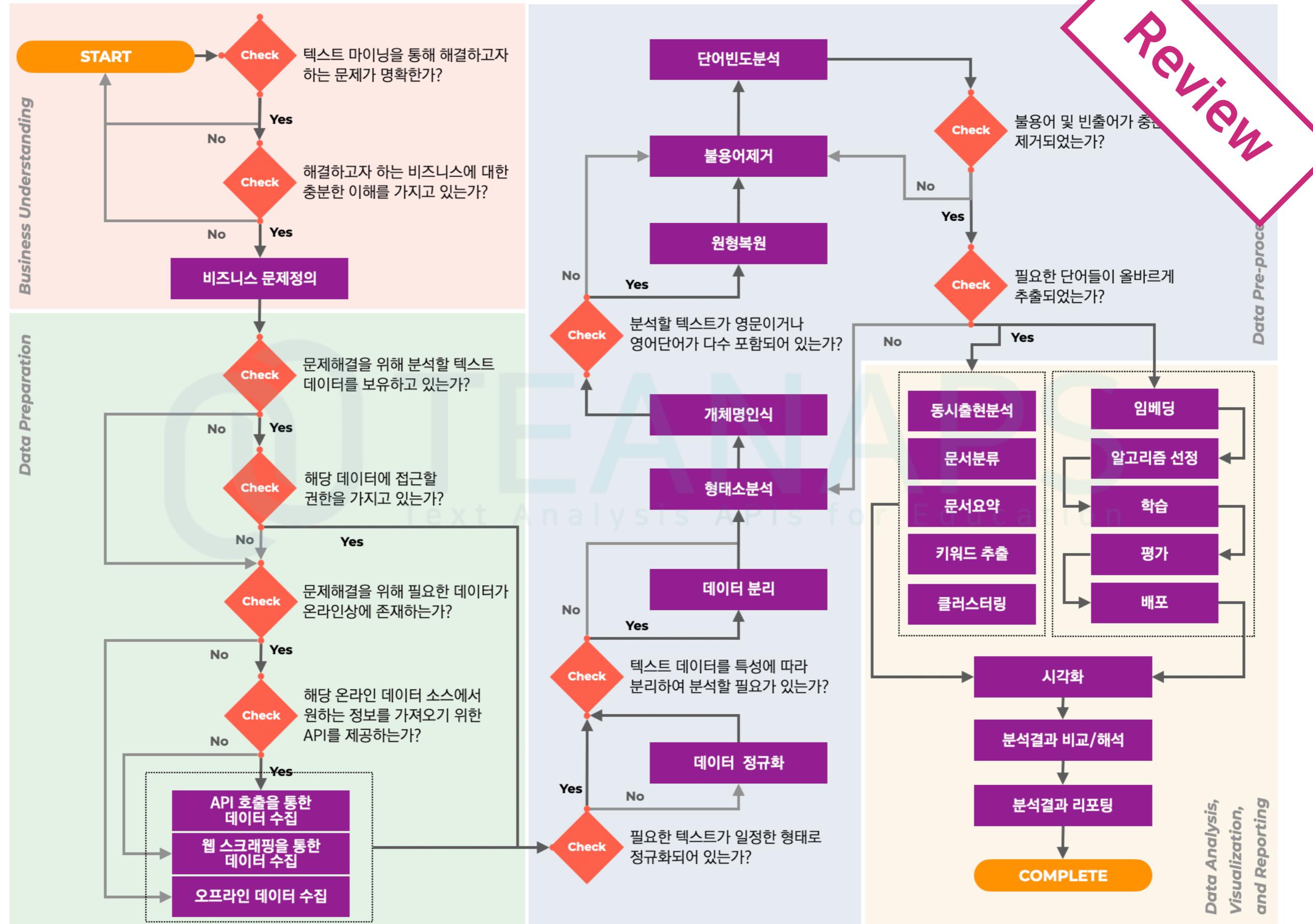


ADVANCED TEXT MINING

by FINGEREDMAN (fingeredman@gmail.com)



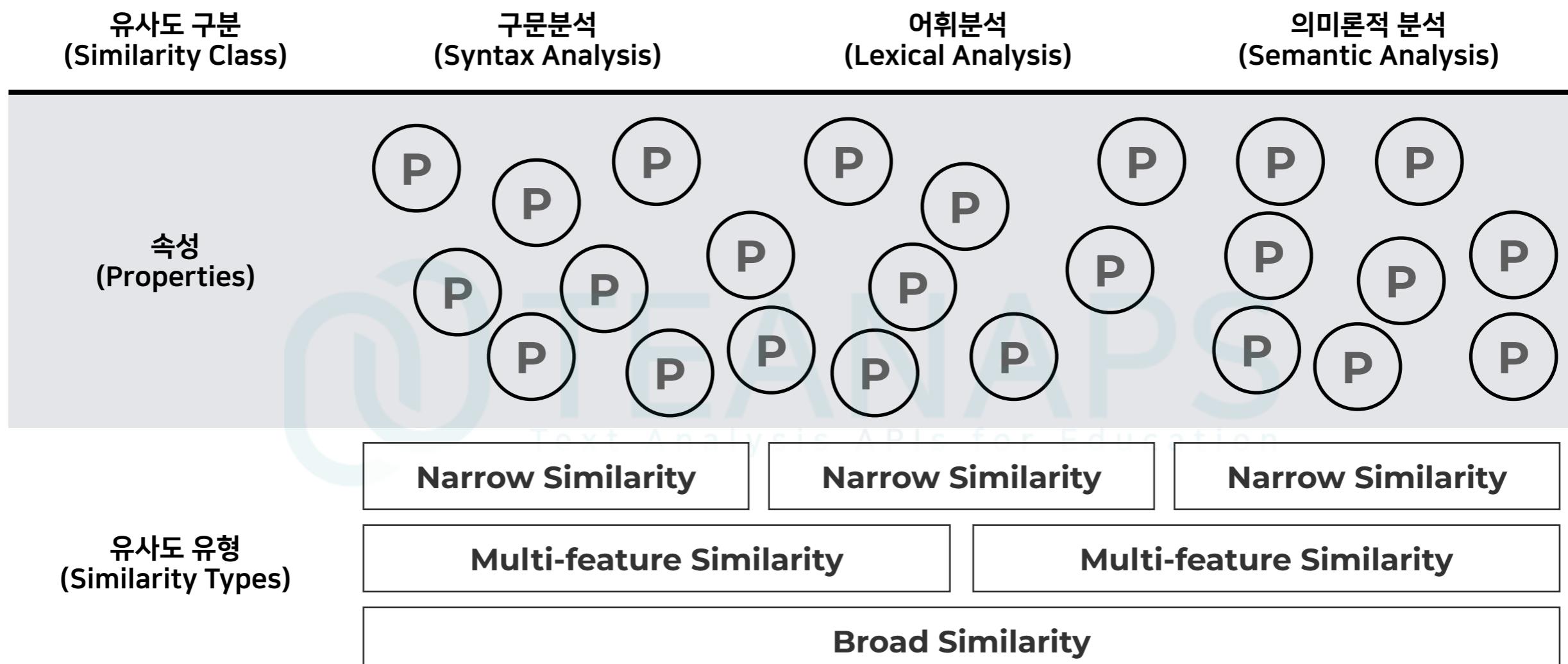
WEEK 10

Clustering

문서 유사도 (Document Similarity)

유사도 (Similarity)

- 서로 다른 두 객체 사이의 공통점을 통해 서로 공유하는 속성의 수에 따라 증가하는 유사한 정도
- 서로 공유하는 속성은 그 기능에 따라서 매우 많이 존재할 수도 있으며 없을 수도 있음



문서 유사도

(Document Similarity)

문서 간의 유사도

- 문서 간의 유사도는 텍스트 데이터 검색, 시각화, 필터링, 정렬 등 다양한 분야에 활용됨
- 문서의 유사도 측정은 문서를 벡터로 표현하고 벡터 간의 유사도를 통해 결정됨
- 유사도를 측정하는 척도(코사인 유사도, 자카드 유사도, 유클리디안 유사도)는 정해진 것이 없으며, 문헌에 따라 가장 적절한 방법을 분석자가 스스로 판단하여 결정해야함

구분	내용
원문	$D_1 = \text{텍스트 마이닝은 비정형 데이터에 대해 다룹니다.}$ $D_2 = \text{비정형 데이터는 정형 데이터에 비해 복잡하고 어렵습니다.}$ $D_3 = \text{오늘은 단어주머니에 대해 배웁니다.}$
유니그램	$D_1 = ["\text{텍스트}", "\text{마이닝}", "\text{은}", "\text{비정형}", "\text{데이터}", "\text{에}", "\text{대해}", "\text{다루}", "\text{ㅂ니다}", "."]$ $D_2 = ["\text{비정형}", "\text{데이터}", "\text{는}", "\text{정형}", "\text{데이터}", "\text{에}", "\text{비해}", "\text{복잡하}", "\text{고}", "\text{어렵}", "\text{습니다.}"]$ $D_3 = ["\text{오늘}", "\text{은}", "\text{단어주머니}", "\text{에}", "\text{대해}", "\text{배우}", "\text{ㅂ니다}", "."]$
유니그램 (불용어 제거 후)	$D_1 = ["\text{텍스트}", "\text{마이닝}", "\text{비정형}", "\text{데이터}", "\text{다루}"]$ $D_2 = ["\text{비정형}", "\text{데이터}", "\text{정형}", "\text{복잡하}", "\text{어렵}"]$ $D_3 = ["\text{오늘}", "\text{단어주머니}", "\text{배우}"]$

