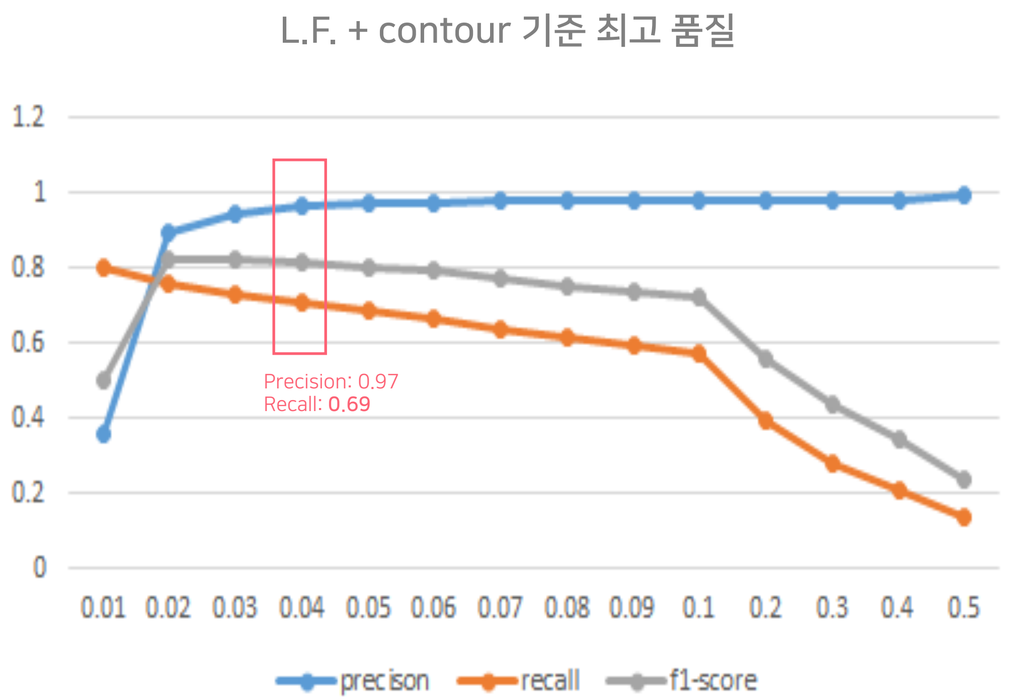
$$Obj = -\sum_j \frac{G_j^2}{H_j + \lambda} + 3\gamma$$

XGBoost Algorithm 예시

https://xgboost.readthedocs.io/en/latest/tutorials/model.html

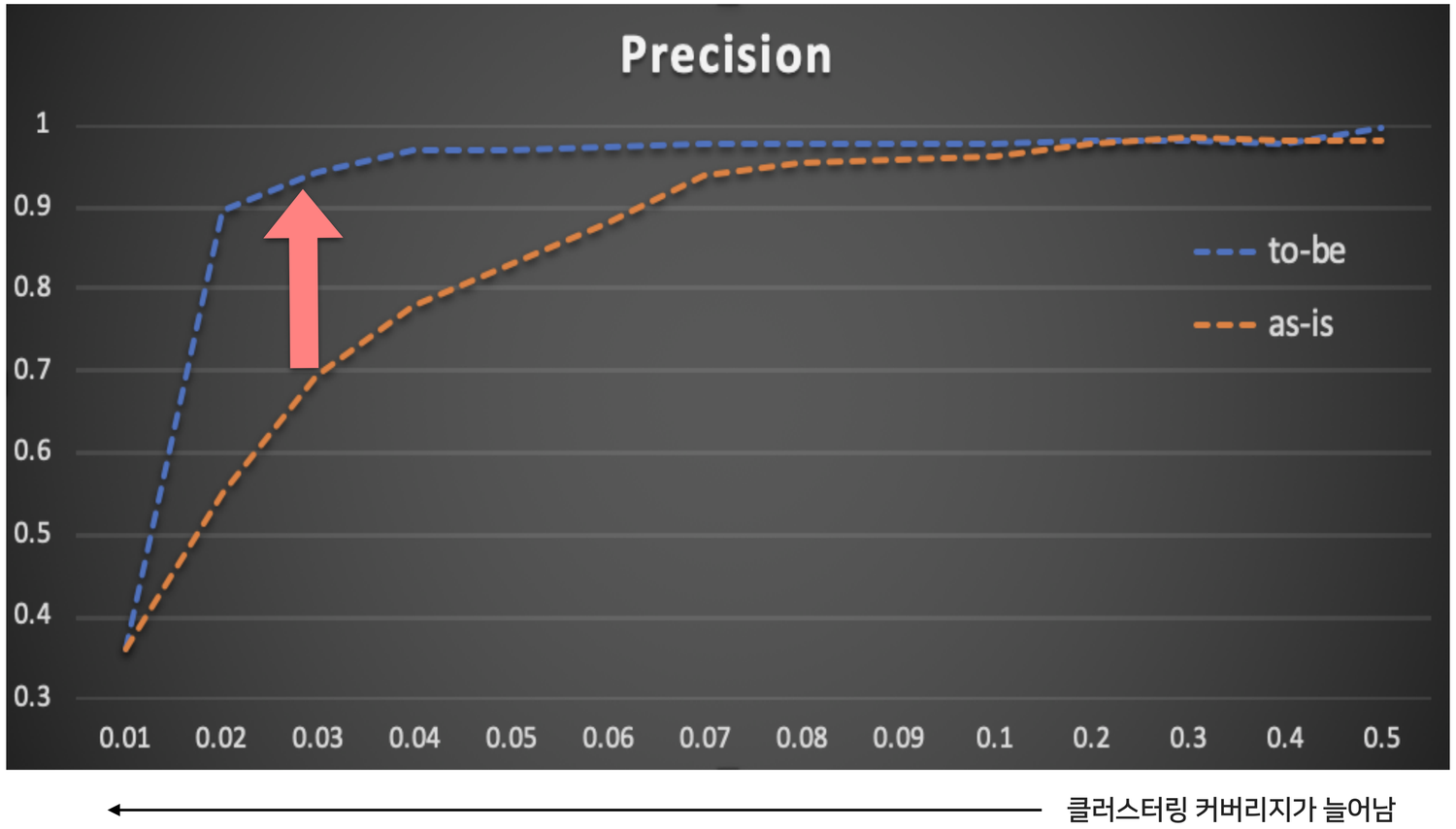
XGBoost: A Scalable Tree Boosting System: https://arxiv.org/abs/1603.02754

6.3 정밀도를 높이는 상품 매칭 방법



이미지 지역 특징으로 매칭 (SIFT-RANSAC)

