





UnSupervised Learning (비지도학습) Clustering


	내 용	비 고
1주 (8/29)	강의소개	오프라인 수업
2주 (9/5)	Overview	실습준비 자기 소개 section, 조편성 (3인 1조)
3주(9/12)	Regression 이해하기	
4주(9/19)	Regression 실습 및 사례로 배우기	
5주(9/26)	Regression 이해를 점검하기 주제를 선택해서, Regression으로 분석하기 - 주제 선정 - 주제선택이유 (회귀분석이 가능한 이유, 목적 명시) - 분석 - 결과 및 인사이트	Take Home : 중간고사 오프라인으로 카운팅
6주(10/3)	개천절	휴강
7주(10/10)	Midterm Recital	Peer-Review
8주(10/17)	Supervised Learning 이해하기	
9주(10/24)	Supervised Learning / 실습 및 사례로 배우기 Unsupervised Learning 이해하기	
10주(10/31)	Unsupervised Learning 실습하기	
11주 (11/7)	Recommender System 이해하기 / 실습 및 사례로 배우기	
12주(11/14)	프로젝트 설명 및 Proposal 준비시간	프로젝트 설명 및 이전 예제
13주(11/21)	Final Project Proposal	조별 제출 및 발표
14주(11/28)	프로젝트 준비 시간 및 질의/응답시간	
15주(12/5)	Final Presentation	기말고사: 오프라인 수업


비슷한 문서 찾기


 **Sports** Football Tennis Golf Olympics More


Audio 


Video





Atlético Madrid midfielder's comeback from brain cancer and mother's paralysis 




World's only openly gay active pro footballer is concerned for LGBTQ community ahead of Qatar 2022 





Meet the man who introduced blind football to Uganda 





Nadia Nadim on women's football in Afghanistan one year on from Taliban takeover 




Freestyle football great makes history with ninth world title 





Andriy Shevchenko: War changed everything 





Euro 2022 winner Alessia Russo on making history, inspiring a generation and that viral backheel goal 





'Maybe we are alone': Jude Bellingham questions whether authorities 'care' about racist abuse directed at Black footballers 





Patrice Evra speaks out on racist abuse and how to combat it 




Aurélien Tchouaméni: Meet the French soccer star everyone is talking about 



Patrice Evra: Former France star opens up about sexual abuse 



How Afghanistan women's football team made it to Australia 



비슷한 문서 찾기

- 어떻게 유사한 문서를 찾을까?
- 유사한 문서의 기준이 무엇일까?
- 많은 문서에서 유사한 문서를 어떻게 효과적으로 찾을 수 있을까?
- 어떻게 추천해 줄 것인가?



비슷한 문서 찾기

중요한 점은

사람들은 어떤 기사를 좋아하면, 함께 읽을 기사를 검색하고자 한다!!

수많은 기사 중 어떤 것을 먼저 보고 싶을까?

- 하나하나 직접 읽어가며 ‘관심 있고 없는지’ 일일이 알려줄 수는 없다.
- 관심을 가질 만한 문서를 자동으로 검색할 수단이 필요하다.
- 그렇다면 기사 간의 유사도를 어떻게 측정할까요?
- 어떤 기사가 어떤 면에서 지금 읽는 기사와 유사하다고 판단할 것인가?
- 어떤 방법으로 다음 기사들을 추천할까?



문서 유사도 검색을 위한 방법론

1. Word Count
2. TF-IDF

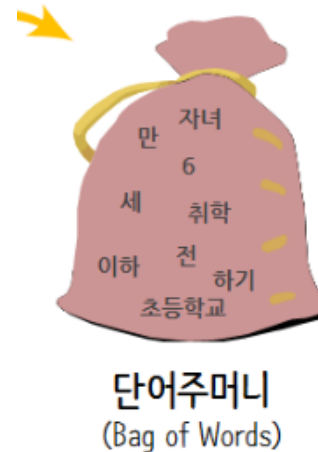
Word Count is from Bag of Words

문서 표현에 가장 인기 있는 방법

- 단어의 순서는 무시하고, 단어의 유무/개수만으로 문서를 표현하는 방법

‘Bag-of-Words’ 모델이라 불리는 이유:

- 주머니에다 문서의 모든 단어를 쏘어 넣고 뒤섞은 후, 해당 단어만 잘 조합하면,
다시 원래 문서를 만들 수 있다.
- 즉, 문서를 만들기 위한 재료!!!
- 순서대로 단어를 잘 정리하면, 정확히 같은 표현 및 문서를 만들 수 있다.