문서 유사도 (Document Similarity)

자카드 유사도 (Jaccard Similarity)

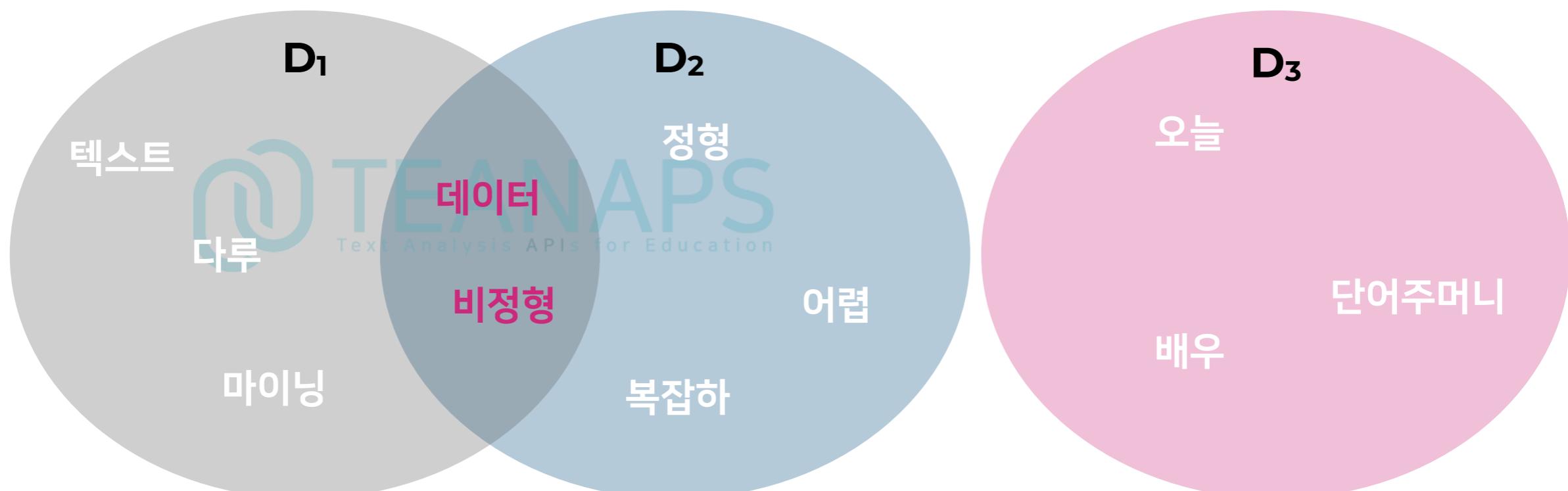
- 문서를 각각 단어의 집합_(BoW)으로 만든 후, 문서 간의 공통된 단어의 개수를 활용하여 유사도를 계산하는 방법

$$\text{Sim}_{\text{Jaccard}}(D_1, D_2) = \frac{|D_1 \cap D_2|}{|D_1 \cup D_2|} = \frac{2}{8} = 0.25 = 25\%$$

$D_1 = ["\text{텍스트}", "\text{마이닝}", "\text{비정형}", "\text{데이터}", "\text{다루"}]$

$D_2 = ["\text{비정형}", "\text{데이터}", "\text{정형}", "\text{복잡하}", "\text{어렵}"]$

$D_3 = ["\text{오늘}", "\text{단어주머니}", "\text{배우}"]$



문서 유사도

(Document Similarity)

유클리디언 유사도 (Euclidean Similarity)

- 두 벡터 사이의 최단거리를 계산하여 유사도를 계산하는 방법
- 단순히 두 벡터 사이 거리를 의미하기 때문에 값이 0보다 크며, 벡터 정규화를 통해 0~1 사이의 값을 가지도록 함

$$\text{Sim}_{\text{Euclidean}}(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2}$$

$D_1 = ["\text{텍스트}", "\text{마이닝}", "\text{비정형}", "\text{데이터}", "\text{다루}"]$

$D_2 = ["\text{비정형}", "\text{데이터}", "\text{정형}", "\text{복잡하}", "\text{어렵}"]$

$D_3 = ["\text{오늘}", "\text{단어주머니}", "\text{배우}"]$

단어	D_1	D_2	D_3	D_4	D_5
텍스트	1	0	0	0	1
마이닝	1	0	0	0	1
비정형	1	1	0	1	2
데이터	1	2	0	4	0
다루다	1	0	0	2	1
정형	0	1	0	1	1
복잡하다	0	1	0	0	1
어렵다	0	1	0	0	0
오늘	0	0	1	0	0
단어주머니	0	0	1	2	1
배우다	0	0	1	0	1

$$\text{Sim}_{\text{Euclidean}}(D_1, D_2)$$

$$\begin{aligned}
 &= \sqrt{(1 - 0)^2 + (1 - 0)^2 + (1 - 1)^2 \\
 &\quad + (1 - 2)^2 + (1 - 0)^2 + (0 - 1)^2 \\
 &\quad + (0 - 1)^2 + (0 - 1)^2 + (0 - 0)^2 \\
 &\quad + (0 - 0)^2 + (0 - 0)^2} = 7
 \end{aligned}$$

문서 유사도

(Document Similarity)

코사인 유사도 (Cosine Similarity)

- 유사도 계산에 가장 널리 쓰이는 방법으로, 두 벡터 사이의 코사인 각도를 통해 유사도를 구하는 방법

$$\text{Sim}_{\text{Cosine}}(D_1, D_2) = \frac{D_1 \cdot D_2}{|D_1| |D_2|}$$

$D_1 = ["\text{텍스트}", "\text{마이닝}", "\text{비정형}", "\text{데이터}", "\text{다루}"]$

$D_2 = ["\text{비정형}", "\text{데이터}", "\text{정형}", "\text{복잡하}", "\text{어렵}"]$

$D_3 = ["\text{오늘}", "\text{단어주머니}", "\text{배우}"]$

단어	D_1	D_2	D_3	D_4	D_5
텍스트	1	0	0	0	1
마이닝	1	0	0	0	1
비정형	1	1	0	1	2
데이터	1	2	0	4	0
다루다	1	0	0	2	1
정형	0	1	0	1	1
복잡하다	0	1	0	0	1
어렵다	0	1	0	0	0
오늘	0	0	1	0	0
단어주머니	0	0	1	2	1
배우다	0	0	1	0	1

$$\text{Sim}_{\text{Cosine}}(D_1, D_2)$$

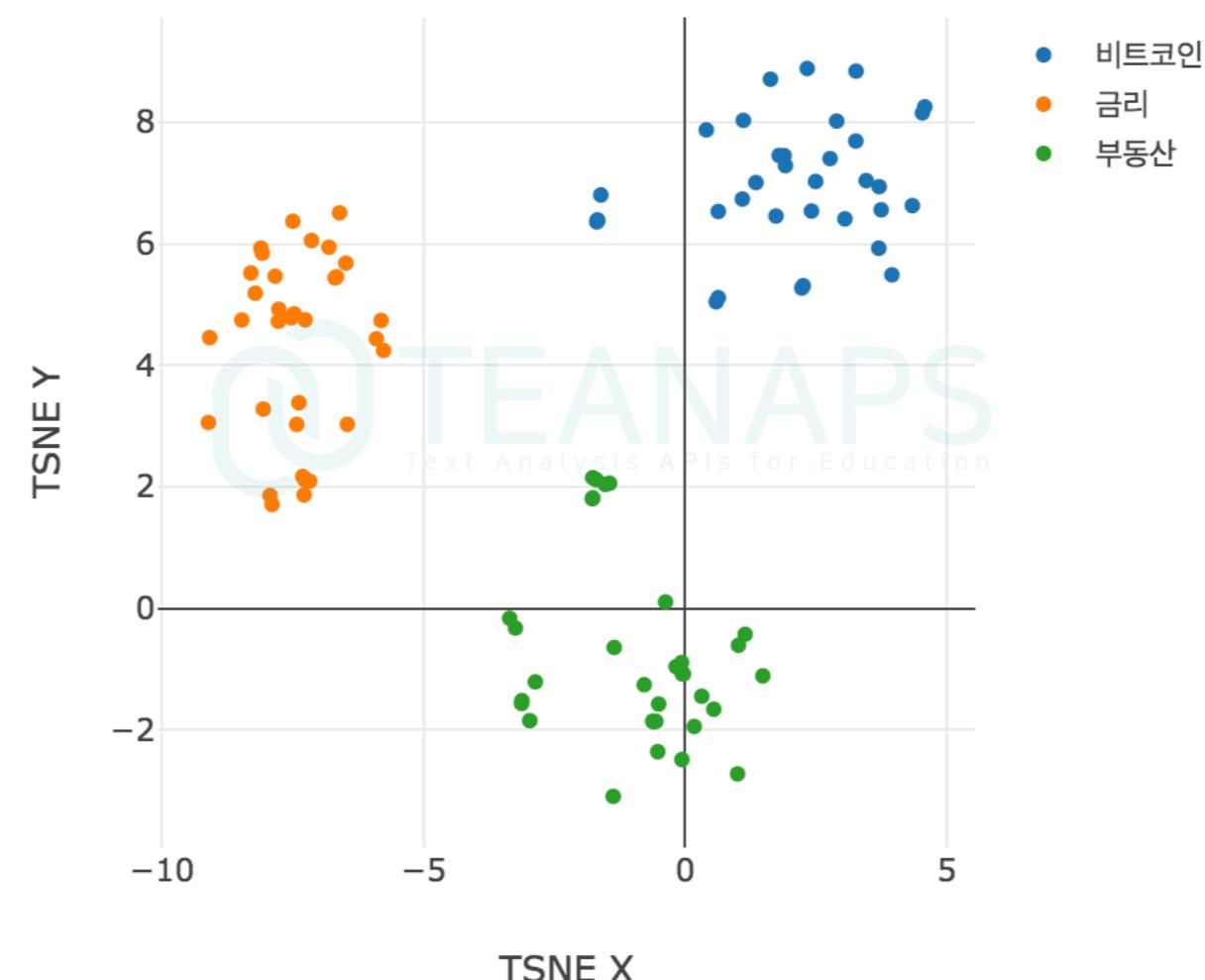
$$\begin{aligned}
 &= \sqrt{(1 - 0)^2 + (1 - 0)^2 + (1 - 1)^2 \\
 &\quad + (1 - 2)^2 + (1 - 0)^2 + (0 - 1)^2 \\
 &\quad + (0 - 1)^2 + (0 - 1)^2 + (0 - 0)^2 \\
 &\quad + (0 - 0)^2 + (0 - 0)^2} = 7
 \end{aligned}$$