6.3 정밀도를 높이는 상품 매칭 방법



낮은 threshold에서도 높은 precision을 가질 수 있는 알고리즘 연구

6.4 Data skewness


AEG RX9-1-SGM Robot Hoover

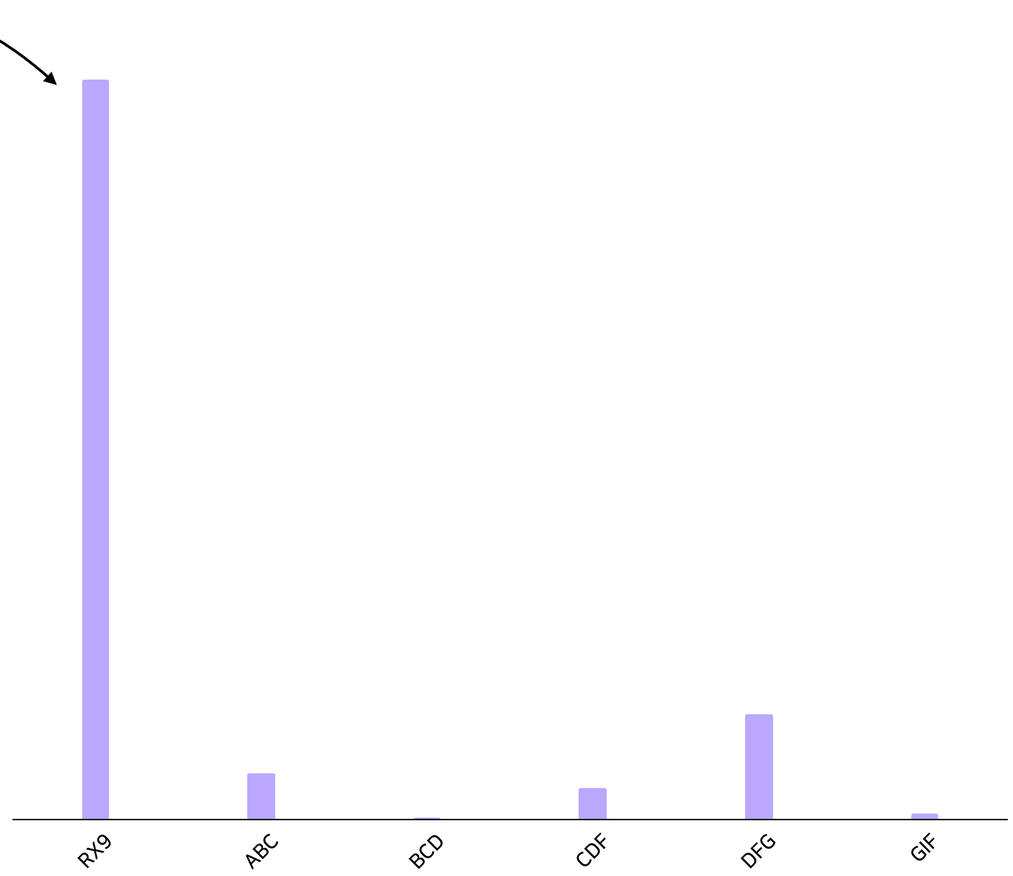

라이스타 RX9


[LISTAR] 라이스타 로봇청소기 RX9

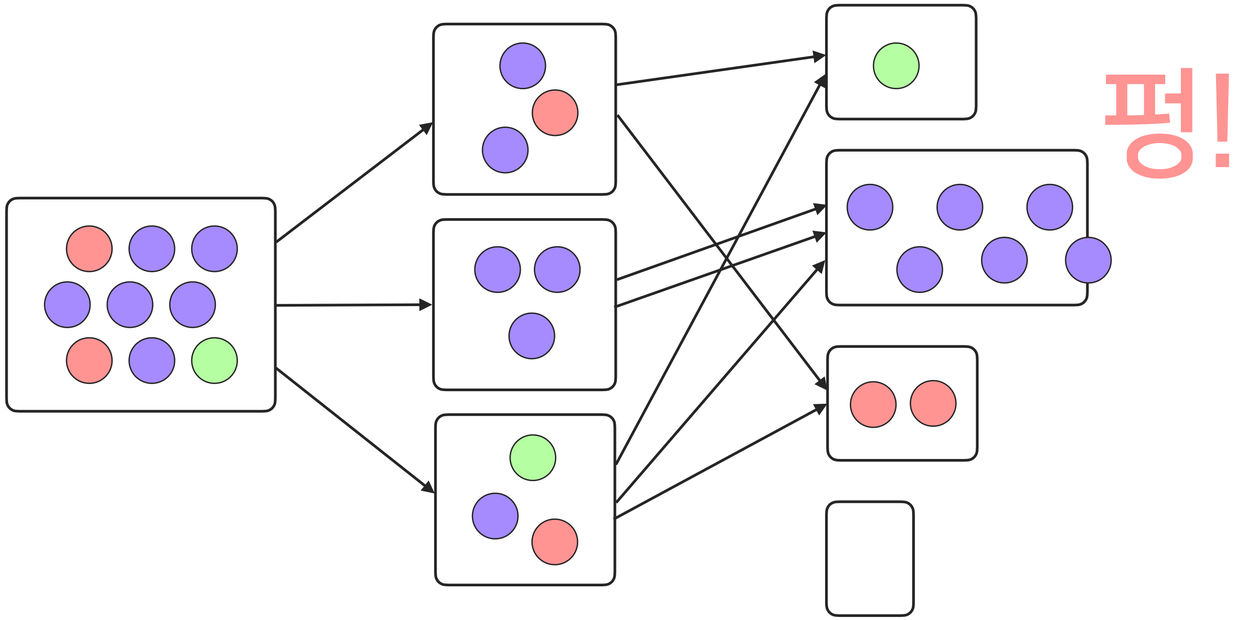

House cleaningAeg RX9-1-SGM


AEG RX9-2-4ANM

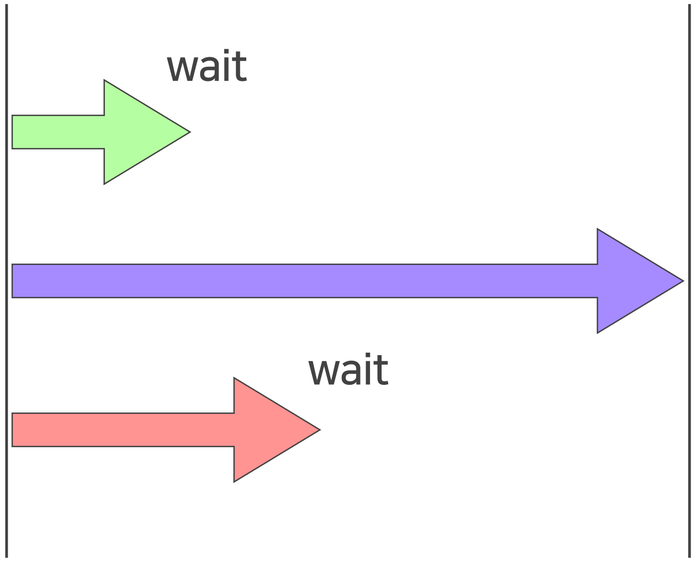

단후이 RX9



6.4 Data skewness

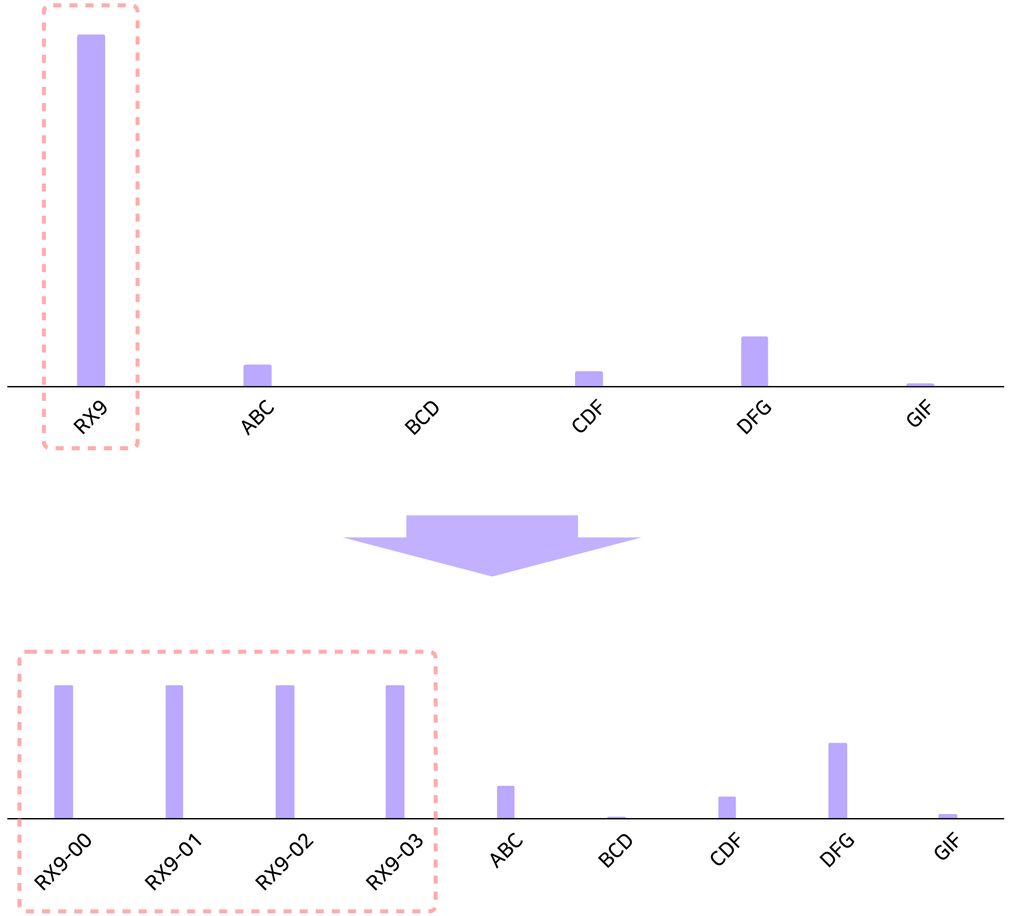


메모리 부족

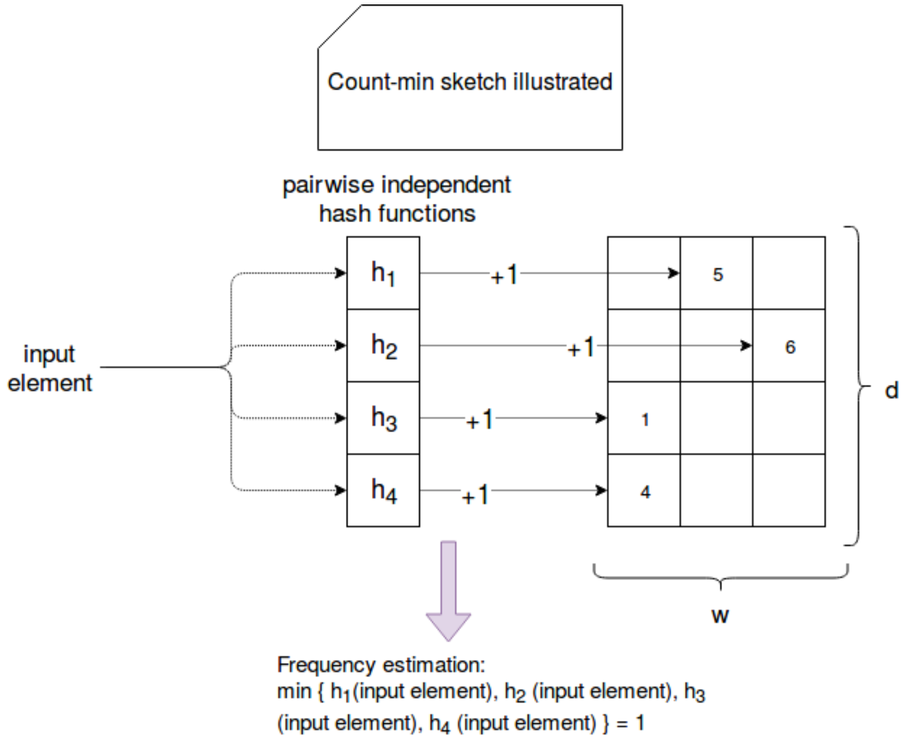


Execution Time

6.4 Data skewness



6.4 Data skewness

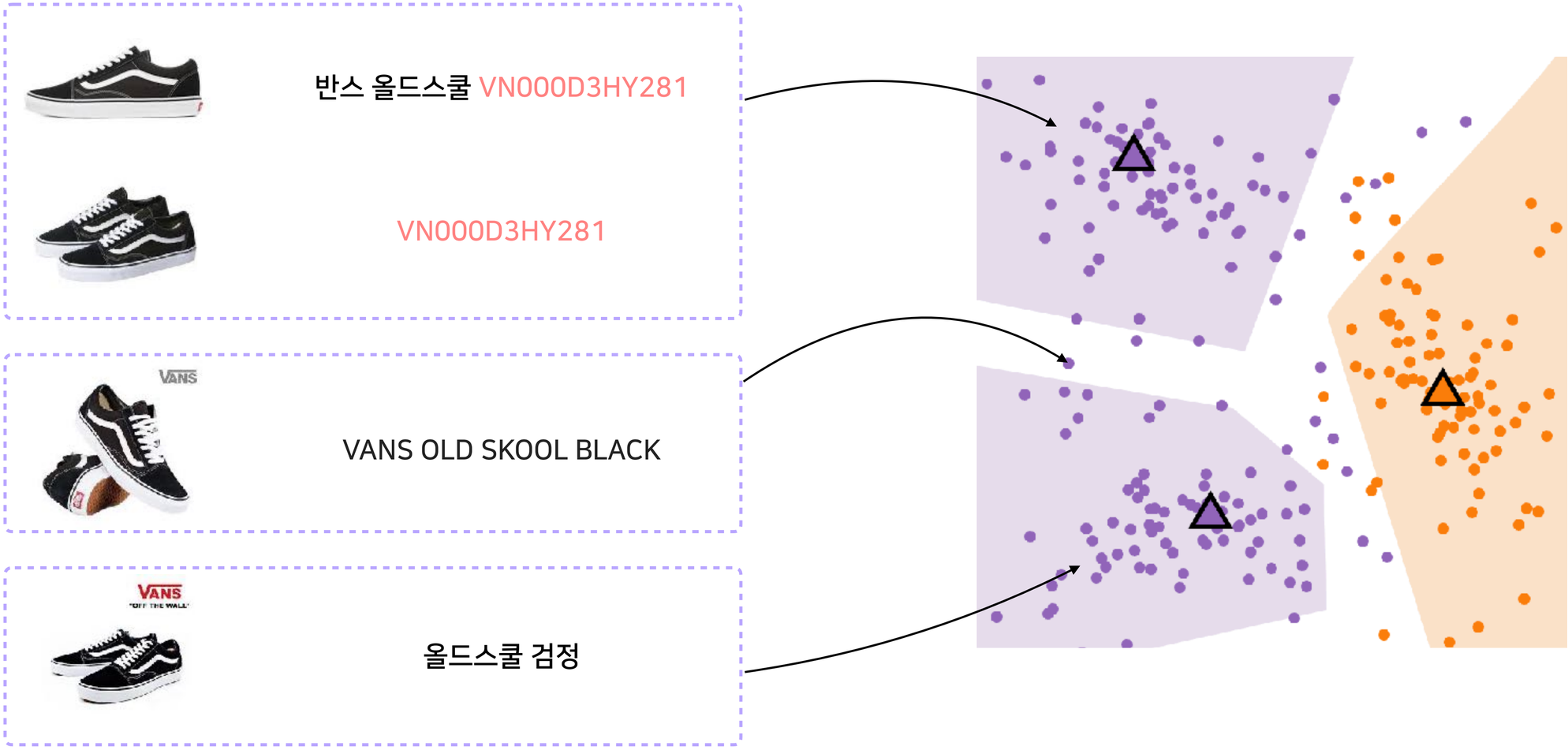


- Initialization: $\forall i \in \{1, \ldots, d\}, j \in \{1, \ldots, w\}, : \text{count}[i, j] = 0$
- Increment count (of element a): $\forall i \in \{1, \ldots, d\} : \text{count}[i, h_i(a)] \mathrel{+}= 1$
- Retrieve count (of element a): $min_{i=1}^{d} \text{count}[i, h_i(a)]$

Count-min Sketch 알고리즘

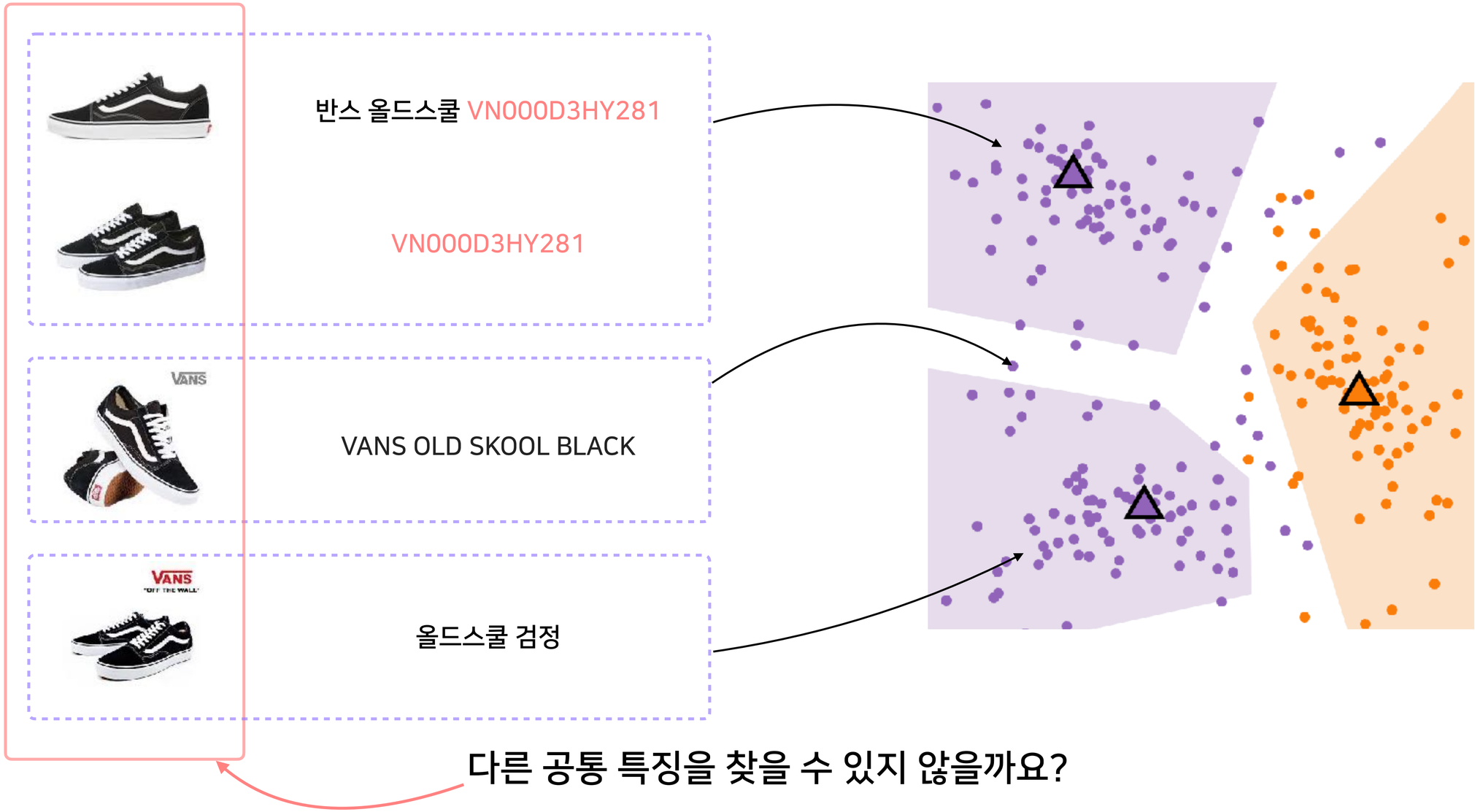
https://www.waitingforcode.com/big-data-algorithms/frequency-estimation--count-min-sketch/read

6.5 클러스터 파편화



동일한 상품이 여러 서브 스페이스로 나뉘어짐

6.5 클러스터 파편화


반스 올드스쿨 VN000D3HY281
VN000D3HY281
VANS OLD SKOOL BLACK
올드스쿨 검정
다른 공통 특징을 찾을 수 있지 않을까요?