Word Count (1): 문서에서 단어 수만 고려하기

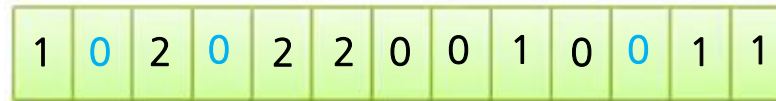
단어 순서보다는 문서에 등장하는 모든 단어의 빈도 수를 고려



“Carlos calls the sport football,
Emily calls the sport soccer.”



[단어 빈도 → Vector == Index]



Carlos Cat the Dog Sport Calls

football Tree soccer Emily

0: 사전에는 있으나, 문서에는 등장하지 않는 단어

Word Count (2): 단어 수를 고려해서 문서 간의 유사도 구하기

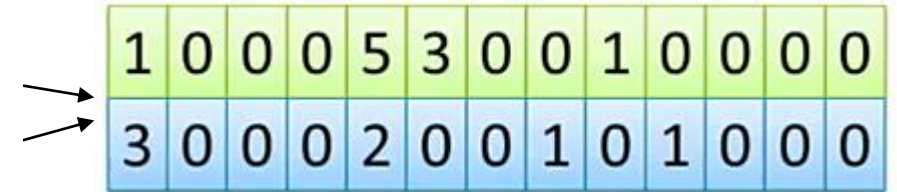
[아르헨티나 메시에 관한 기사의 Count Vector]



[브라질 펠레에 관한 기사의 Count Vector]



문서의 유사도 :
벡터의 스칼라 곱의 합
→ 각각의 항을 곱해서
더합니다.



축구 기사 2개의 유사도:

→ $1 \times 3 + 5 \times 2 = 13$

→ Value가 클수록 더 유사도가 높다!!

Word Count (3): 단어 수를 고려해서 문서 간의 유사도 구하기

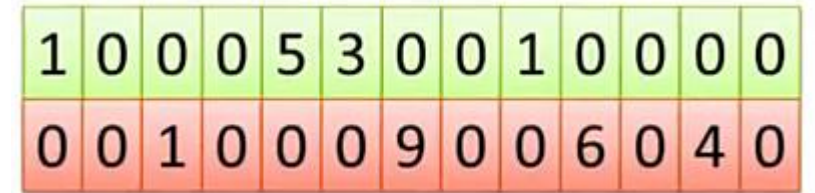
[아르헨티나 메시에 관한 기사의 Count Vector]



[아프리카 분쟁 기사의 Count Vector]



문서의 유사도 :
벡터의 스칼라 곱의 합
→ 각각의 항을 곱해서
더합니다.



축구와 아프리카 기사의 유사도:

$$\rightarrow 1*0+0*1+5*0+3*0+0*9+1*0+0*6+0*4=0$$

→ 2 기사의 유사도 = 0

Word Count (5): 문서의 길이는 유사도를 구할 때 어떤 영향력을 주는가?

→ 길이가 길어짐에 따라 유사도가 더 높게 나타난다.

→ 길이가 두 배라고 더 비슷? → 문서 길이(긴 문서)에 편향되어 있음. → 정규화 필요



1	0	0	0	5	3	0	0	1	0	0	0	0
3	0	0	0	2	0	0	1	0	1	0	0	0

Similarity = 13



2	0	0	0	10	6	0	0	2	0	0	0	0
6	0	0	0	4	0	0	2	0	2	0	0	0

Similarity = 52



Word Count (6): 문서의 길이가 다를 때? 단위를 맞추어 준다. 정규화가 필요하다.