문서 군집화 (Clustering)

벡터공간 모델 (Vector Space Model)

- 각 문서를 N 차원 상의 벡터공간에 표현하는 방법
- N은 문서를 표현하는 특성^(feature)의 수에 따라서 결정됨

K-Means Clustering Graph - label



기계학습 절차: 데이터 준비

Review

레이블링 (Labeling)

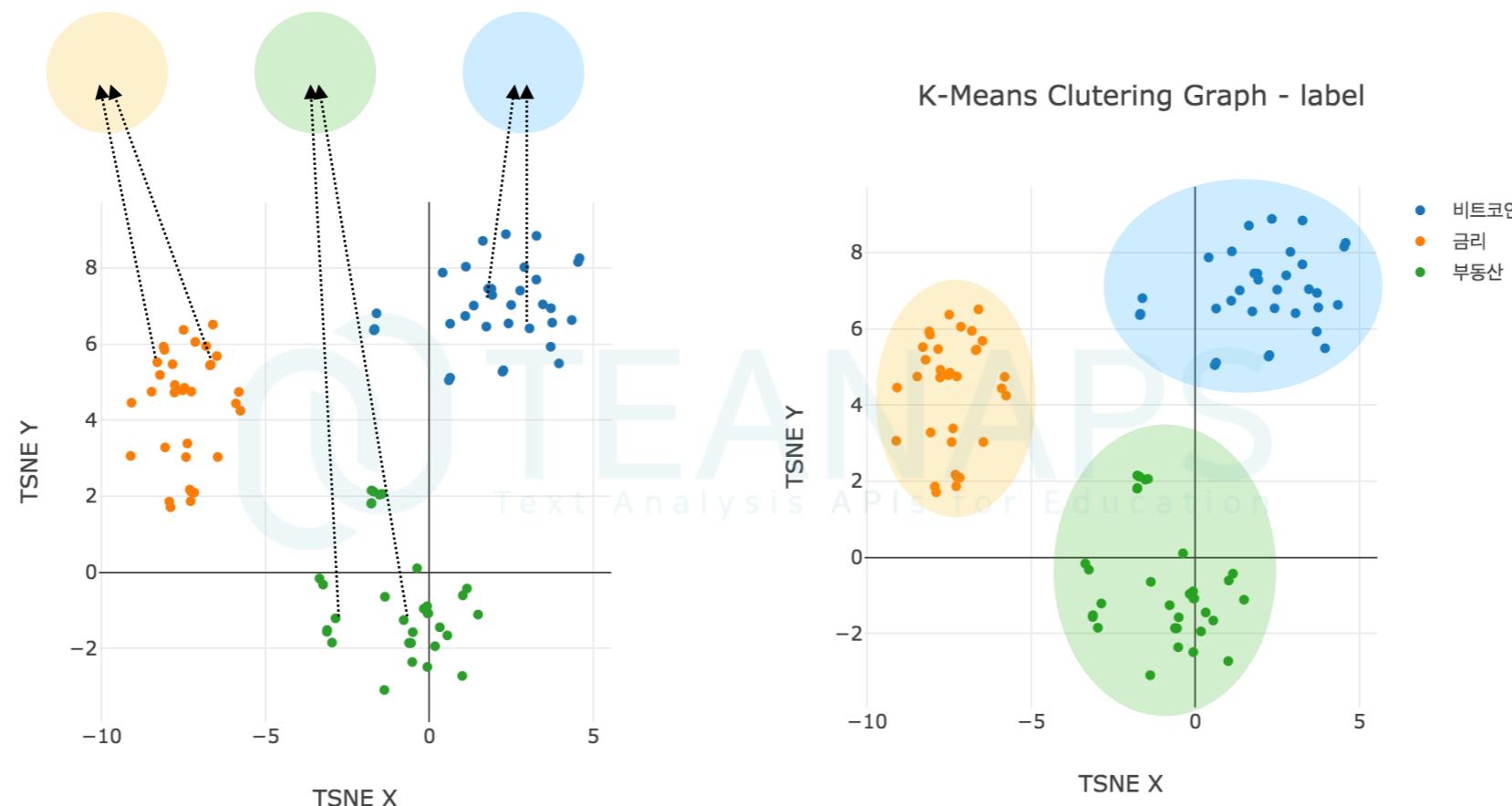
- 준비된 학습데이터에 지도학습을 위한 라벨(label)을 부착하는 과정
- 지도학습을 활용하는 경우, 학습데이터의 양과 레이블링의 정확도가 모델의 성능에 큰 영향을 미칠 수 있음
- 효율적으로 라벨을 부착하는 방법을 찾는 것도 데이터 분석 준비 과정 중 매우 중요한 요소로 작용함

구분	메시지	특징 (Feature)						라벨 (Label)
		메시지 길이	URL 여부	특수문자 개수	해외수신 여부	의심단어 개수	광고문자 표시여부	
1	[국제발신] 하루 30만 달 1천만 만원으로 이렇게! http://bit.ly/3f~	40	1	6	1	1	0	TRUE
2	팀장님 이보람 선임입니다. 출근하시면 결재 부탁드립니다.	20	0	1	0	0	0	FALSE
3	(광고)웰컴박하라 vc⑤47③.co 코드 wc1004 무료수신거부 01084510000	45	1	5	0	1	1	TRUE
4	(광고)신한과 함께하는 소중한 미래 따뜻한 금융 [신한]입니다. 2019년에 힘들었던 모든 일들은 다 잊어버리시고, ~	80	0	4	0	0	1	TRUE
5	[WEB발신] 갤럭시 노트20/노트20 울트라 사전예약 오늘이 마지막날입니다-!! 구매를 망설이고 ~	75	0	6	0	0	0	TRUE
6	(광고)등촌역스톤힐 ★더블역세권 9호선 ~ ★선착순으로 동호수 지정분양 가능 ★인근주변 아파트 시세보다 4~5억 저렴 ★~	120	0	8	0	0	1	TRUE
7	[한진택배] 상품 배송 안내 안녕하세요 고객님. ★상품 수령이 편하신 장소를 선택 ~ ①직접수령 ②경비실 ③문앞 ~	110	0	9	0	0	0	FALSE

문서 군집화 (Document Clustering)

군집화 (Clustering)