6.6 클러스터 파편화를 완화하는 key 생성

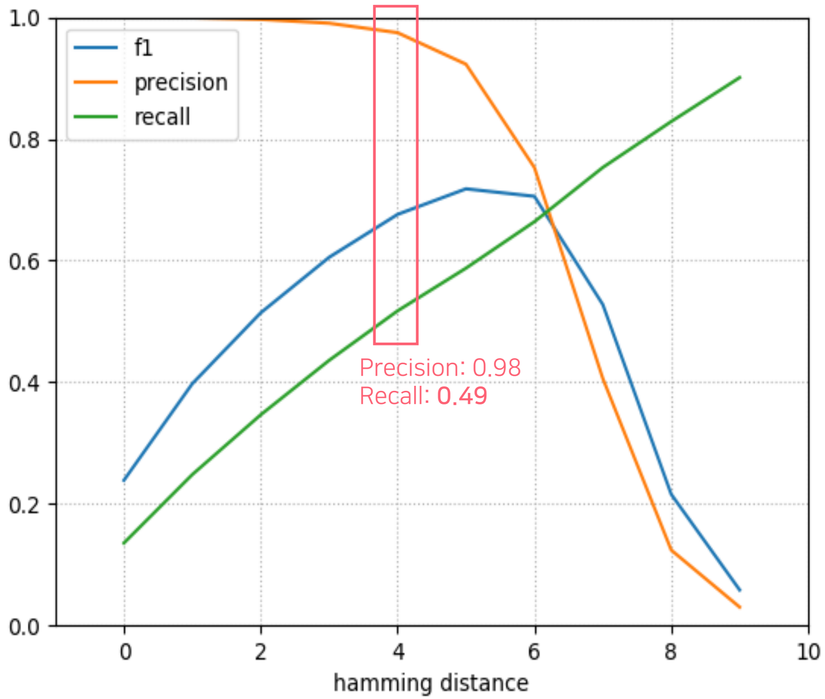
image hash: 041ef0aaf9cd3

Hamming distance = 0

image hash: 041ef0aaf9cd3

Hamming distance = 5

image hash: 0416f0aae9ed3



이미지 해시 생성 기법

http://www.hackerfactor.com/blog/index.php?/archives/432-Looks-Like-It.html

6.6 클러스터 파편화를 완화하는 key 생성

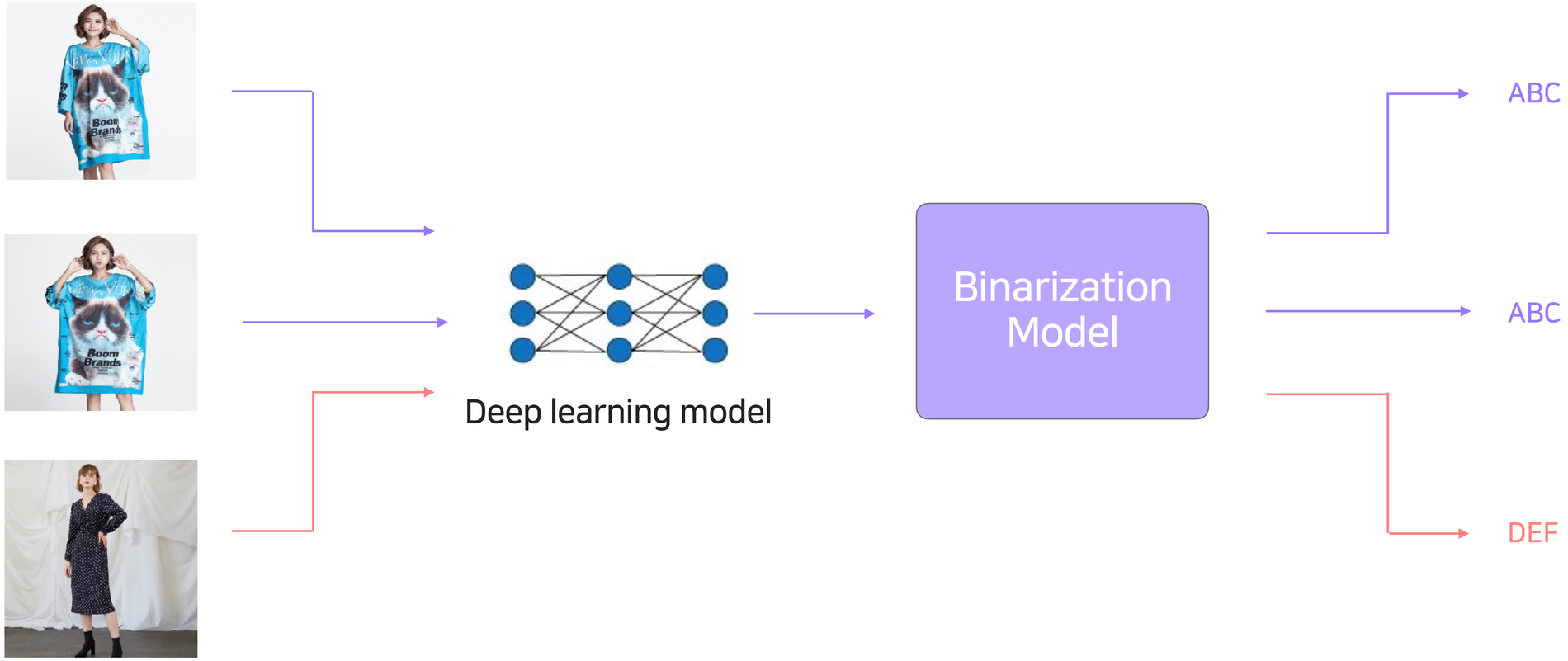
모델의 자세가 다른 이미지

이미지의 크롭이 다른 이미지

이미지 구도가 다른 이미지

패션 상품(rigid body)에 불리함

6.6 클러스터 파편화를 완화하는 key 생성

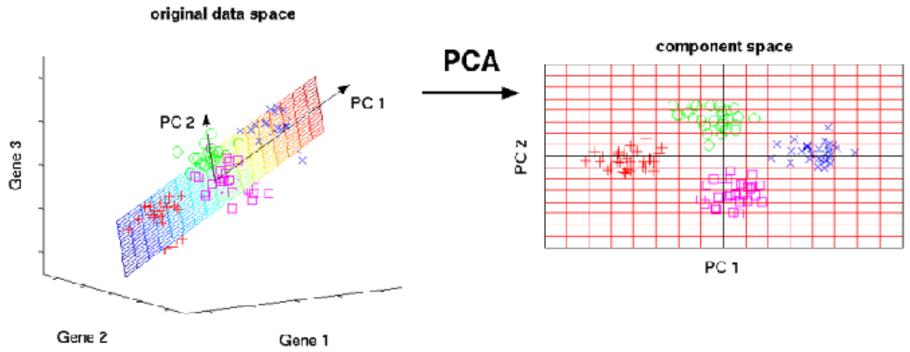


Deep feature를 binarize 하는 방법

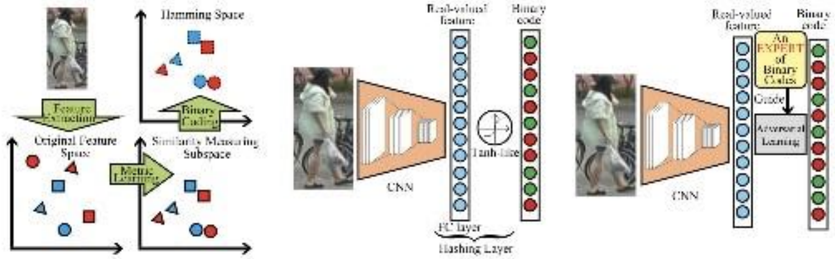
6.6 클러스터 파편화를 완화하는 key 생성

DEVIEW 2021

PCA를 사용한 deep feature 차원 축소



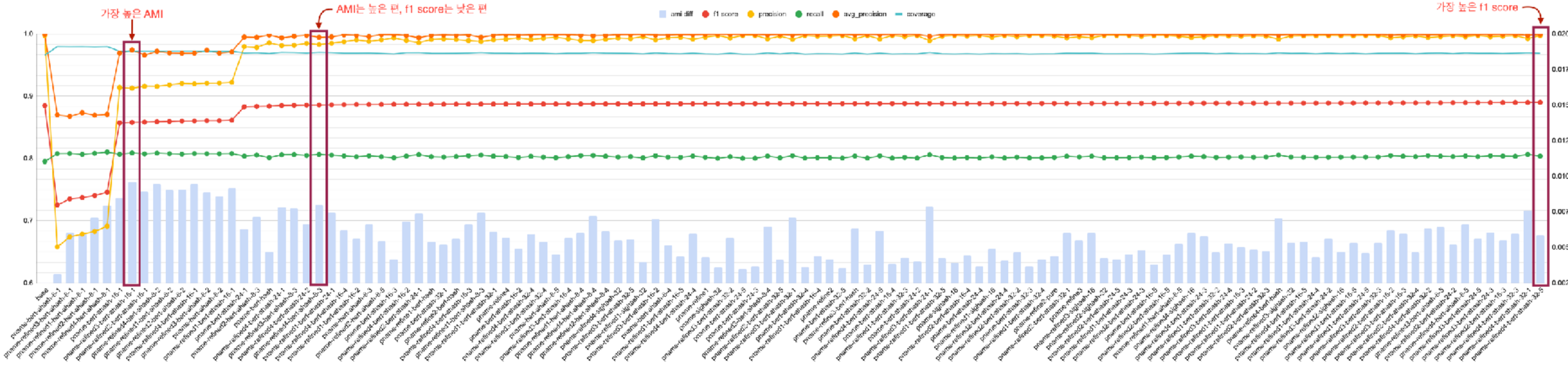
Deep hash 구조



sighash	상품 명
"4336424425236656"	"GOLD 반지 방울 콤보링 여성금반지 RCS2709 18K반지 26호 반지"
"4336424425236656"	"GOLD 반지 방울 콤보링 여성금반지 RCS2709 18K반지 17호 반지"
"4336424425236656"	"GOLD 반지 방울 콤보링 여성금반지 RCS2709 18K반지 22호 반지"
"4336424425246656"	"GOLD 반지 방울 콤보링 여성금반지 RCS2709 18K반지 3호 반지"
"4336424425236656"	"GOLD 반지 방울 콤보링 여성금반지 RCS2709 18K반지 14호 반지"
"4326435535337655"	"패션플러스 GOLD 반지 방울 콤보링 여성금반지 RCS2709 18K반지 22호 반지"
"4336435535236655"	"패션플러스 GOLD 반지 방울 콤보링 여성금반지 RCS2709 18K반지 13호 반지"
"4336435535236655"	"패션플러스 GOLD 반지 방울 콤보링 여성금반지 RCS2709 18K반지 10호 반지"
"4336435535236655"	"패션플러스 GOLD 반지 방울 콤보링 여성금반지 RCS2709 18K반지 4호 반지"
"4336434435236655"	"AK온라인몰 GOLD 반지 방울 콤보링 여성금반지 RCS2709 18k반지"