- 벡터를 정규화 → 단어 수 벡터에서 Norm 구하기
- Norm → 모든 항목을 제곱해서 다시 제곱근을 분모로 사용
- 모든 기사(길이에 관계 없이) 동등한 위치에서 유사도 비교 가능
- Norm: 벡터의 길이



1	0	0	0	5	3	0	0	1	0	0	0	0
---	---	---	---	---	---	---	---	---	---	---	---	---

$$\sqrt{(1^2 + 5^2 + 3^2 + 1^2)}$$

1				5	3			1				
/	0	0	0	/	/	0	0	/	0	0	0	0
6				6	6			6				



TF-IDF

Term Frequency– Inverse Document Frequency

중요한 단어와 조사 구별하기

[아르헨티나 메시에 관한 기사의 Count Vector]



- General Words: “the”, “player”, “field”, “goal”
- Rare words: “football”, “Messi”
- Football, Messi등의 단어는 적게 등장하기 때문에 묻혀 버릴 수 있다. 그러나,
- General << 중요한 Rare Words (such as Football, Messi) 가 문서를 검색하는데 더 중요한 Point!!

Document Frequency

- 어떤 단어가 Rare Words 중에서도 중요한 Rare Words일까?
 - 어떻게 정의해야 될까?
 - 단순히 Infrequency words가 중요한 Rare Words 일까??
 - 무조건 자주 안 나타난다고 그 단어가 중요한 단어일까?
 - 답) 아닐 것이다.
 - Rare Words일 수는 있지만, 중요한 단어는 아닐 수 있다. 그렇다면, 중요한 단어는?
 - 특정 문서에만 나타나지만,
 - 해당 문서 내에서는 자주 발생하고,
 - 의미가 있는 단어가 정말 중요한 단어!!

이런 단어들에 가중치를 높여서 적은 수의 문서에만 등장하지만, 해당 단어를 강조한다.

Important Words

- What characterizes an important word?
 - Appears frequently in a document (common locally)

특정 문서에 무척 자주 등장하면

읽고 있는 문서와 관련이 있는 건지

- Appears rarely in corpus (rare globally)

전역적으로 희귀해야 한다.

→ The, is, a 등은 전역적으로 자주 등장하기 때문



Trade off

Trade off → between local frequency and global rarity !!

TF-IDF(Term Frequency-Inverse Document Frequency)

Common locally 과 Rare Globally 를 나타내는 방법 중 하나

- TF : 단어 빈도
→ 지금 읽고 있는 문서에서 단순히 단어 수를 센다. → 단어 수 벡터
- IDF: 역문서 빈도
→ 벡터의 가중치를 조정

TF-IDF(Term Frequency-Inverse Document Frequency)

Term frequency = WORD COUNTS



TF-IDF(Term Frequency-Inverse Document Frequency)

Inversed document frequency

$$= \log(\# \text{ DOCS} / (1 + \# \text{DOCS using WORD}))$$

→ 벡터 가중치를 낮춘다.



$$\log \frac{\# \text{ docs}}{1 + \# \text{ docs using word}}$$



TF-IDF(Term Frequency-Inverse Document Frequency)

Inversed document frequency

$$= \log(\# \text{ DOCS} / (1 + \# \text{DOCS using WORD}))$$

→ 벡터 가중치를 낮춘다.



$$\log \frac{\# \text{ docs}}{1 + \# \text{ docs using word}}$$

[희소 단어] Rare word → $\log(\text{large \#} / 1 + \text{small \#}) \rightarrow \text{large \#}$

#docs는 클 것이다. (#docs: 총 문서수)

1+ #docs using word (단어가 있는 문서 수는 작을 것이다)

따라서, $\log(\# \text{ docs} / (1 + \# \text{ docs using word})) \rightarrow \text{큰 수} / \text{작은 수} = \text{큰 수}$

[참고: 1을 더해 주는 이유]

문서의 모든 단어가 말뭉치에 있다고 가정할 수 없기 때문!!

어떤 문서에도 없다면, 분모가 0이 되므로, 수식이 성립이 되지 않는다.

→ 1을 더해준다 → 대세에 지장이 없다.