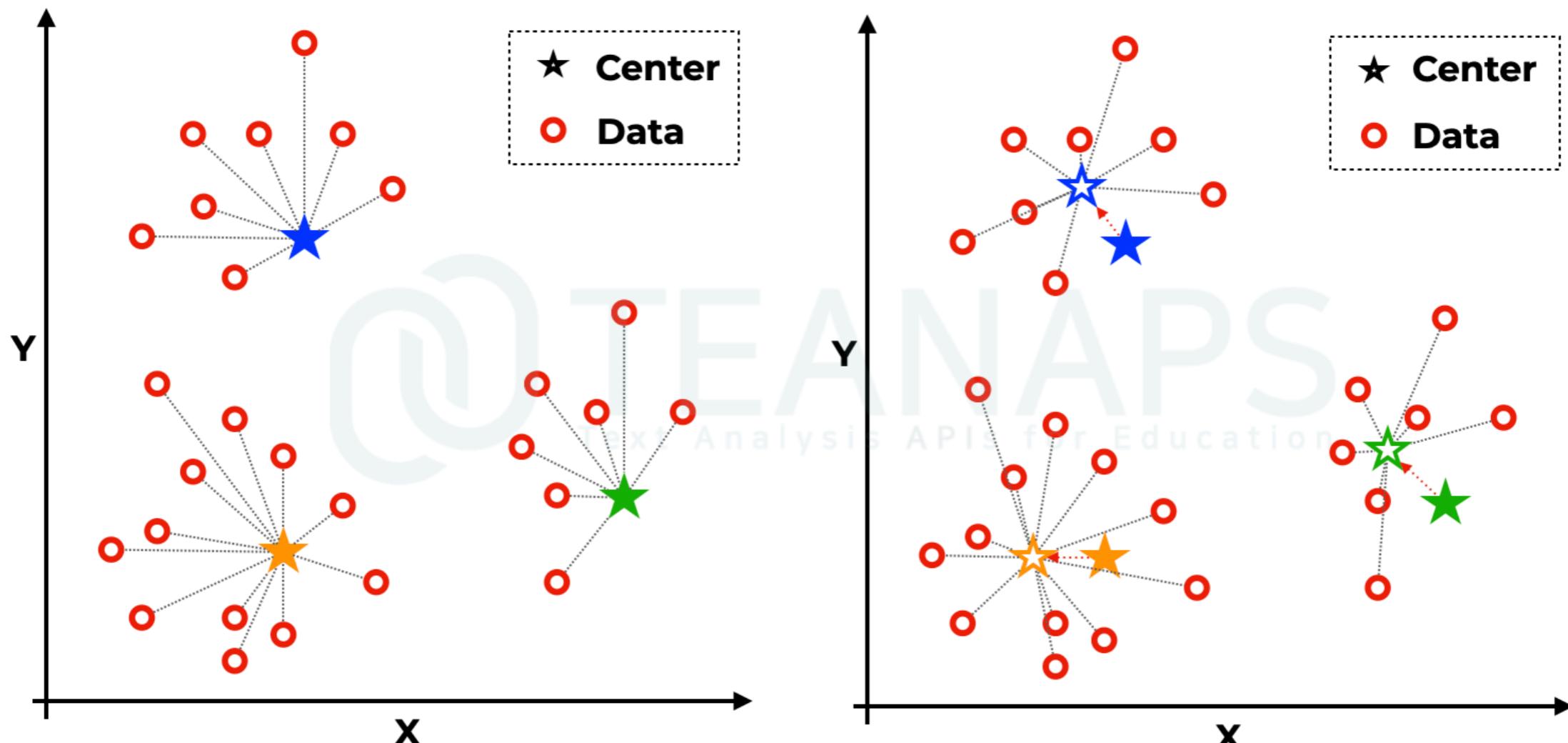
- 데이터 간에 발견되는 자연스러운 그룹(군집)을 발견하고 주제를 제시하는 분석방법
- 분류와 군집화의 차이점
 - 1) **분류** (Classification) : 문서의 집합을 이미 정해진 유형의 개수와 속성에 따라서 분류하는 방법
 - 2) **군집화** (Clustering) : 사전에 유형의 개수와 속성이 알려지지 않은 상태로 그룹을 발견하는 방법



군집화 알고리즘: K-평균

K-평균 군집화 (K-means Clustering)

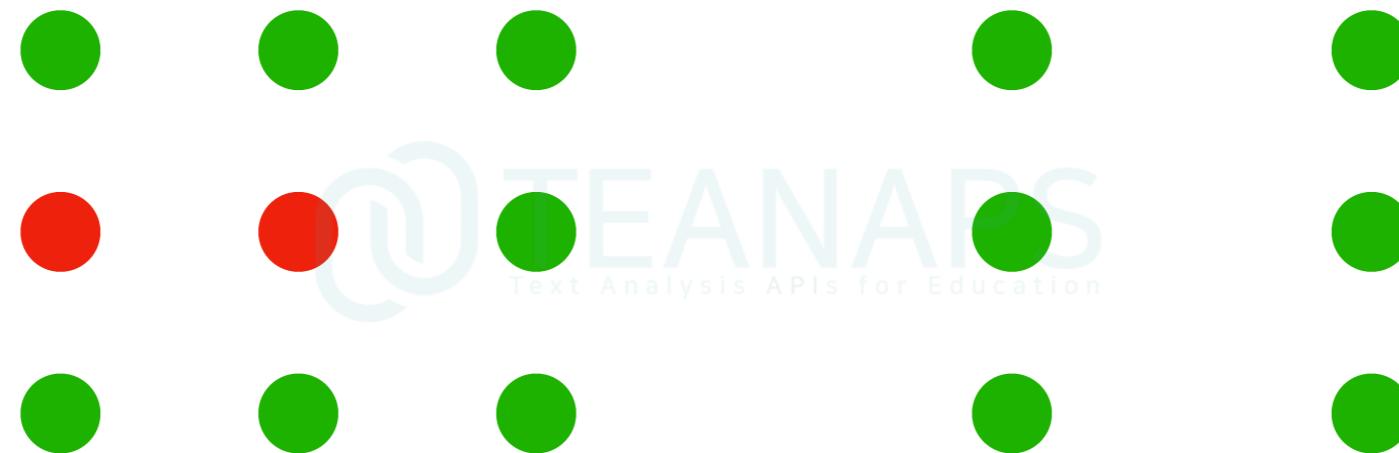
- 벡터공간 상에 K개의 가상의 중심점(centroid)을 설정한 후, 중심점 위치를 변경해가면서 중심점과 군집의 각 데이터 간의 거리를 최소로 하는 방향으로 군집을 생성하는 방법
- 알고리즘이 매우 단순하여 군집을 생성하는 속도가 매우 빠르고 좋은 성능을 보임