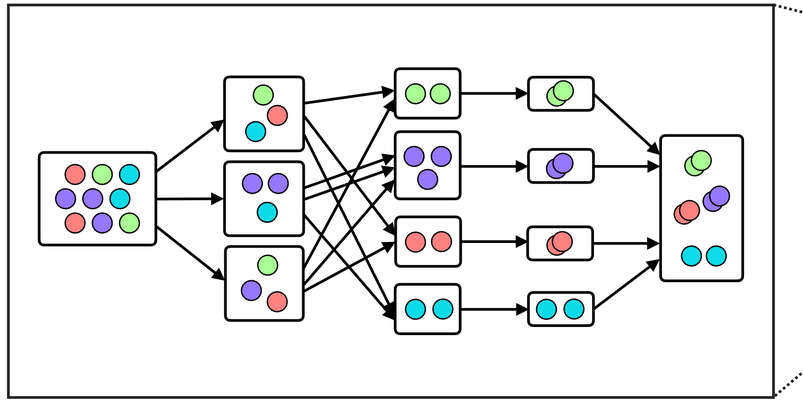
PCA / Deep hash를 사용한 binarization 기법

6.6 클러스터 파편화를 완화하는 key 생성



한 달간 생성한 key들과 품질 검증

6.7 단계 별 클러스터링



Stage 1

key: 상품명 임베딩 해시

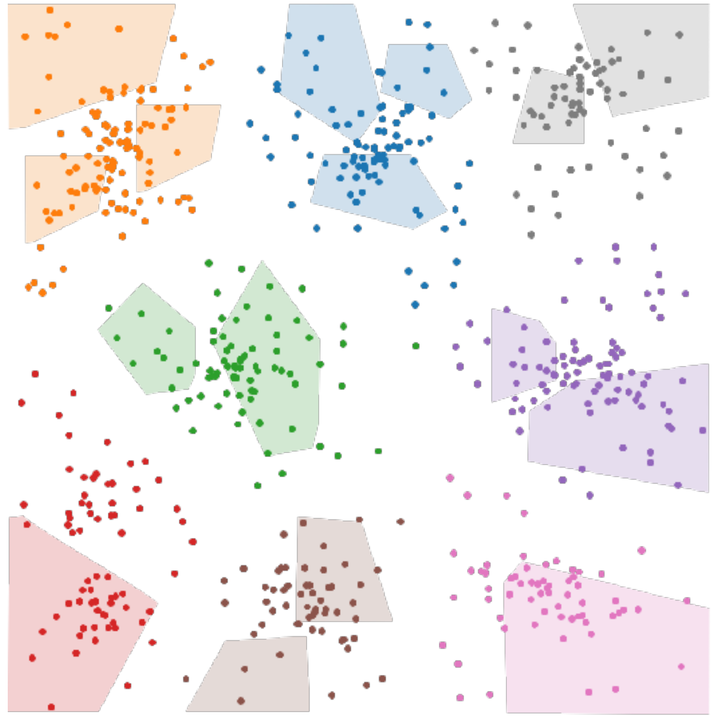
Stage 2

key: 이미지 임베딩 해시

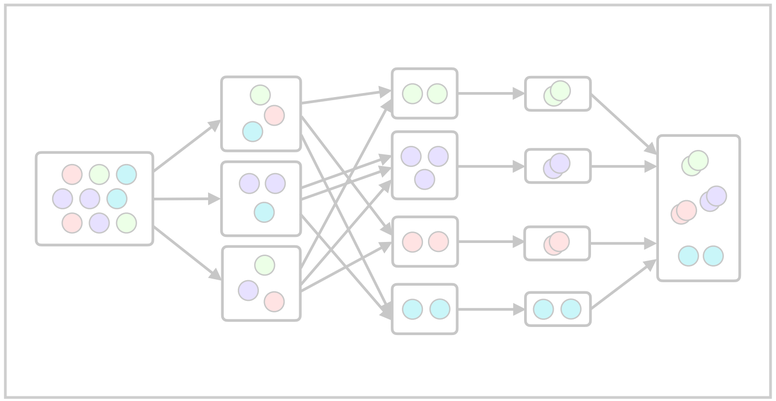
Stage 3

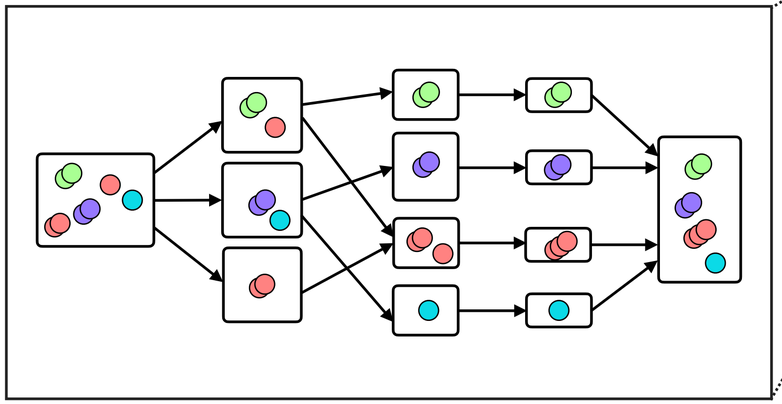
key: 상품 이름

⋮



6.7 단계 별 클러스터링





Stage 1

key: 상품명 임베딩 해시

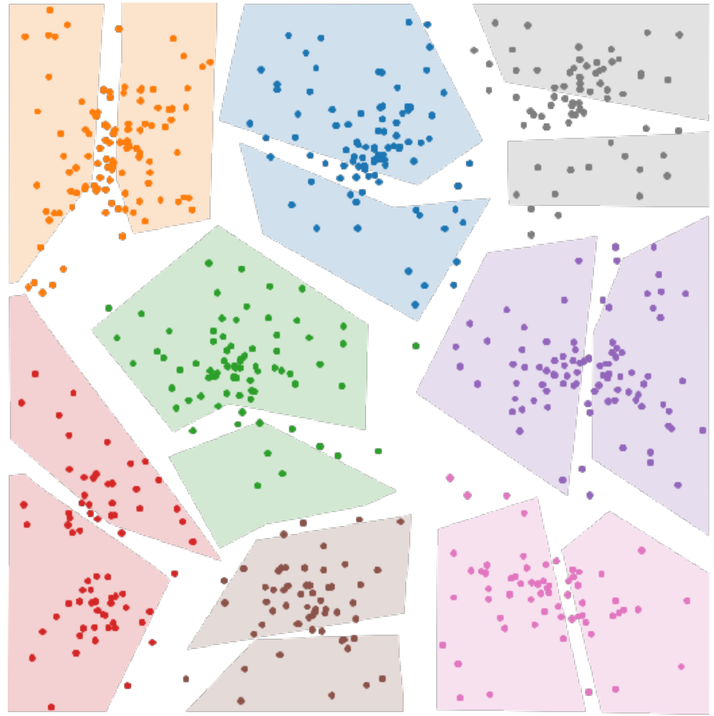
Stage 2

key: 이미지 임베딩 해시

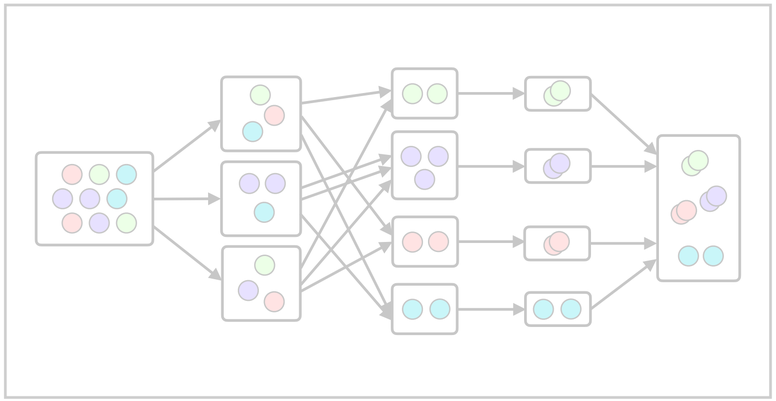
Stage 3

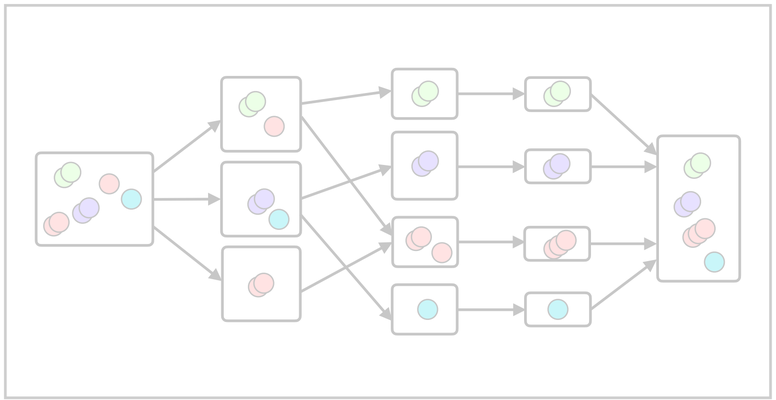
key: 상품 이름

⋮



6.7 단계 별 클러스터링





Stage 1

key: 상품명 임베딩 해시

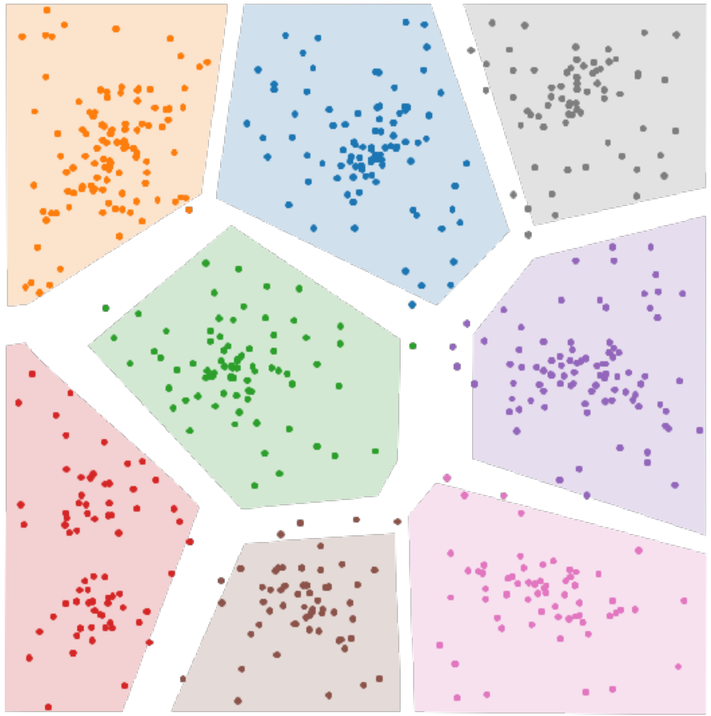
Stage 2

key: 이미지 임베딩 해시

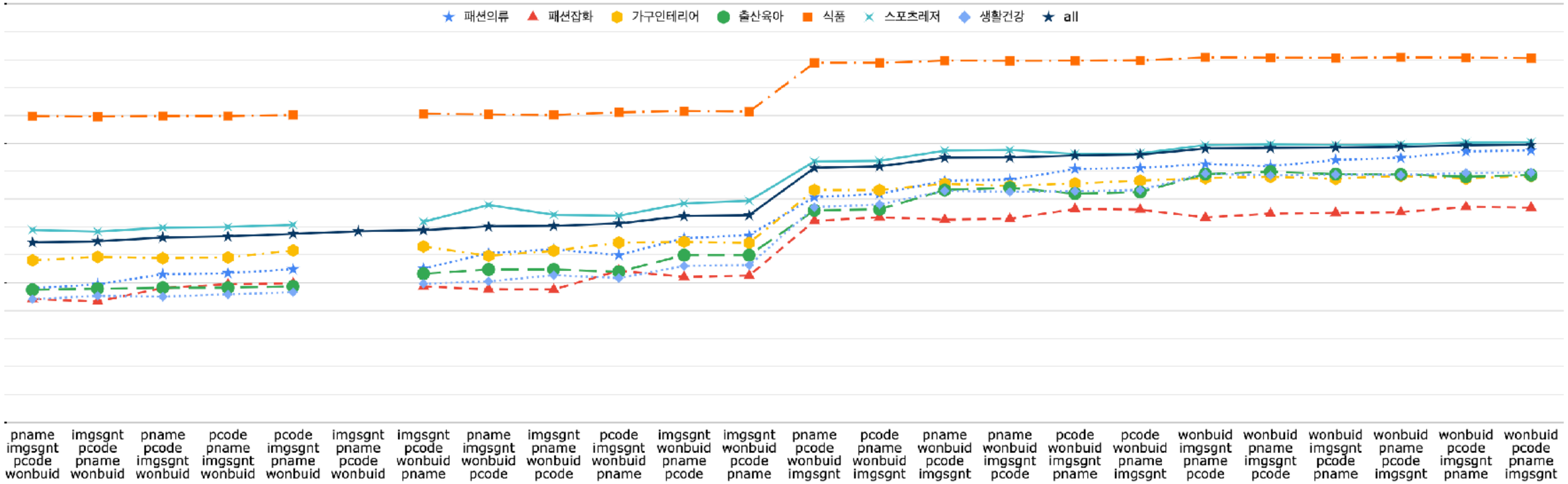
Stage 3

key: 상품 이름

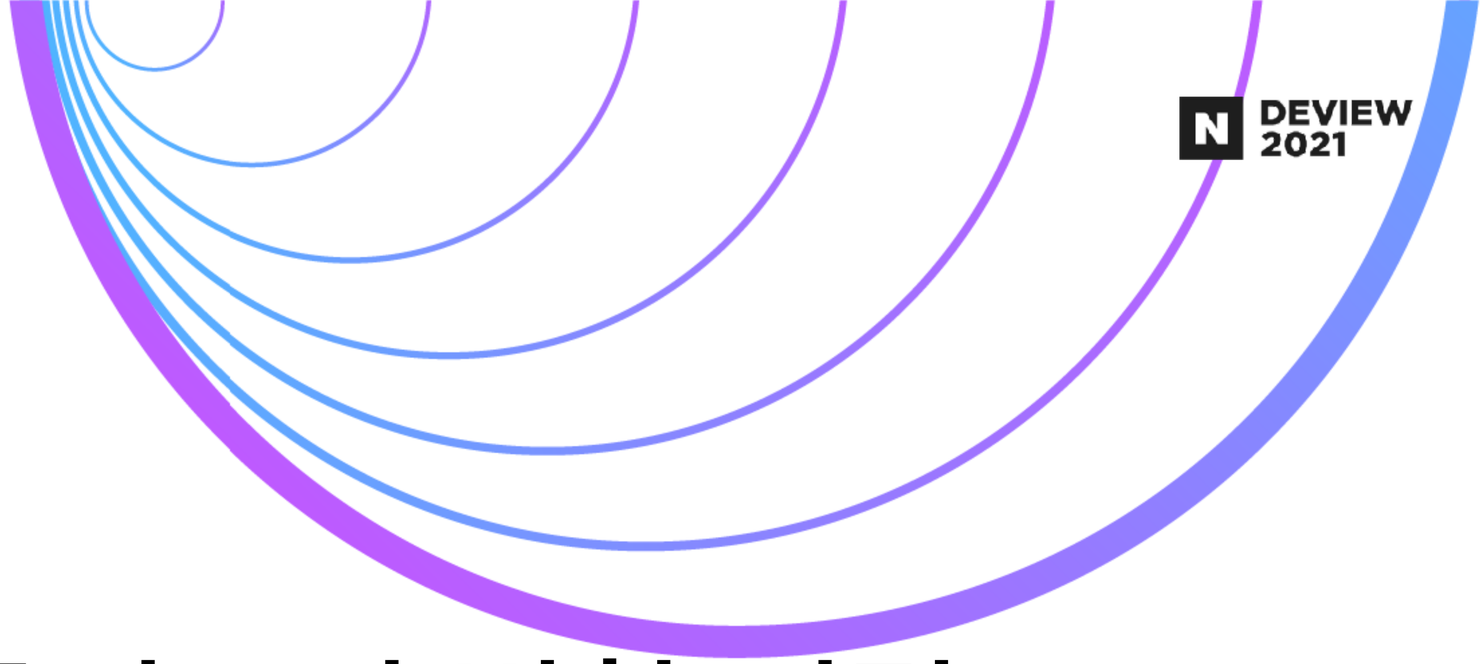
⋮



6.7 단계 별 클러스터링



key가 클러스터링에 사용되는 순서에 따라 품질이 달라짐



7. 대형 클러스터 병합 전략

7.1 대형 클러스터 간 머지

DEVIEW 2021

클러스터 A

id:25525022701
category_name:식품>음료>청량/탄산음료>사이다
product_name:[롯데칠성] 칠성사이다 210mlx30캔
po:BRAND=롯데칠성음료=|QTY=30=캔|VOL=210=ml
brand:롯데칠성음료
maker:
price:14900