TF-IDF (Term Frequency-Inverse Document Frequency)

[Term Frequency : 대상문서에서 해당 단어가 나타나는 숫자]



[Inverse Document Frequency]



$$\log(64/1+63) = 0$$

$$\log(64/1+3) = \log 16$$

tf * idf ← 두 벡터의 곱

$$1000 * 0 + 5 * 4 = 20$$



[가정] “the” 모든 문서에 나타난다.

- 문서 수: 64개 로 가정
- 계산을 간단히 하기 위해, 하나를 뺀 모든 문서에 “the”가 포함
- 가중치 = 0
- 문서 검색의 변별력이 적다.

[가정] “messi” 는 희소 단어다.

- 문서 수: 64개 로 가정
- 총 64개의 문서 중에 3개의 문서에 등장
- $\log 16 = 4$
- 문서 검색의 변별력이 높다.

- Common word 인 “the”의 가중치는 내려가고,
 - 희귀하지만 중요단어인 “messi”의 가중치는 올라간다.
- [결론] 문서 검색에 효율적이다.

Word in many docs → $\log(\text{large \#} / 1 + \text{large \#}) \approx \log 1 = 0$, Rare word → $\log(\text{large \#} / 1 + \text{small \#}) \rightarrow \text{large \#}$



유사한 검색들의 묶음 : Clustering (kNN)

Nearest Neighbors Search

지금 읽고 있는 기사와 가장 근접한/유사도가 높은 기사를 검색하는 방법
→ 거리 기준이 있어야 한다. 거리 기준을 현재 읽고 있는 책과 가장 유사한 k개!!

Query Article :



Corpus :



Specify : Distance metric

Output : Set of most similar articles

K- Nearest Neighbors Search


지금 읽고 있는 기사와 가장 관련 있는 기사를 출력하는 대신 가장 관련 있는 k ($k=8$)개의 기사 모음을 보여준다.

Input : Query article



Output: *List of k* similar article





유사한 검색들의 묶음 : Clustering (K-means)