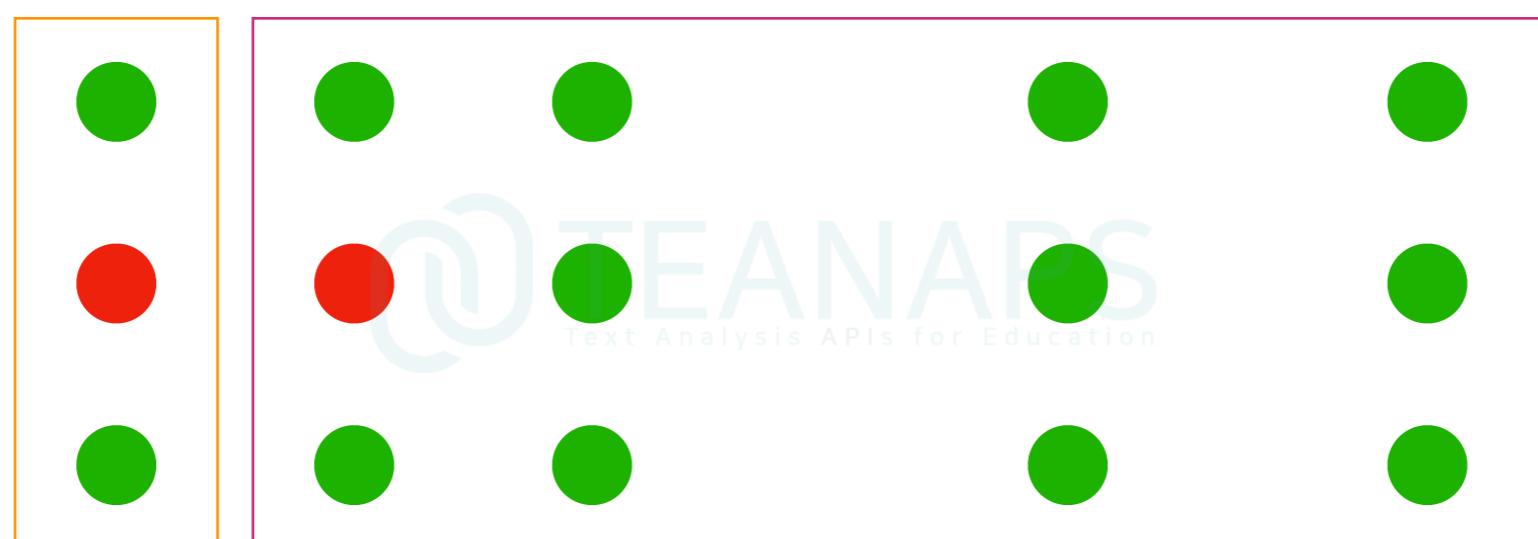
군집화 알고리즘: K-평균

K-평균 군집화 알고리즘

- Step 1 : 임의로 k개의 데이터 포인트를 시드로 선택



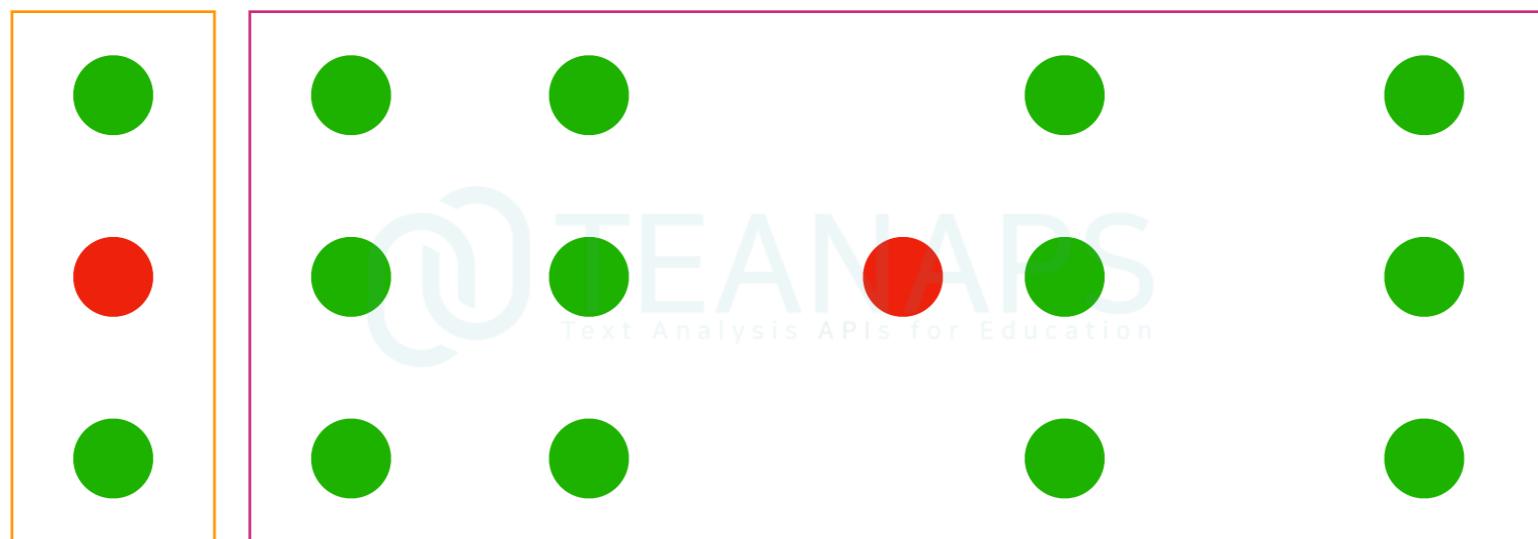
- Step 2 : 각 레코드를 가장 가까운 시드에 배정



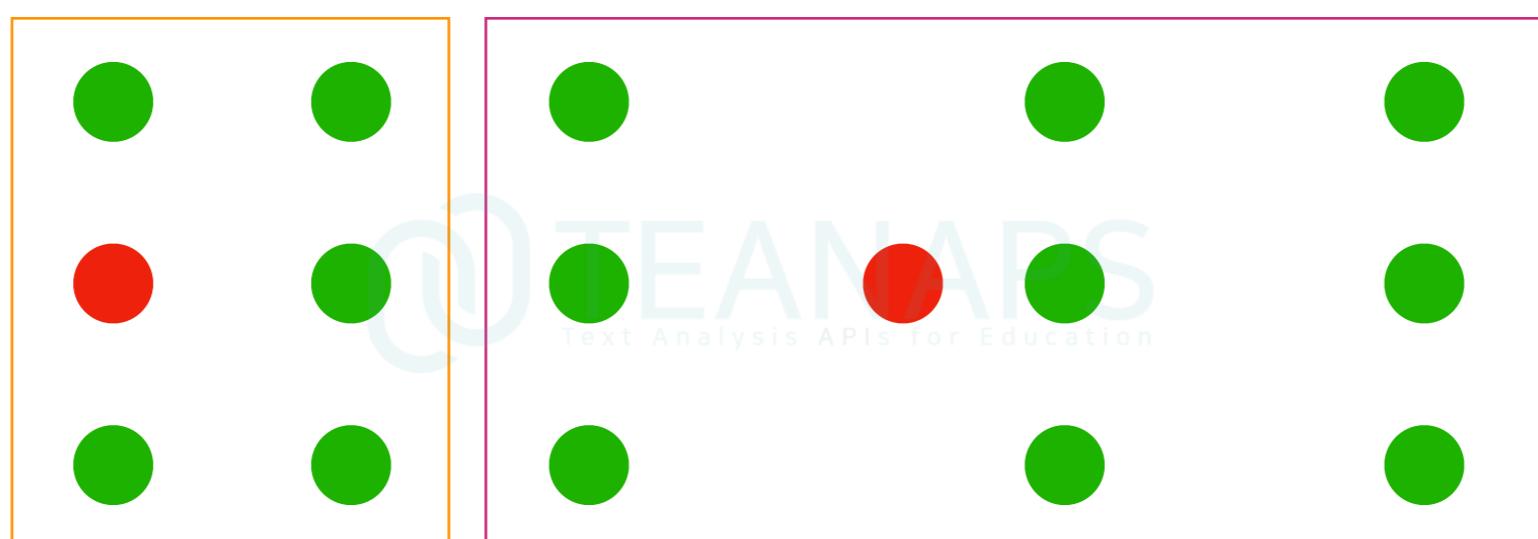
군집화 알고리즘: K-평균

K-평균 군집화 알고리즘

- Step 3 : 군집의 중심점 찾기 (1st loop)



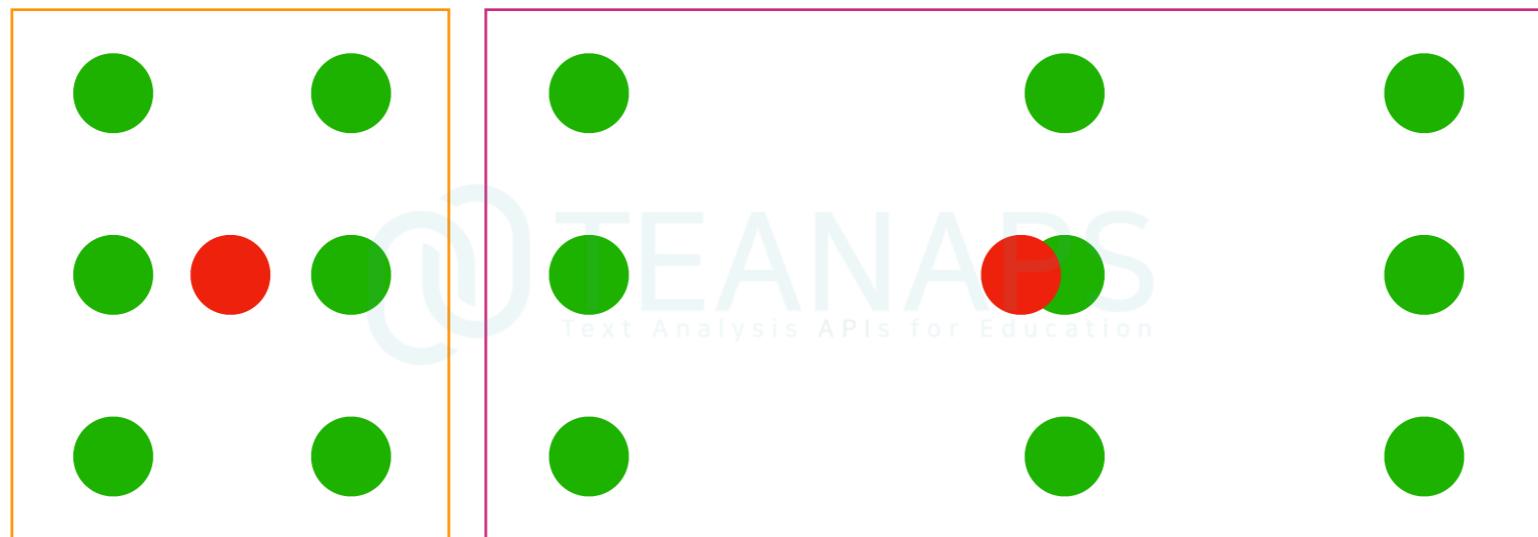
- Step 2 : 각 레코드를 가장 가까운 시드에 배정



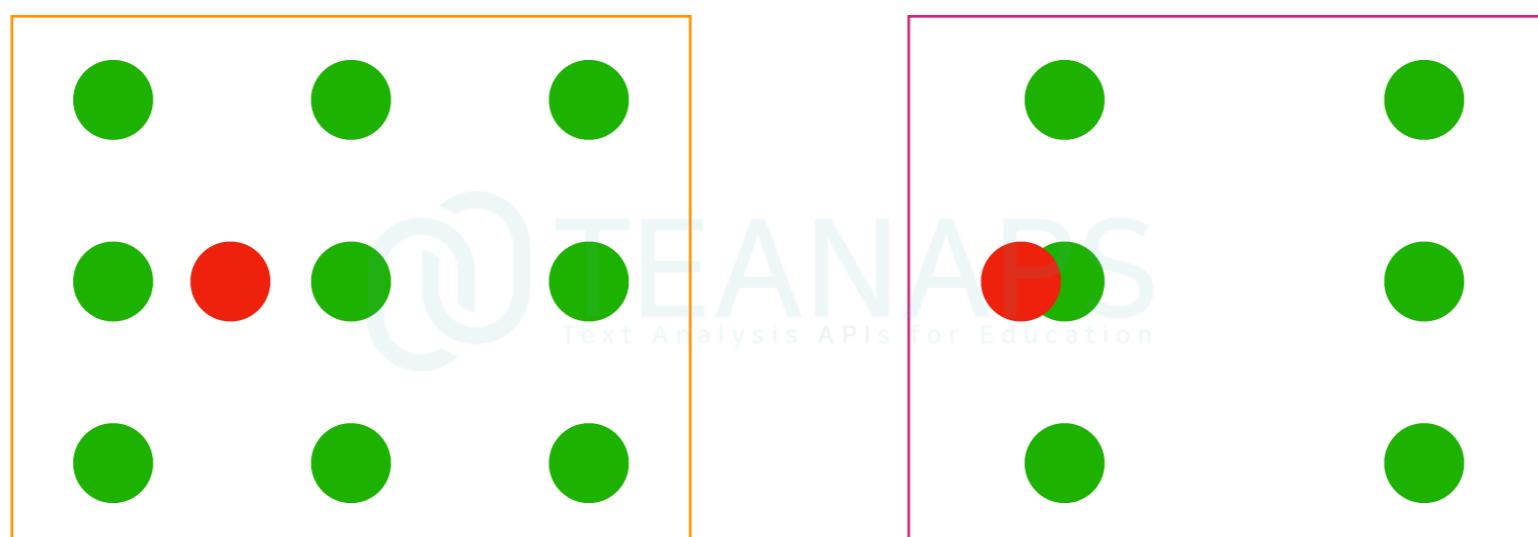
군집화 알고리즘: K-평균

K-평균 군집화 알고리즘

- Step 3 : 군집의 중심점 찾기 (2nd loop)