id:23097629008
category_name:식품>음료>청량/탄산음료>사이다
product_name:[롯데닷컴]롯데칠성 칠성사이다 210mlx30캔/ 롯데칠성
po:BRAND=롯데칠성음료=|QTY=30=캔|VOL=210=ml
brand:롯데칠성음료
maker:
price:15190

id:26529851664
category_name:식품>음료>청량/탄산음료>사이다
product_name:[롯데백화점] 롯데칠성 칠성 사이다(210ml*30캔) / 무료배송
po:BRAND=롯데칠성음료=|QTY=30=캔|VOL=210=ml
brand:롯데칠성음료
maker:
price:14970

id:23775088469
category_name:식품>음료>청량/탄산음료>사이다
product_name:무료배송 칠성사이다 210mlx30캔/ 롯데칠성
po:BRAND=롯데칠성음료=|QTY=30=캔|VOL=210=ml
brand:롯데칠성음료
maker:
price:15820

id:26314465722
category_name:식품>음료>청량/탄산음료>사이다
product_name:롯더칠성 [본사직영] 롯데 칠성사이다 210mlx30캔
po:BRAND=롯데칠성음료=|QTY=30=캔|VOL=210=ml
brand:롯데칠성음료
maker:
price:14450

⋮

클러스터 B

id:21959196860
category_name:식품>음료>청량/탄산음료>사이다
product_name:낙더마트/칠성사이다 210mlx30캔
po:BRAND=롯데칠성음료=|QTY=30=캔|VOL=210=ml
brand:롯데칠성음료
maker:
price:13670

id:23936497384
category_name:식품>음료>청량/탄산음료>사이다
product_name:[롯데칠성음료] 칠성사이다 탄산음료 210ml X30캔
po:BRAND=롯데칠성음료=|QTY=30=캔|VOL=210=ml
brand:롯데칠성음료
maker:
price:15900

id:20135574181
category_name:식품>음료>청량/탄산음료>사이다
product_name:칠성사이다 210ml X 30캔
po:BRAND=롯데칠성음료=|QTY=30=캔|VOL=210=ml
brand:롯데칠성음료
maker:
price:20390

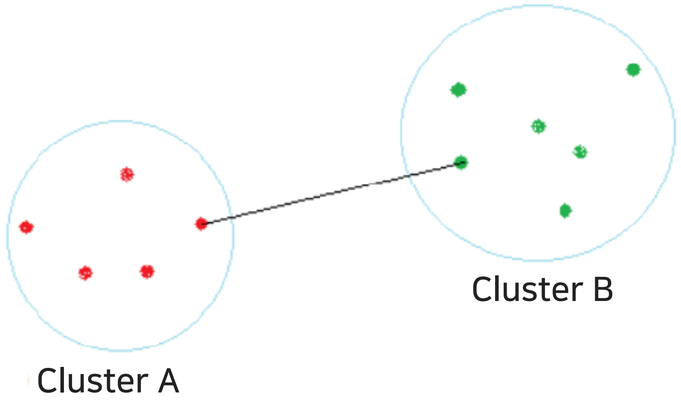
id:24318432488
category_name:식품>음료>청량/탄산음료>사이다
product_name:칠성사이다 210mlx30캔
po:BRAND=롯데칠성음료=|QTY=30=캔|VOL=210=ml
brand:롯데칠성음료
maker:
price:15590

id:23623917929
category_name:식품>음료>청량/탄산음료>사이다
product_name:칠성사이다 칠성사이다 210mlx30캔
po:BRAND=롯데칠성음료=|QTY=30=캔|VOL=210=ml
brand:롯데칠성음료
maker:
price:16950