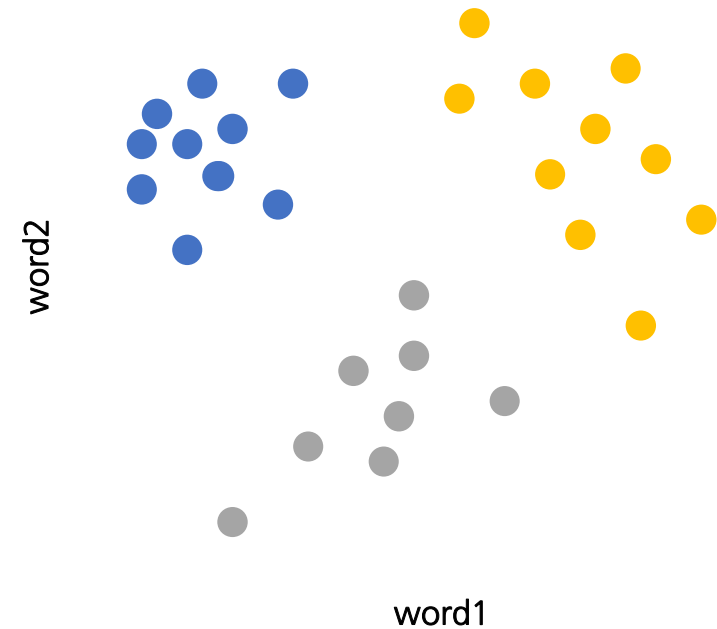
Clustering : Unsupervised Learning

Labeling이 되어 있지 않은 데이터를 다룰 때

Input: Documents안에 Words들의 분포 → 단어 수 벡터

Output: Cluster labels

→ 문서에 대한 라벨링



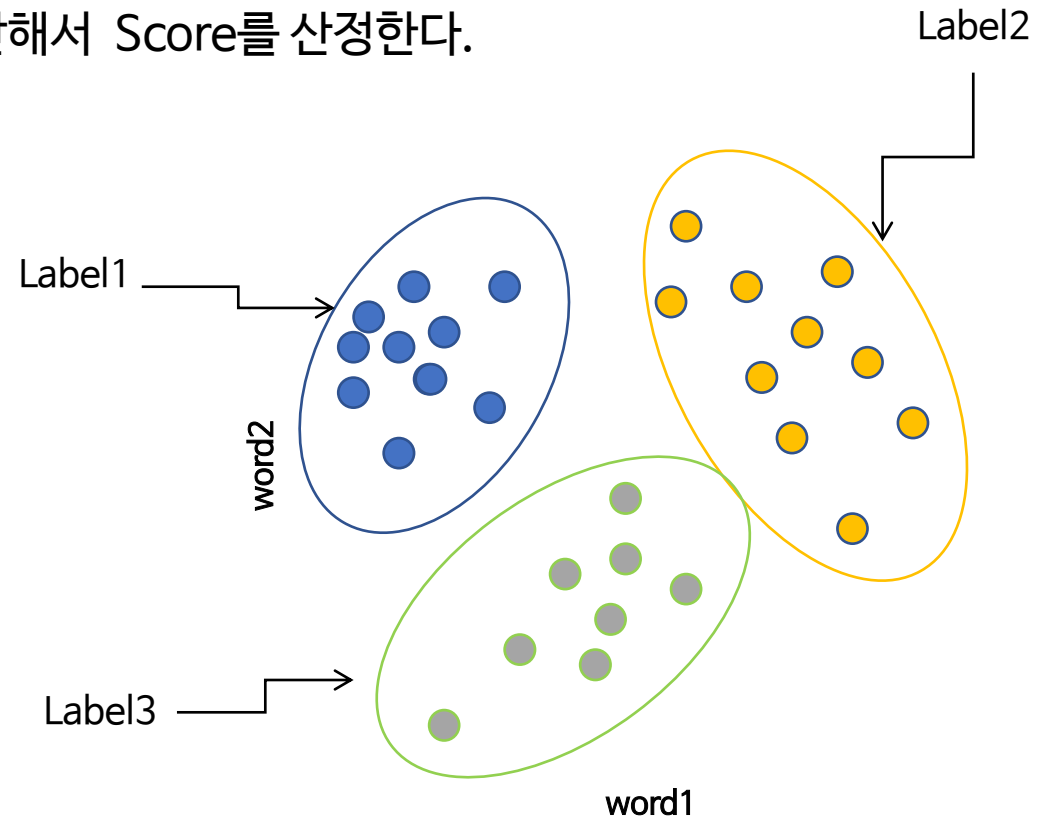
Clustering : k-means

문서가 Word1 과 Word2로만 이루어졌다고 가정할 때, 물론 실제로는 다양한 Words 로 이루어져 있음.

- Word들의 분포(Word Vector)를 보고 Topic별로 Clustering 한다.
- 보통 중앙을 기준으로 분포도를 보고 거리(유사도)에 기반해서 Score를 산정한다.
- 비슷할수록 센터에 가깝게 분포되어 있다.

Input: Documents안에 Words들의 분포 → 단어 수 벡터

Output: Label1, Label 2, Label 3



Clustering : k-means

Input (Observation) : Documents안에 Words들의 분포 → 단어 수 벡터

Output (Topic Label) : Label1, Label 2, Label 3

클러스터 중심 : ★

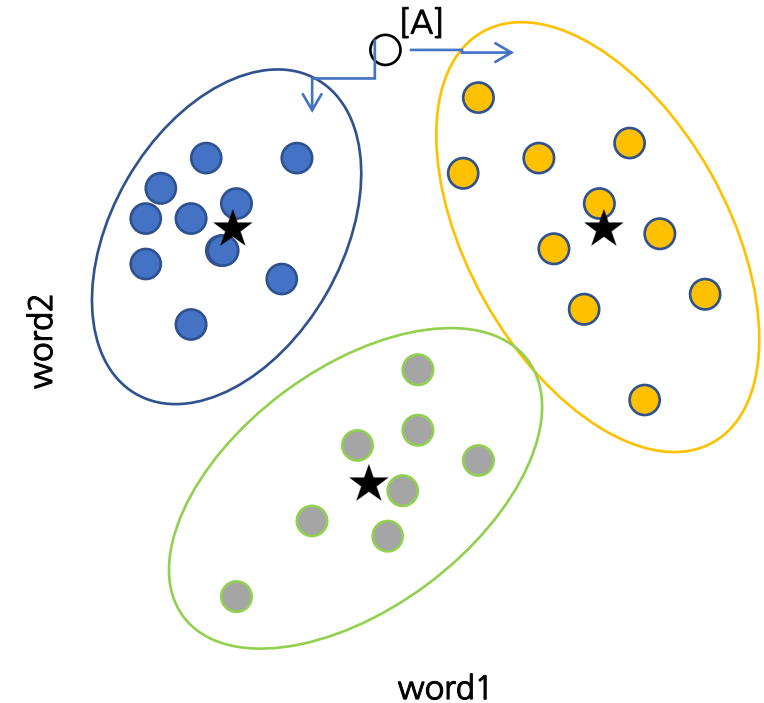
[A] 가 어디에 Blue Cluster 인지, Yellow Cluster인지 확실하지 않음.