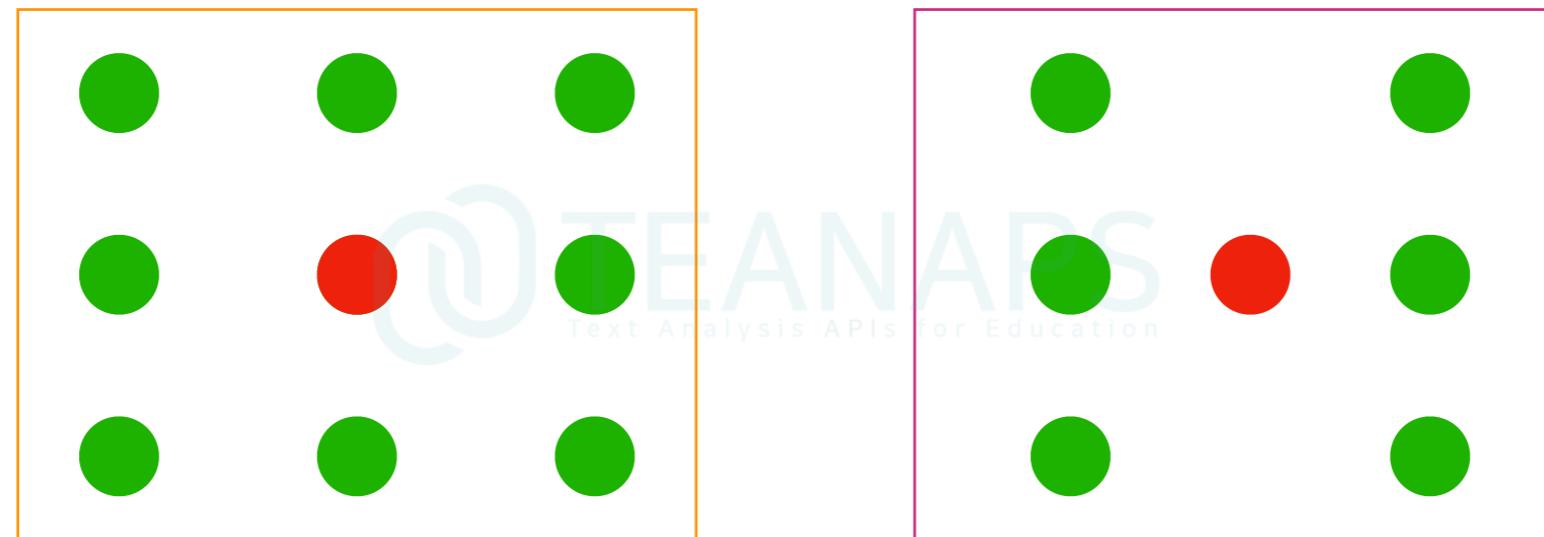
- Step 2 : 각 레코드를 가장 가까운 시드에 배정



군집화 알고리즘: K-평균

K-평균 군집화 알고리즘

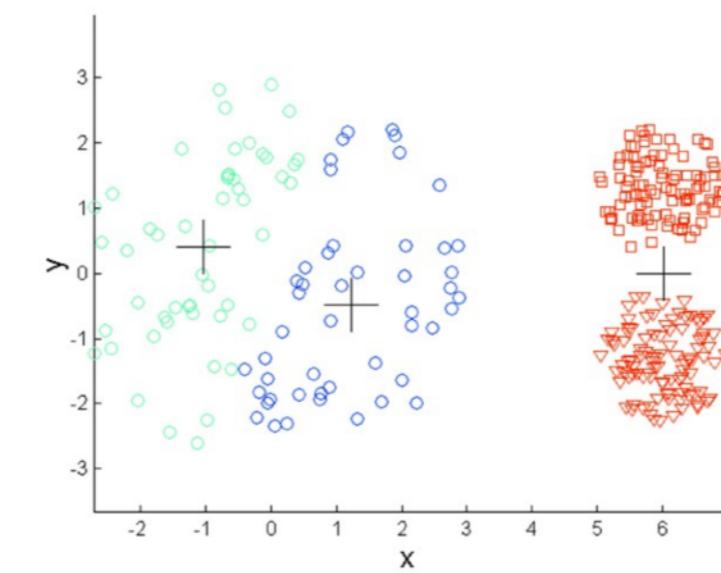
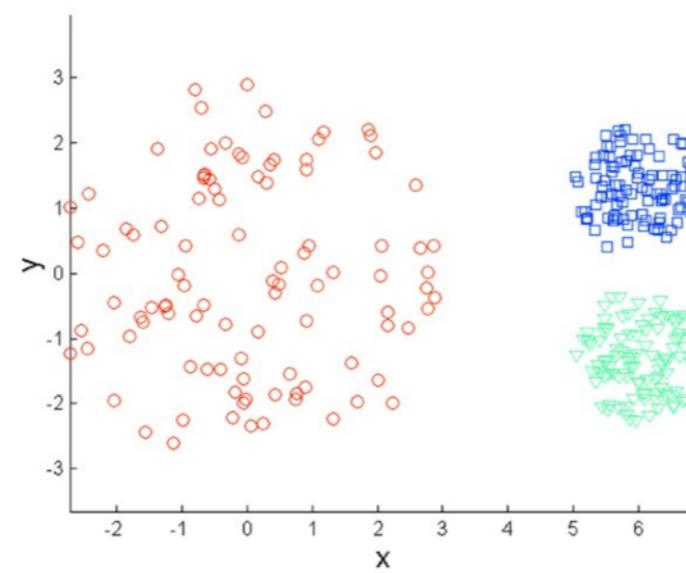
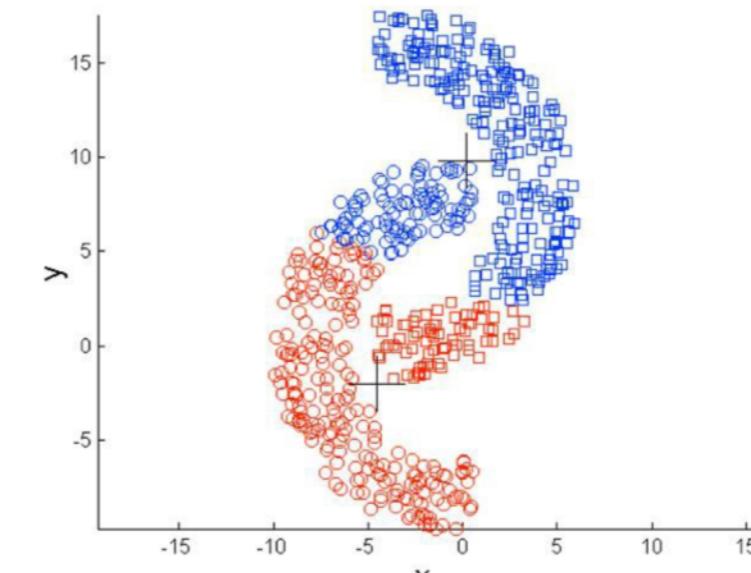
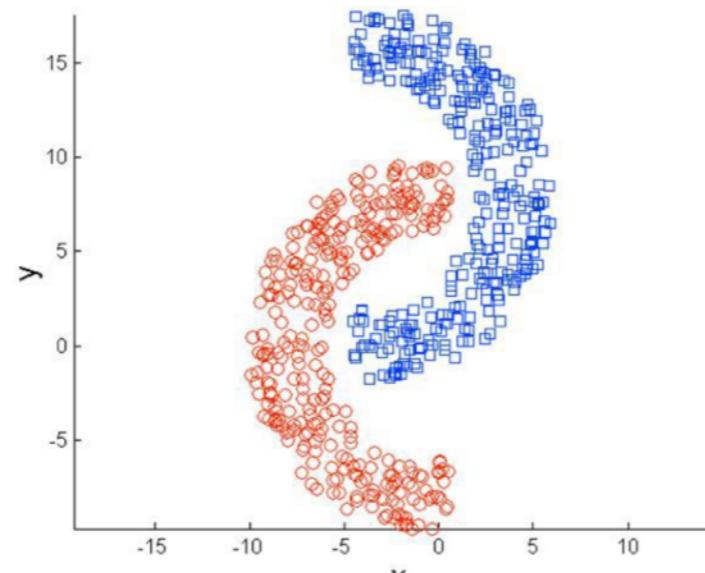
- Step 3 : 군집의 중심점 찾기 (3rd loop)



군집화 알고리즘: K-평균

k-평균 군집의 한계점

- 구형이 아닌 데이터 분포에 대해서 잘 구분할 수 있을까?
- 밀도가 다른 군집을 분할 수 있을까?