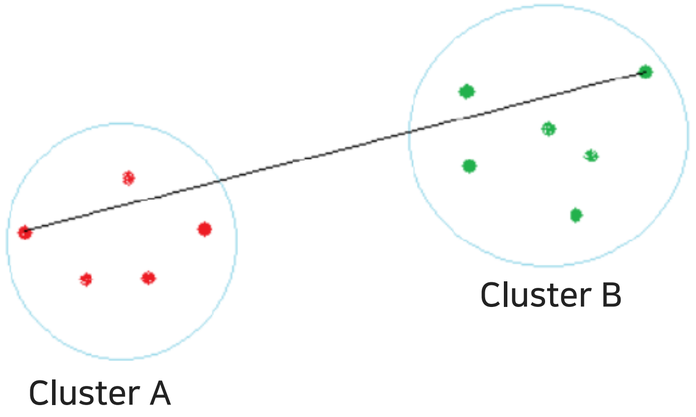
⋮

7.1 대형 클러스터 간 머지

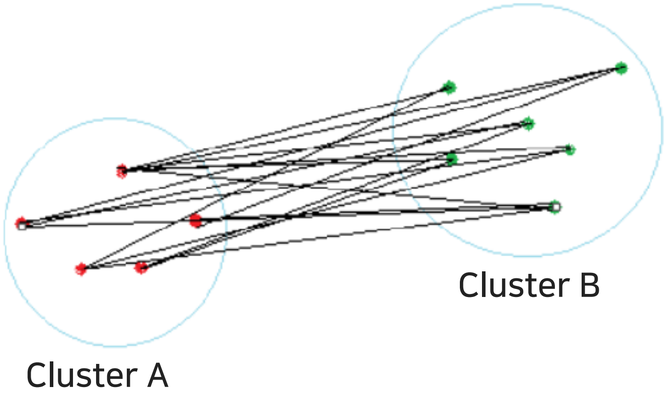
Single Linkage



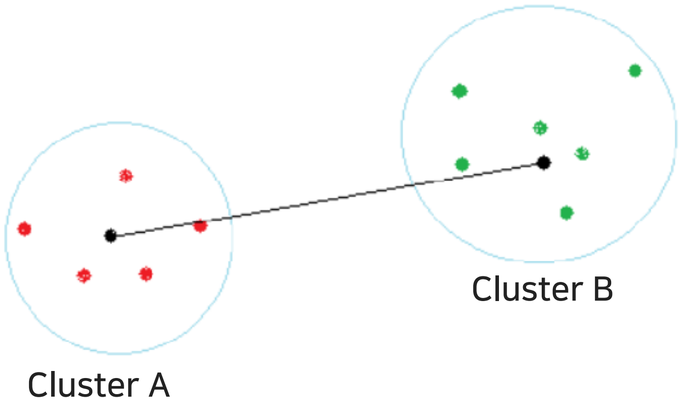
Complete Linkage



Average Linkage



Centroid Linkage



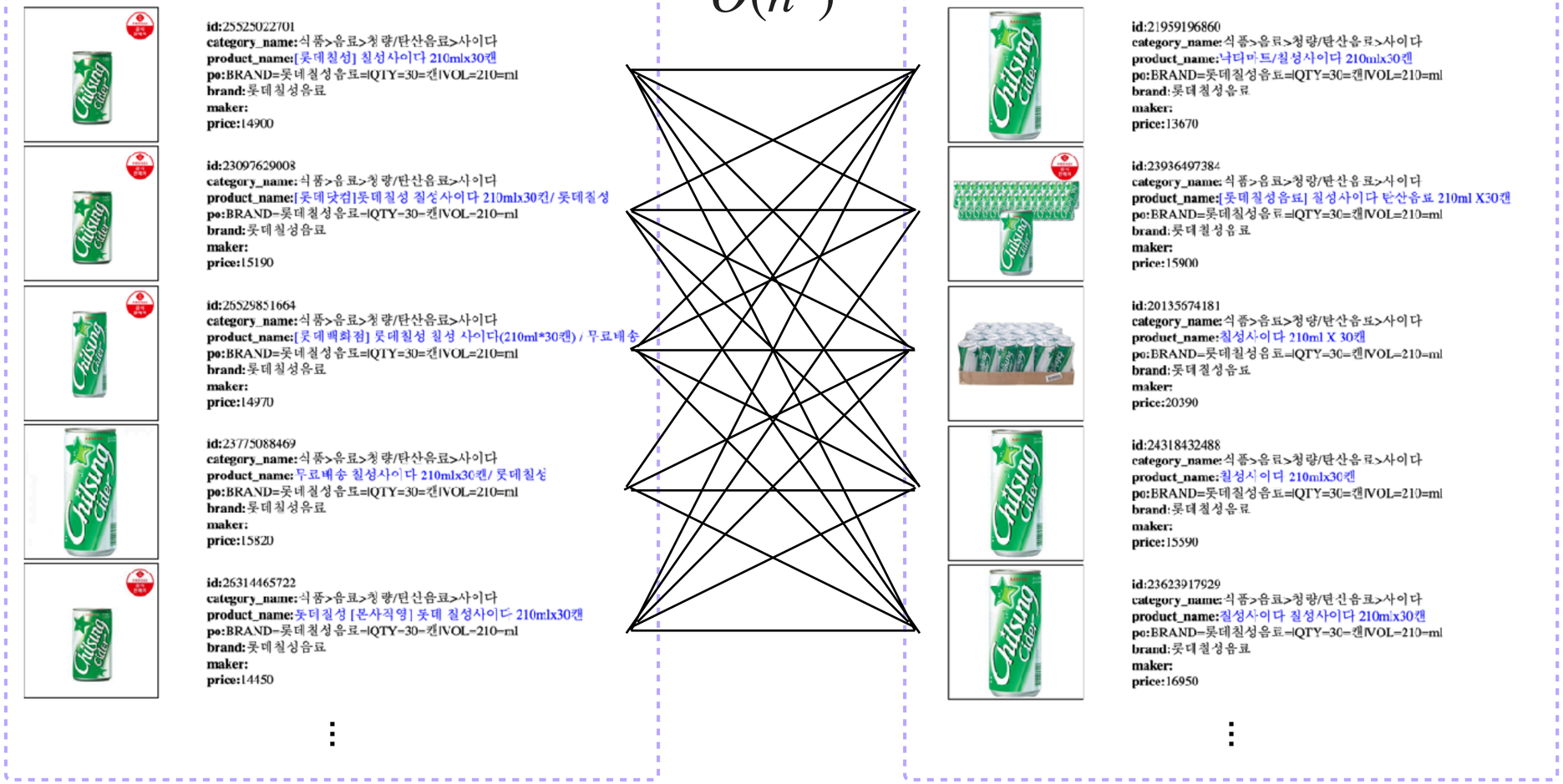
7.1 대형 클러스터 간 머지

DEVIEW 2021

클러스터 A

$O(n^2)$

클러스터 B



7.1 대형 클러스터 간 머지


```
{"id":
["1943887200","1934023619","1952980881","4039203022","3185548215","3278417718","4190130002","1938808826","1933778843","5768457840","1933772840","5768457843","6995494072","327839220
2","5591227348","4170163648","4092448477","3525945835","6715113334","3449656722","3008660924","6730078541","3504883548","1944370540","3462341552","4211097410","4028541125","4211818
596","4157806355","5854391168","4181181266","4186218365","3278906457","3186303604","2964320578","3179076820","4154293372","1934288433","5914890167","2686883176","1946891435","33597
63978","3039718782","4046510976","4022366065","3242677728","3008491088","2026566166","4016156520","4216063546","2026528718","3091477262","2016171970","3278733835","1952974728","195
2982262","3449794153","3727267257","5576905586","1943035847","3490720136","4091159662","3622284755","1936318612","1945797784","1943872414","3184452025","1934273778","3805197971","2
015573008","4061665409","4182744214","6598760403","1942473641","4047972582","3184078713","4048284000","2677008711","3278978417","1934302776","3007776259","3727303558","1945788048",
"3449678874","3449611682","3504851850","4213042910","2963346353","3614990785","4209855023","4082834718","1942477972","4130690816","6615072447","4182744223","4154376370","3475069558
","6919889207","3665413217","4181394820","4212406608","3406563638","5602785564","4213452145","1953136989","4161951032","3504867318","2965179642","1944376936","4145284210","42135791
68","2000254178","3614108472","4216578244","1940305963","2065271889","6841977725","4215486896","4093276171","6615095801","3462413608","3148931187","4184389957","4212979996","405034
3500","1956724009","3462721098","2803192222","4186173743","3607205559","4214244955","1934023620","4031017143","1944370542","4186230691","4088244979","2800796493","4136814763","3278
733019","1940196356","1940196357","3462345571","3022600256","6598747038","4145898649","5632663291","3185326273","1956724010","327881870S","3187047910","4053151962","4041992882","19
36334435","4184138092","2026563467","3406583913","1944401425","3186696605","7048564554","3091158218","2065276772","1946569666","1946569667","6881892903","3244855274","3620665241","
3148071587","4135810219","3186951238","4137985495","6995461756","3462570751","2000254251","6519842449","1943949523","4217136619","2689610999","4181275893","3509461811","34631214O5"
,"1938797288","4145804200","3184725187","6919963561","6841977722","3621410358","3449529671","5793566777","3504881925","2691590472","6715119226","3041071711","3008284348","318414056
0","5568532538","3662203197","1953132284","5835114775","3311106876","1943007862","3278413588","3178647389"],"upc":
["4987316025232","4987300058619","0100598","0100597","ssp","4987300058602"],"catId":["10009183"],"brand":["エスエス製薬"],"pname":["ハクビ c 180錠 セット 第3類 医薬品","エスエス製 薬 株式 会社
ハイチ オール c ホワイ ティア 120錠x セット","薬 エスエス製薬 ハイチ オール c ホワイ ティア 40錠","エスエス ハイチ オール c ホワイ ティア 120錠","第3類 医薬品 ハイチ オール c ホワイ ティア 120錠","ハイチ オール c ホワイ ティア
120錠","ハイチ オール c ホワイ ティア 120錠x しみ 日 やけなどの 色素 沈着に","ハイチ オール c ホワイ ティア 120錠 定形外 郵便 代引き 不可 シミ そばかす 1- システイン 第3類 医薬品","ハイチ オール c ホワイ ティア シミ そばかすに
ハイチ オール","ハイチ オール c ホワイ ティア","ハイチ オール c ホワイ ティア 40錠 第3類 医薬品","第3類 医薬品 ハイチ オール c ホワイ ティア 40錠","ハイチ オール c ホワイ ティア 120錠 エスエス製薬 エスエス製薬 しみ そばかす 全
身 倦怠 二日 酔 第3類 医薬品","エスエス製薬 ハイチ オール c ホワイ ティア 120錠","エスエス製薬 ハイチ オール c ホワイ ティア 40錠","ハイチ オール c ホワイ ティアx セット","ハイチ オール c ホワイ ティア 120錠 エスエス製薬","ハ
イチ オール c ホワイ ティア 40錠x そばかす しみに 効く 薬","定形外 ハイチ オール c ホワイ ティア 40錠","ハイチ オール c ホワイ ティア 40錠x しみ 色素 沈着に 効く","エスエス製 薬 株式 会社 ハイチ オール c ホワイ ティア 120
錠","ハイチ オール c ホワイ ティア 40錠 入","ハイチ オール c ホワイ ティア 120錠 シミ そばかす セット 第3類 医薬品","ハイチ オール c ホワイ ティア 40錠 種別 b","t 相当 エスエス製薬 株式 会社 ハイチ オール c ホワイ ティア 120
錠","ハイチ オール ホワイ ティア 120錠","ハイチ オール c ホワイ ティア 120錠 第3類 医薬品","ハイチ オール c ホワイ ティア 120錠 セット","薬 エスエス製薬 ハイチ オール c ホワイ ティア 120錠","ハイチ オール c ホワイ ティア 120
錠入り× 箱 しみ ソバカス","ハイチ オール c ホワイ ティア 120錠x しみ そばかすの 改善薬","ハイチ オール c ホワイ ティア 120錠x セット","ハイチ オール c ホワイ ティア 120錠x 3コ セット","しみ そばかす 治療 セット ハイチ オール c
ホワイ ティア 120錠 メラノcc 薬用 しみ 集中 対策 美容液","ハイチ オール c ホワイ ティア 40錠x 日 やけ 色素 沈着に","ハイチ オール c ホワイ ティア 120錠 配送 分類","ハイチ オール c ホワイ ティア 120錠 二日 酔い 疲れに 効く
薬","ハイチ オール c ホワイ ティア 120錠 他の 商品と 同時 購入は 不可","ハイチ オール c ホワイ ティア 120錠 第3類 医薬品 1点まで メール 便 発送可","ハイチ オール c ホワイ ティア ハイチ オール","ハイチ オール c ホワイ ティア し
み そばかす","ハイチ オール c ホワイ ティア 120錠 入","エスエス ハイチ オール c ホワイ ティア 40錠","第3類 医薬品 ハイチ オール c ホワイ ティア 120錠","ハイチ オール ハイチ オール c ホワイ ティア ハイチ オール c ホワイ ティ
ア","ハイチ オール c ホワイ ティア 120錠 箱 セット エスエス製薬 エスエス製薬 しみ そばかす 全身 倦怠 二日 酔 第3類 医薬品","ハイチ オール c ホワイ ティア 40錠 エスエス製薬 しみ そばかす 全身 倦怠 二日 酔 第3類 医薬品","t 相当
エスエス製薬 株式 会社 ハイチ オール c ホワイ ティア 120錠x セット","ハイチ オール c ホワイ ティア 120錠x","ハイチ オール c ホワイ ティア 40錠 第3類 医薬品 3点まで メール 便 発送可","定型外 郵便 代引き 時間 指定 できません エス
エス製薬 激安 ハイチ オール c ホワイ ティア 120錠","ハイチ オール c ホワイ ティア 120錠x しみ そばかすの 治療薬","ハイチ オール c ホワイ ティア 120錠 種別 b","ハイチ オール c ホワイ ティア 40錠x しみ そばかすに 効く","ハイチ
オール c ホワイ ティア 120錠 シミ そばかす 第3類 医薬品","定型外 郵便 代引き 時間 指定 不可 激安 ハクビ c 180錠 佐藤製薬 ハイチ オール c ホワイ ティアより ビタミン 多い","ハイチ オール c ホワイ ティア 40錠 シミ そばかす セット 第
3類 医薬品","ハイチ オール c ホワイ ティア 40錠 にきび そばかすの 治療薬","ハイチ オール ビタミン剤 第3類 ハイチ オール c ホワイ ティア","ハイチ オール c ホワイ ティア 40錠","ハイチ オール c ホワイ ティア 120錠 栄養
剤"],"pcode":["c100mg"],"price":["800.0","16879.0"],"imgsqnt":
["1b26872fa98c40ff0fe08337","051fa101221380030fe0fa37","1dc74627ae0bb0030fe0fa00","0517f0e0a427b0030fe0fe00","1d57c6a5aecbbd0030fe0fe00","0f1fa12002be30030ff0fa37","0d17f0a0c4a3b003
0fe0fa37","1f8028fdea3440fd0fe00d37","1f19ffbef25580ff0030f737","0517f0d02273b0030fa0fe00","1d17f6e3f4cff0fe0030fa37","1f187ebe189400ff0fe0f600","1fc345768703b0030fe0ff37","1c17bea
1dad250ff0fe0fb37","1f1ffebef21520fe0030e337","1f187ebe721460fe0ff00337","1f1ff50450bd20fb0030f737","1dd741e3a3a5b0fe0030ff37","0517f0d0a463b0030fe0fa00","15d741f2e3a3b0fe0030ff37"
,"0f1fa12052ba30030fb0f737","0f1fb1200a33800301b01f37","1b1e77b633c66e0fe0fa0e300","0537a155aa50e0030fe0fa00","0f1fa12102ba30030fe0ff37","1b3f75b637f7a0fe0fa0e323","0d17f0e0a4a73003
0fe0fa37","1f1fb0e4ad90b0030fe0ff37","133ff7bd92a6809b09f0ad00","1f1efebef25da0fe0030ff37","1937a1a9eda0f0030fe0fa37","1f1ffebef255a0fb0030f737","155661e3c64360030fe0fa00","0d47d6a
2a40bb0fe0030b300","1f19febe721d00ff0fe00337","1d8af5b2ec4ea0030fa0fe00","1b3f75b437f7a0fe0fa0e323","131e77f633a2e0fe0fa0e300","1dd64521c603b0fe0030fa00","0517f0d0a46330030ff0fb00"
,"133f75be33a6e0fe0fa0e323","1dc74627ae0bb0fe0030fa00","1b1e35b63386e0fe0fa0e300","1b1542f453c5e0fe0ff00337","1dc74627ae0bb0030fa0fe00","041fa1012213a0030ff0fe37","1e17b0d9ed4c50ff
0fe00337","1f1fb12052be30c30fb0f737","155661e3c643e0030fe0fa00","1dc6046fdc0970fe0fa0e323","1f8858e162bd40fb0f70f337","131f7fbe23f3b0fe0fa0e300","1b0958f122bd40ff0fb0f737","051fa14
1221380030fe0ff37","0517f0d0a463b0030ff0fb00","1c1760c15c4940fe0030fa37","1b3f37b60b9760fe0fa0e323","1f2f79b64fdfb0fe0fa0e323","1c17a0e0d49240fe0030fa37","1b19e6b45bddc0fe0ff00337"
,"051fa1412213800301b0f711","1e17f0e0fca250ff0fe01b11","1b3e77b62f9760fe0fa0e300","131e77b633e2e0fe0fa0e300","051fa1412213800301b0fa37","1e176e1a0fc8ba0fa0030fe23","1949cd0740427 0fa
0fe0f600","0d47d626a40bb0fe0030b300","051fa1412213800301b0ff00","1f1fa13452bca0030ff0f737","1f1fa12052be30c30ff0fb37","1f2f79b64b9f30fe0fa0e323","1c86c66b877f20fa0fe0e30c","1d17d6e
7b9c330fe0030ff37","1f1fa165129ca0030fa0fe47","155c6161d64320fe0030fa00","1f0cffbe5d7c0030ff0fe37","0d07d6a6a40b90fe0030b300","1f1876be721460ff0030fc37","1f0bb6f763052030fe03237"
,"1f1fa1205ab0300309d09e37","1b09e6deeb4050030fe0ff47","133f77be33e6e0fe0fa0e323","1d8346a2accbf0fe0030b300","1dc74567af83b0fe0030ff37","1236a78b8da050ff0fe0fb37","1dd64523c603f003
0fe0fa00","1b0c7efef05d10fe0fb0fa37","1d17f6e3f4cff0fe0030ff37","051fa121221300030fe0fa37","131a7bb643f7b0fe0fa0e300","1b1d77be4b7aa0fe0fa00323","1b1df7be4b78b0fa0fe00332","1f1dffb
ff25580ff0030f737","1b8064ee3e86400f0ff0fb12","0b1bfbbe729da0fe0030e337","051fa121029a30030ff0fb11","1937a1a1eda0f0030fe0fa37","1d834622accbf0fe0030b300","051fa0592273c0030fe07300"
,"1dd743e3a18e30fe0030ff37","1b0958f122bd40ff0fb0fa37","1d8346a2accbb0fe0030b300","0c127ed25474e00030270831a","1f1ffbef3752ff0030fb37","1d17f6e3f4cfd0fe0030fa37","1b1e77b63386e0fc
0fa0e300","3c21813e55b230af0fe02f37","1d17f6e3f4cfd0fe0030ff37","1f1fa12052be30030ff0fe37","1b1a7bb643f7b0fe0fa0e300","19e2943fac4400fe0030ff37"],"wonbui
d":["5576905586"],"ahash":
["ffb5110d45010101","ffbd1d0d45010101","ffff809391808087","7b7f80c5c1c1c080","7b7f80c1c0c0c0c0","ffc7c3c3c3c3c3fc","03d7c3e3e3e3431c","ffff100e64000000","ffff101d44000000","fb7f80c
1c0c0c080","c7c3c3c3e3e3633c","ffff100d64000000","f7d7c3c3c3c3431d","797f80858180809","ffbd1f0f65010000","00c3c3c3c3c3c3c3","7f7f80c5c0c0c080","fbc381818181815e","ffffff1000000bd
","ffc3c3c3c3c3c3e3","ffff030303030183","ff3e3c7e00089891","fff0f0f0900000c0","1c1c0f8fcb89cfff","ffff1f0f67010000","fd8c8c8d3cdcfbd7","fdfff0f0000000f0","ffffc0a9a0808f8","ffdfc3
e3c3e3633c","ffff101d64010000","f7d7c3c3c3c3c31d","dfff0f0f0808080f","c7c3c3c3c3c3c39c","9f9fc3c3e3634383","ffff101d64000000","ffbd010d61010060","ffbd1f0f45010000","7f7f80c1c0c0c08
```