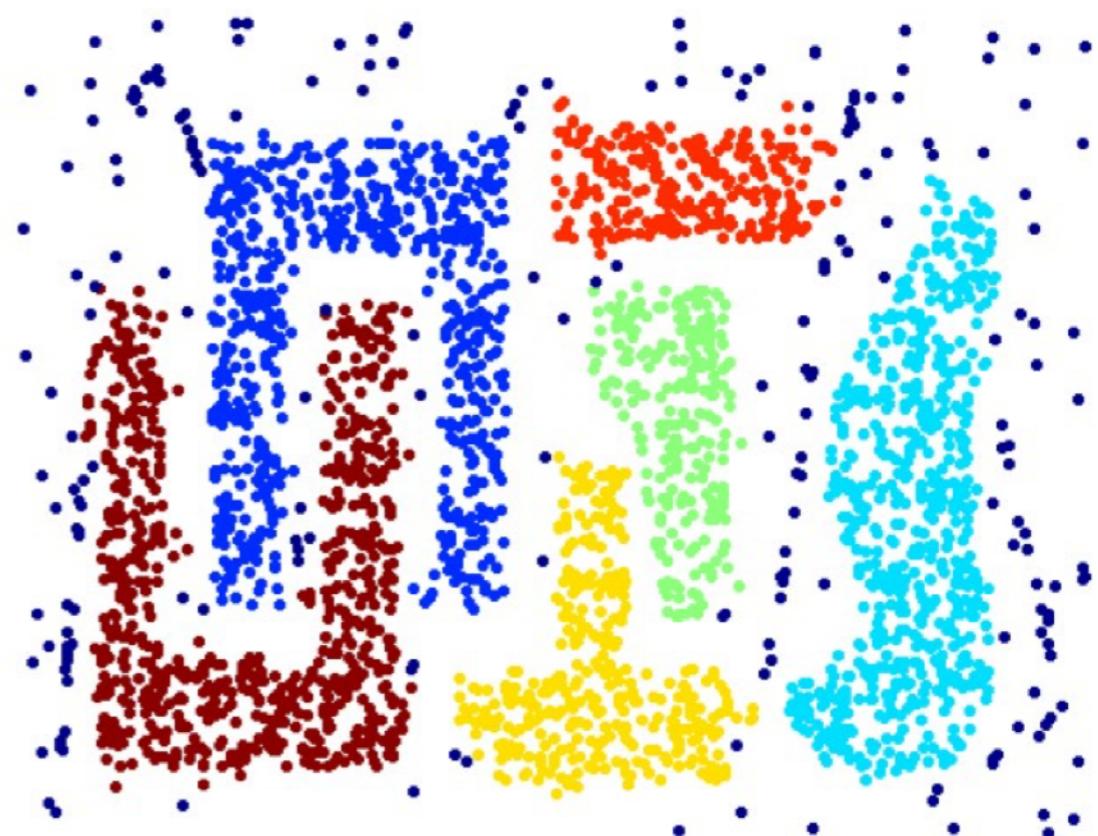


군집화 알고리즘: DBSCAN

Reference

| 밀도기반 군집화 (Density-based Clustering)

- 데이터의 분포와 밀도(density)를 고려하여 군집을 생성하는 방법
- 구형이 아닌 임의의 모양으로 생긴 클러스터도 잘 찾을 수 있음
- 클러스터링 과정에서 노이즈를 제거하는 것이 가능함

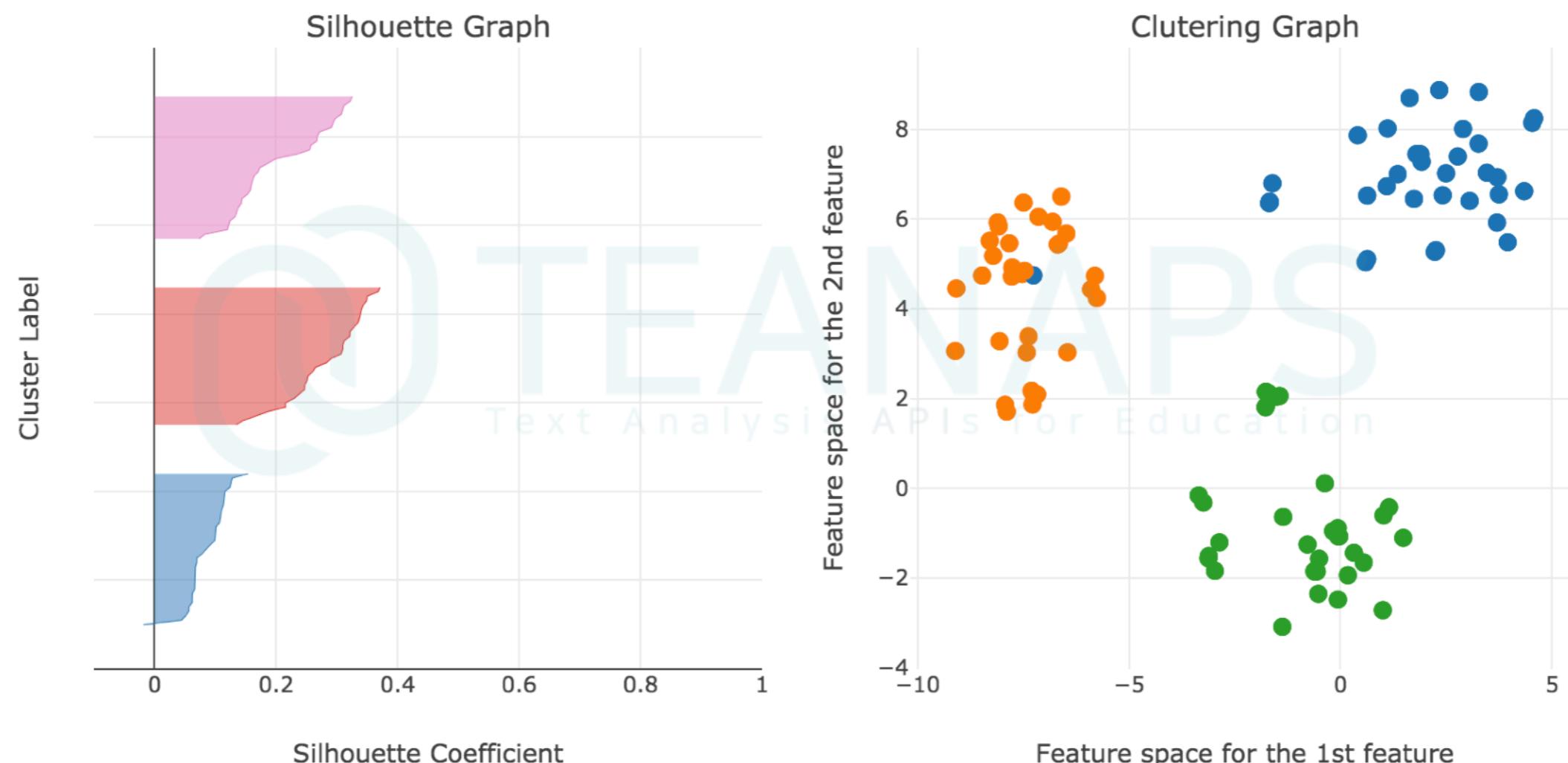


군집의 유효성 검증: 실루엣 스코어

실루엣 스코어 (Silhouette Score)

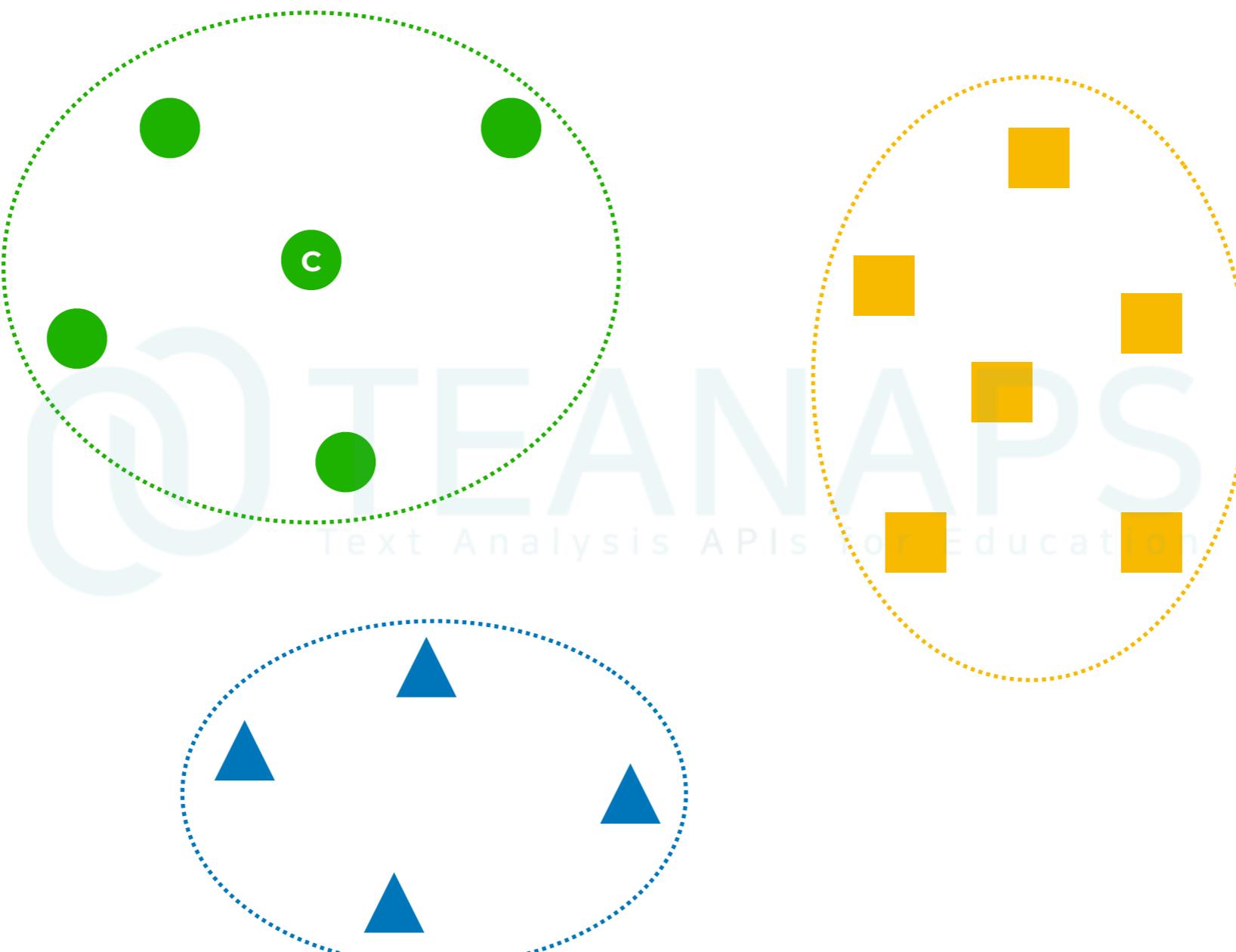
- 클러스터 내의 일관성과 유효성을 검증하는 척도로, 각 객체가 얼마나 잘 분류되었는가를 판단하기 위해 활용됨
- 다른 군집과 비교하여 객체가 자체 클러스터와 얼마나 비슷한지_(cohesion) 또는 분리되어 있는지_(separation)에 대한 척도