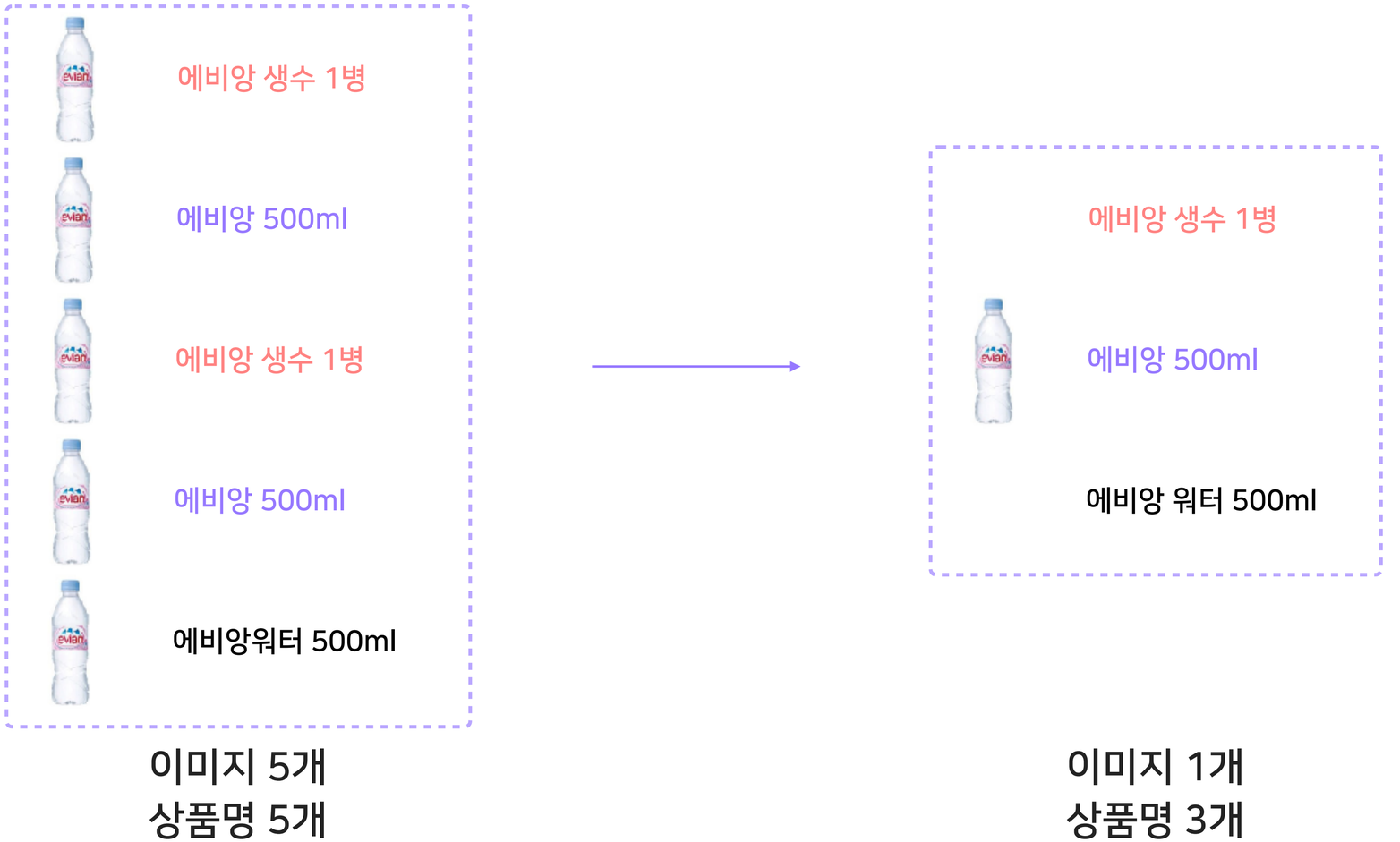

상품 ID, 상품명, 브랜드, 가격, 각종 이미지 해시, 각종 텍스트 해시, …

7.1 대형 클러스터 간 머지

	클러스터 A	클러스터 B
상품 고유 id 리스트	["1935523501","3610001185","1940202154","1945044688","6915753094","6814744228","4155170327","6814726932","4087525877","4152013795","4154609593","6971451629","6971363710","4152307849","3106606348","3106356882"	["1968128520","5816107257","2375744242","5923748586","4136383454","4014138015","4117079876","4100149563","4030532577","6535897027","4098161292","2173141145","4030924456","1966152727","1968128518","2151484199","4191843929","4029551859","2151800501","2151484195","4216975537","5928379883","4025389362","2172796087","4021150999","2181313746"
상품명 리스트	["4904776542619タケヤ化学工業 ディスペンサー ママ クラブ 100cc キャップ付","05- 0150- ママ クラブ 00cc","ママ クラブ キャップ付 ディスペンサー レッド","タケヤ化学工業-takeyakagaku- ママ クラブ","619- キャップ付 ディスペンサー ママ クラブ","キャップ付 ディ	["tokyo comedy キャバレー と とボ イとユ ジ バラエティ","tokyo comedy キャバレー ~ と とボ イとユ ジ~ vol.","tokyo comedy キャバレー ~ と とボ イとユ ジ~ vol.","tokyo comedy キャバレー ~ と とボ イとユ ジ~ vol.","tokyo comedy キャバレー ~ と とボ イとユ ジ~ vol. バラエティ","バラエ
이미지 해시 1 리스트	["051da1285a9a30030a30b308","041e292e62ab300302303337","0e1e4eb6d584300309909500","03562c274e8b500309308300","041e492a6aaa30030a308008","041e092a6aaa10030a30b312","0c2dd2aefbce009209308237","041e092a7a8a30030a302308","042dc04dc409e00309308347","07e82636734ce00303302308","06	["1f1fa1b0d49db00300a00900","1fc7abb3ae81800e00a00c00","1bc22930fc89a00a01400900","1bc20920fc89a00e00c00a00","1bc32930fc898000a01401300","1fc7abb3ae81800a03001400","041fa1412253800300a0d337","314ec5392bf1400a01301400","1fc7abb3ae81800a01402000","354ec5392bf1401400a02000","0e1fa1212211a00300a00e00","041fa141225380030730
이미지 해시 2 리스트	["ef8683ffff87ffff","0c60606078787efe","9fbfbf97030387ff","bf93037b83837b83","003fbfa7ffffc10","fffff3e303030383","0000007c7c7c7c3c","9f9f83fb8783ff83","efc683fbff838383","ffffe7e7fffff8f8","9f93837b8383bb83","003fbfa7ffffc34","0000fff7ffe7e761","9f93837b8383b983","000018f8fffffbe9","0000bfffffffb898",	["fdb18383efcd8100","8f0f0f0e0f2befef","ffff8f9380c0f1f9","870300d0e5ebd8381","f7e38585000881d9","fffdffe180c1c301","0000449f9fe3efff","c7c7c7ffbf0081f3","0000010587b1d7ff","e0e0e0c080a0c0ff","fff1ffc3c3370000","0000000f8fbfffff","0000010585b1d7ff","ffff818581e7c700","07071f1f3fefcb01","87030d0e7efd8381","0
Universal product Code 리스트	["4987067813902","4987067429202","4987067804108","4987067804306","4987067219001"]	[]
브랜드 리스트	["ナチュラグラッセ"]	["ortofon","オルトフォン"]
상품명 해시 1 리스트	["39a50e6c0d93209501701308","1f1fa120d6be300307307200","02424278c78210e30d30e708","1ba88f9202e710130030f32e","129a389087c5e0810960922e","1be8569fbb4190720760732e","041e292a6aaa100301701347","041fa1092a11900301307300","0e1ef1281eba300302307308","06d75cc5c7a260030f30132e","00ba	["0aa6bfa97d3670030a30002e","1e1e318f0f70e0030fe02f00","1b380f13e27420030000a041","0a1fb97f03dd30030a302337","1d16f1880f72e00302f09d00","1f7e19124aa4b0030a300037","0a2e39f9065190030930a32e","03043d711a53c0030000a037","1e16319e2f70e0030a002f00","083e218f2f78c00302b02300","0ea2997be6173
...	...	...

총 비교 횟수 = 각 클러스터 상품 간 비교 x 클러스터 내 속성 개수

7.1 대형 클러스터 간 머지



이미지 5개
상품명 5개

이미지 1개
상품명 3개

유일한 원소(unique key)의 개수 = Cardinality

7.2 유일한 원소의 개수, Cardinality

클러스터 A		클러스터 B		클러스터 A U B
[ABC, DEF, JHI, JKL]	+	[ABC, XYZ]	=	[ABC, DEF, JHI, JKL, XYZ]
$Cardinality_{img_A} = 4$		$Cardinality_{img_B} = 2$		$Cardinality_{img_A \cup img_B} = 5$

$$Cardinality_{img_A} + Cardinality_{img_B} - Cardinality_{img_A \cup img_B} = 4 + 2 - 5 = 1$$

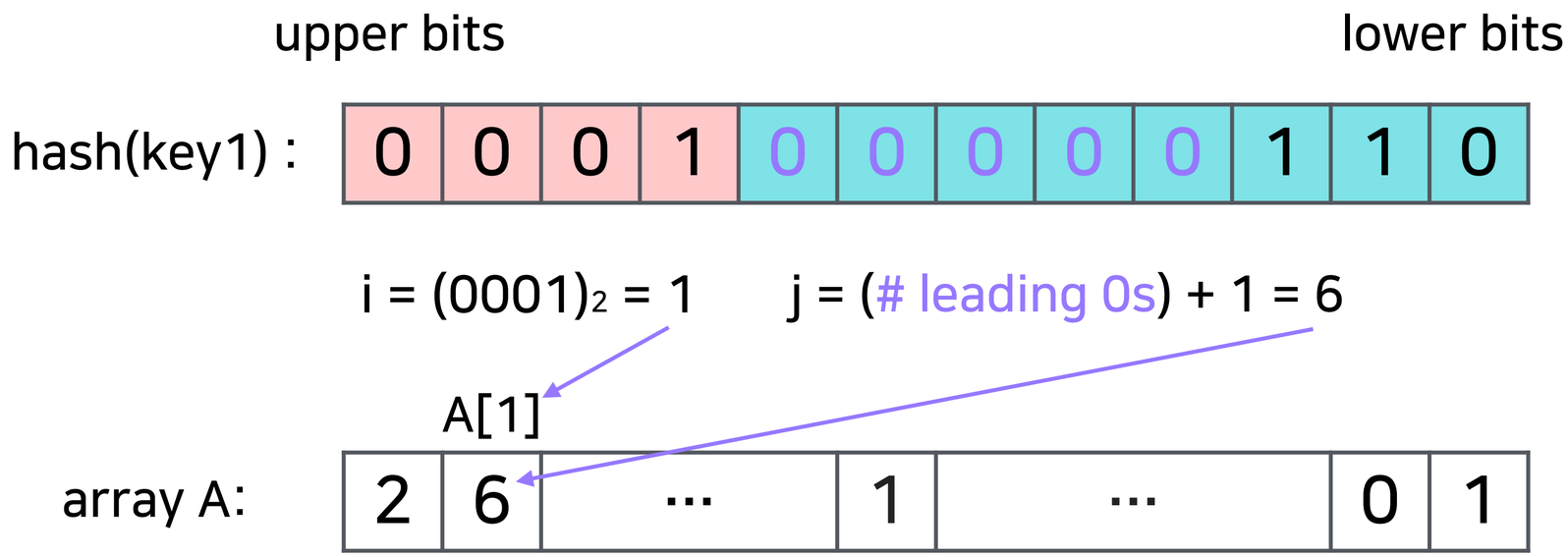
1 이상이면 공통된 정보를 가지고 있다는 의미

Cardinality를 비교하면 '중복 여부'를 판단 할 수 있음

데이터를 보지 않고 Cardinality만 알 순 없을까?

7.3 HyperLogLog

1. Set i to upper bits
2. Set A[i] to max(j, A[i])



3. Estimate # unique items(Cardinality) from $E = \frac{1}{\sum 2^{-A[i]}}$

Philippe Flajolet and Éric Fusy and Olivier Gandouet and Frédéric Meunier, "HyperLoglog: the analysis of a near-optimal cardinality estimation algorithm", DMTCS Proceedings, 2007

Stefan Heule, Marc Nunkesser, Alex Hall. HyperLogLog in Practice: Algorithmic Engineering of a State of The Art Cardinality Estimation Algorithm. Proceedings of the EDBT 2013 Conference, ACM

7.3 HyperLogLog

DEVIEW 2021

클러스터 A

[ABC, DEF, JHI, JKL]

$Cardinality_{img_A} = 4$

+

클러스터 B

[ABC, XYZ]

$Cardinality_{img_B} = 2$

=

클러스터 A U B

[ABC, DEF, JHI, JKL, XYZ]

$Cardinality_{img_A \cup img_B} = 5$

$$Cardinality_{img_A} + Cardinality_{img_B} - Cardinality_{img_A \cup img_B} = 4 + 2 - 5 = 1$$