Silhouette Analysis for KMeans Clustering - 3 Cluster



군집의 유효성 검증: 실루엣 스코어

실루엣 스코어 계산



군집화 알고리즘: 토픽 모델링

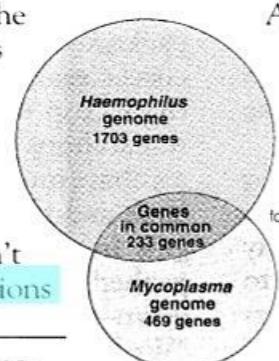
토픽모델링 (Topic Modeling)

- 뉴스, 블로그, 웹페이지, 기사 등 구조화되지 않은 방대한 문서(비정형데이터)에서 주제를 찾아내기 위한 방법
- 맥락과 관련된 단서들을 이용하여 유사한 의미를 가진 단어들을 클러스터링하는 방식으로 주제를 추출하며, 같은 맥락에서 나타날 가능성이 있는 단어들을 그룹화함
- 활용범위 : 문서 요약, 검색 등

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions



Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains

Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



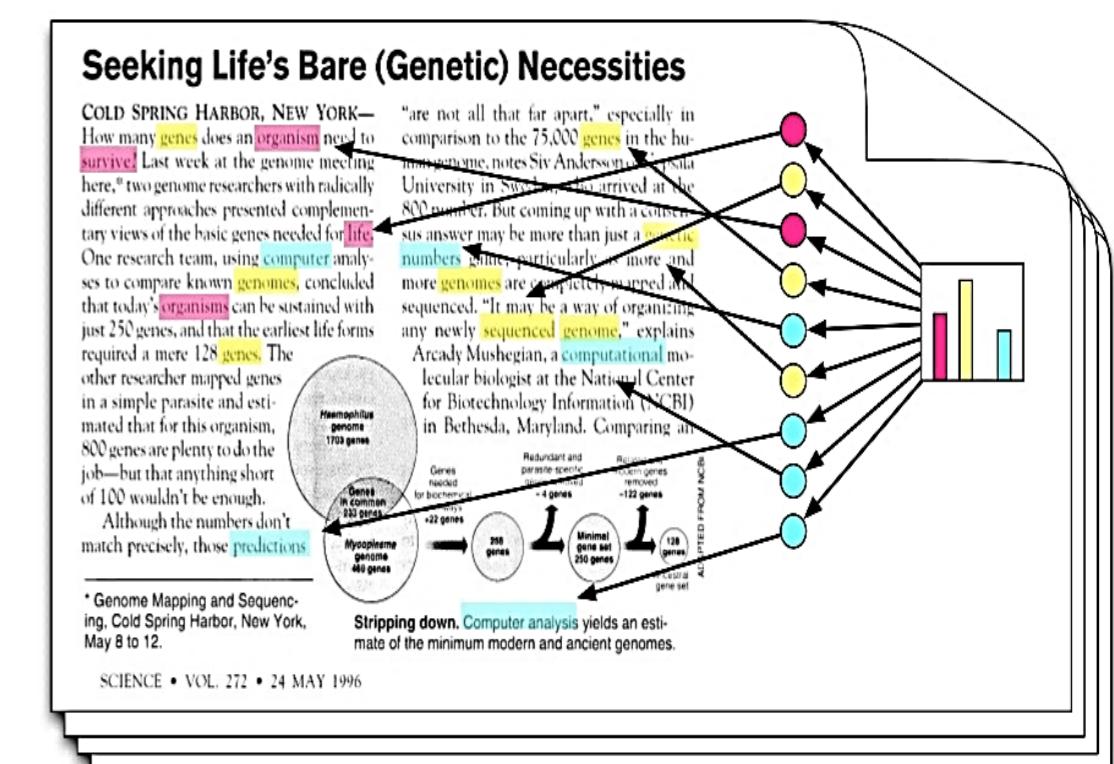
Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

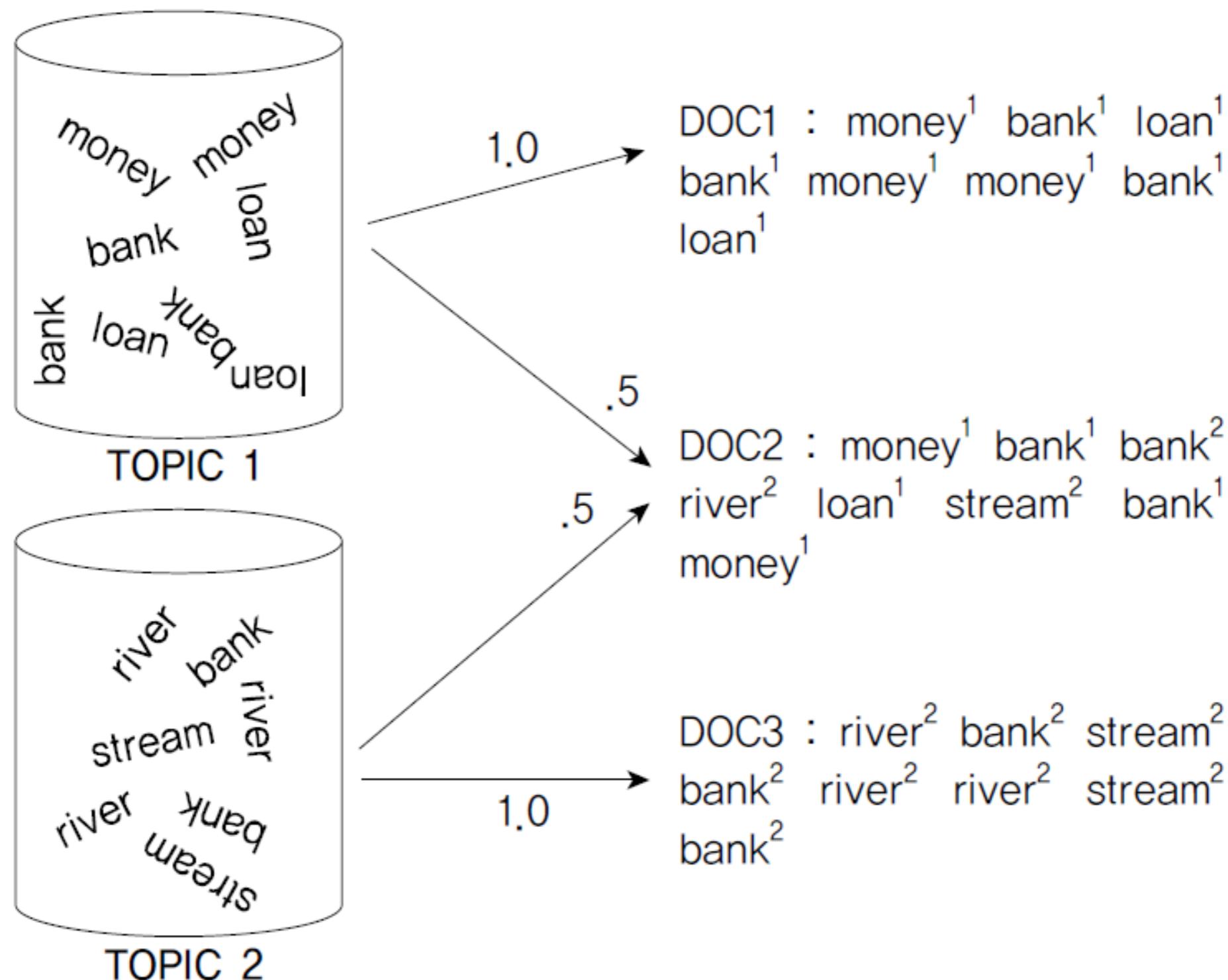
Although the numbers don't match precisely, those predictions

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Topic proportions and assignments



군집화 알고리즘: 토픽 모델링



E.O.D

Contact

-  <http://www.teanaps.com>
-  fingeredman@gmail.com