7.3 HyperLogLog

클러스터 A

HLL_A

$Cardinality(HLL_A) = 4$

+

클러스터 B

HLL_B

$Cardinality(HLL_B) = 2$

=

클러스터 A U B

$HLL_A + HLL_B$

$Cardinality(HLL_A + HLL_B) = 5$

$$Cardinality(HLL_A) + Cardinality(HLL_B) - Cardinality(HLL_A + HLL_B) = 1$$

7.4 HyperLogLog 머지

	클러스터 A	클러스터 B
상품 id	["1935523501","3610001185","1940202154","1945044688","6915753094","6814744228","4155170327","6814726932","4087525877","4152013795","4154609593","6971451629","6971363710","4152307849","3106606348","3106356882","3106509176","6828811177","2036901090	["1968128520","5816107257","2375744242","5923748586","4136383454","4014138015","4117079876","4100149563","4030532577","6535897027","4098161292","2173141145","4030924456","1966152727","1968128518","2151484199","4191843929","4029551859","2151800501","2151484195","4216975537","5928379883","4025389362","2172796080","3158993462","217
상품명	["4904776542619タケヤ化学工業 ディスペンサー ママ クラブ 100cc キャップ付","05- 0150- ママ クラブ 00cc","ママ クラブ キャップ付 ディスペンサー レッド","タケヤ化学工業- takeyakagaku- ママ クラブ","619- キャップ付 ディスペンサー	["　　tokyo comedy キャバレ　　と　とボ　イとユ　ジ バラエティ","tokyo comedy キャバレ　~　と　とボ　イとユ　ジ~ vol.　　","　　tokyo comedy キャバレ　~　と　とボ　イとユ　ジ~ vol.","tokyo comedy キャバレ　~　と　とボ　イとユ　ジ~ vol.","tokyo comedy キャバレ　~　と　とボ
이미지 해시 1	["051da1285a9a30030a30b308","041e292e62ab300302303337","0e1e4eb6d584300309909500","03562c274e8b500309308300","041e492a6aaa30030a308008","041e092a6aaa10030a30b312","0c2dd2aefbce009209308237","041e092a7a8a30030a302308","042dc04dc409e00309308347","07e82636734	["1f1fa1b0d49db00300a00900","1fc7abb3ae81800e00a00c00","1bc22930fc89a00a01400900","1bc20920fc89a00e00c00a00","1bc32930fc89800a01401300","1fc7abb3ae81800a03001400","041fa1412253800300a0d337","314ec5392bf1400a01301400","1fc7abb3ae81800a01402000","354ec5392bf1401400a02000","0e1
이미지 해시 2	["ef8683ffff87ffff","0c60606078787efe","9fbfbf97030387ff","bf93037b83837b83","003fbfa7fffffc10","fffff3e303030383","0000007c7c7c7c3c","9f9f83fb8783ff83","efc683fbff838383","ffffe7e7fffff8f8","9f93837b8383bb83","003fbfa7fffffc34","0000fff7ffe7e761","9f93837b8383b983","000018f8fffffbe9"	["fdb18383efcd8100","8f0f0f0e0f2befef","ffff8f9380c0f1f9","87030d0e5ebd8381","f7e38585000881d9","fffdffe180c1c301","0000449f9fe3efff","c7c7c7ffbf0081f3","0000010587b1d7ff","e0e0e0c080a0c0ff","fff1ffc3c3370000","0000000f8fbfffff","0000010585b1d7ff","ffff818581e7c700","07071f
Universal product Code	["4987067813902","4987067429202","4987067804108","4987067804306","4987067219001"]	[]
브랜드	["ナチュラグラッセ"]	["ortofon","オルトフォン"]
상품명 해시 1	["39a50e6c0d93209501701308","1f1fa120d6be300307307200","02424278c78210e30d30e708","1ba88f9202e710130030f32e","129a389087c5e0810960922e","1be8569fbb4190720760732e","041e292a6aaa100301701347","041fa1092a11900301307300","0e1ef1281eba300302307308","06d75cc5c7a2	["0aa6bfa97d3670030a30002e","1e1e318f0f70e0030fe02f00","1b380f13e27420030000a041","0a1fb97f03dd30030a302337","1d16f1880f72e00302f09d00","1f7e19124aa4b0030a300037","0a2e39f9065190030930a32e","03043d711a53c0030000a037","1e16319e2f70e0030a002f00","083e218f
…	…	…

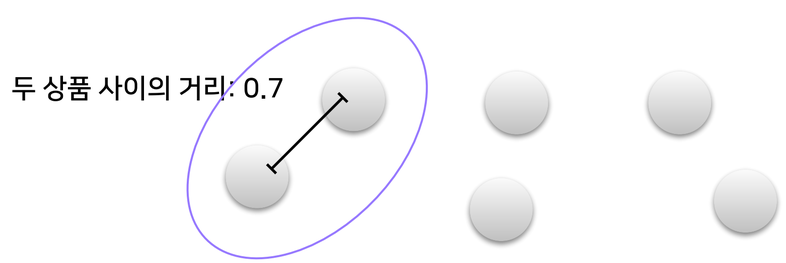
7.4 HyperLogLog 머지

DEVIEW 2021

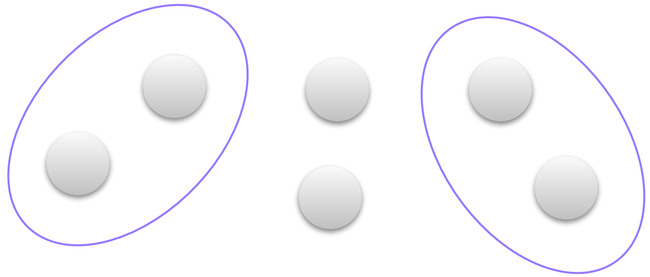
	클러스터 A	클러스터 B
상품 id	["1935523501","3610001185","1940202154","1945044688","6915753094","6814744228","4155170327","6814726932","4087525877","4152013795","4154609593","6971451629","6971363710","4152307849","3106606348","3106356882","3106509176","6828811177","2036901094	["1968128520","5816107257","2375744242","5923748586","4136383454","4014138015","4117079876","4100149563","4030532577","6535897027","4098161292","2173141145","4030924456","1966152727","1968128518","2151484199","4191843929","4029551859","2151800501","2151484195","4216975537","5928379883","4025389362","2172796080","3158993462","217
상품명	HLL_{title_A}	HLL_{title_B}
이미지 해시 1	HLL_{img1_A}	HLL_{img2_B}
이미지 해시 2	HLL_{img2_A}	HLL_{img2_B}
Universal product Code	["4987067813902","4987067429202","4987067804108","4987067804306","4987067219001"]	[]
브랜드	["ナチュラグラッセ"]	["ortofon","オルトフォン"]
상품명 해시 1	HLL_{title1_A}	HLL_{title2_B}
...	...	...

8. 오차의 전파

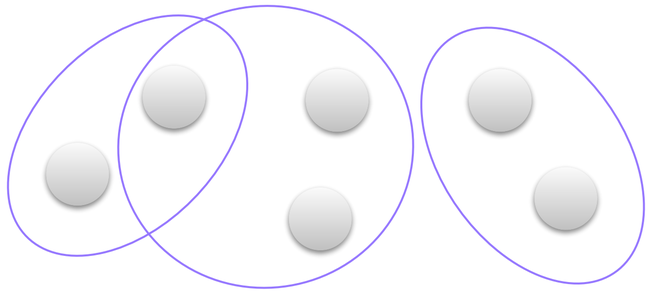
8.1 오차의 전파



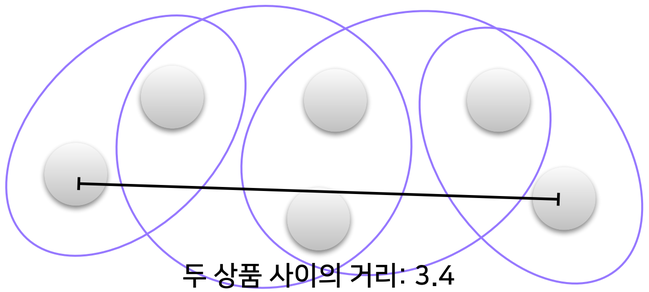
Stage 1



Stage 2



Stage 3



Stage 4

매우 강력한 기준으로 묶는다 하더라도, 오차는 점점 퍼진다

8.1 오차의 전파

Error
Backpropagation

Error
Propagation

- 다변수 함수의 오차 전파
  - 변수 x_i에 대한 입력 오차 Δx_i, 함수 $y = f(x_1, x_2, \ldots, x_n)$

➔ $y + \Delta y = f(x_1 + \Delta x_1,\ x_2 + \Delta x_2,\ \cdots,\ x_n + \Delta x_n)$

$y = f(x_1, x_2, \cdots, x_n)$을 미분하면,

$$dy = \frac{\partial f}{\partial x_1} dx_1 + \frac{\partial f}{\partial x_2} dx_2 + \cdots + \frac{\partial f}{\partial x_n} dx_n$$

오차 계산의 일반식

$$\varepsilon_y = \Delta y = \frac{\partial f}{\partial x_1} \Delta x_1 + \frac{\partial f}{\partial x_2} \Delta x_2 + \cdots + \frac{\partial f}{\partial x_n} \Delta x_n$$

➔ $$|\varepsilon_y| = \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i} \right| |\Delta x_i|$$

최대 오차

Error propagation이란 애도 있어요..

https://slidesplayer.org/slide/15463809/
https://pythonhosted.org/uncertainties/

8.2 클러스터 내의 오차를 지속적으로 모니터링

안정

$\epsilon = 0.02$

$\epsilon = 0.13$

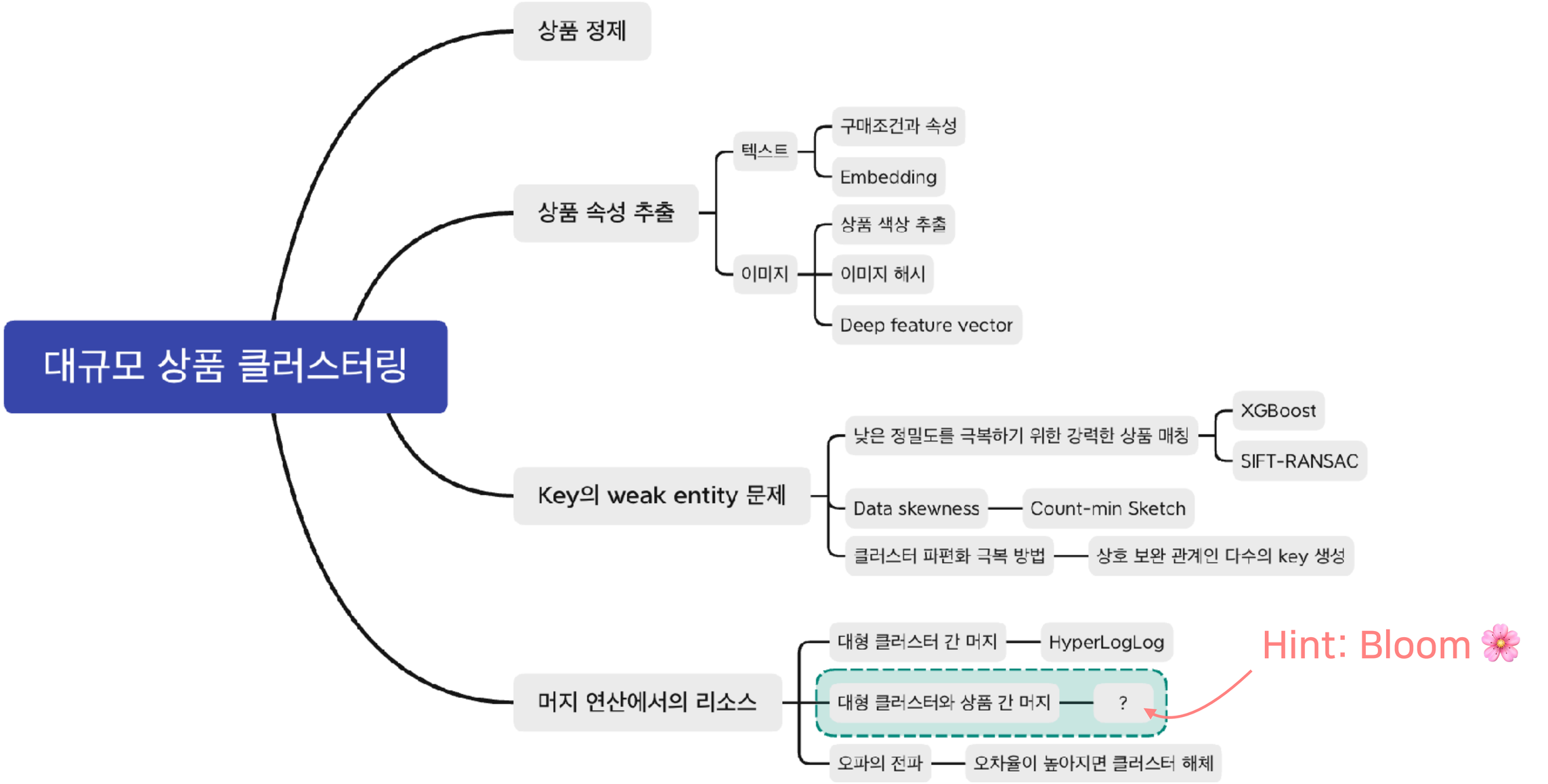
$\epsilon = 0.26$

$\epsilon = 0.42$

불안정

모든 클러스터는 자신의 혼잡도와 오차(불확도)를 지니고 있어야 함

최종 정리


대규모 상품 클러스터링
상품 정제
상품 속성 추출
텍스트
구매조건과 속성
Embedding
이미지
상품 색상 추출
이미지 해시
Deep feature vector
Key의 weak entity 문제
낮은 정밀도를 극복하기 위한 강력한 상품 매칭
XGBoost
SIFT-RANSAC
Data skewness
Count-min Sketch
클러스터 파편화 극복 방법
상호 보완 관계인 다수의 key 생성
머지 연산에서의 리소스
대형 클러스터 간 머지
HyperLogLog
대형 클러스터와 상품 간 머지
?
오파의 전파
오차율이 높아지면 클러스터 해체
Hint: Bloom 🌸

We Are Hiring!

DEVIEW 2021

Information retrieval / extraction / classification

Natural language processing / Query understanding

Machine learning / Deep learning / Data mining

Computer vision / Image processing / Object recognition / Deep tagging

Distributed computing / Large-scale system design / Large-scale data processing

seungkwon.choe@navercorp.com - Catalog Matching

Thank You