[A] 가 어떤 Cluster에 속하는지 알기 위해

1. 유사도 검사가 필수

모든 샘플의 점수를 클러스터 중심과의 거리를 점수화 하여
유사도를 측정

2. [A] 와 중심과의 거리만



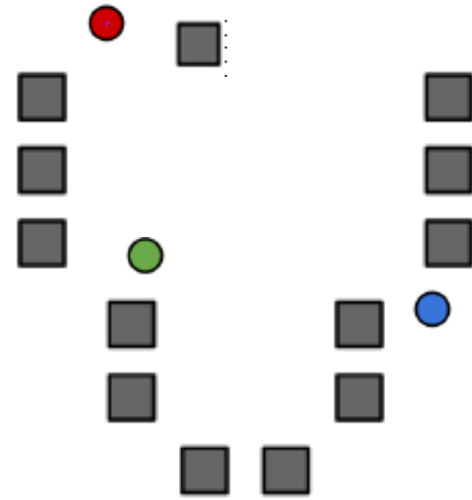
예제로 살펴보기 : k-means

- 클러스터 중심과의 거리만을 기준으로 삼는 클러스터링 알고리즘
→ Similarity metric = **Distance to Cluster Center** (smaller better)
- k 클러스터가 있다는 가정 →
 - k 클러스터가 있고
 - 각 샘플을 클러스터로 할당할 때
→ 클러스터의 평균 = 클러스터 중심

예제로 살펴보기 : k-means (1) 초기화

1. Initialize cluster centers

- 클러스터 중심을 무작위로 놓았다고 가정
- K의 개수 설정하기 → Clusters의 개수를 설정하기
- 여기 예제에서는 k를 3으로 설정 → 3-means 알고리즘



Data to Cluster

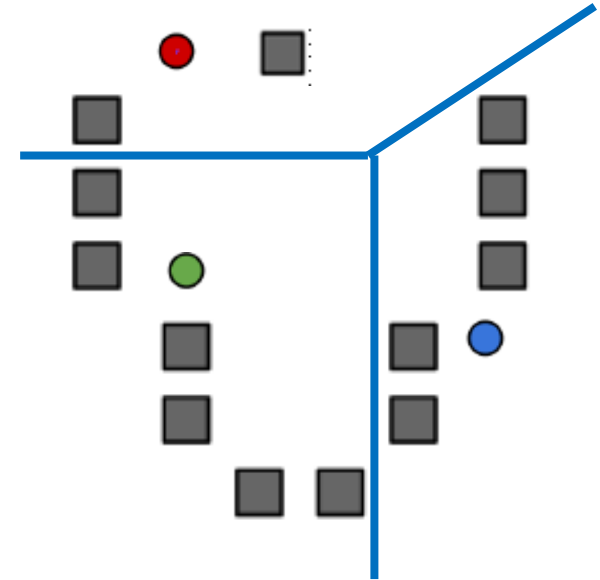
예제로 살펴보기 : k-means (2) 가까운 곳에 중심점 찍기

Similarity metric = Distance to Cluster center (smaller better)

1. Initialize cluster centers

2. Assign observations to closet cluster center →

첫 단계는 모든 샘플을 가장 가까운 클러스터 중심에 할당



예제로 살펴보기 : k-means (3) Similarity 구하기

Similarity metric = **Distance to Cluster center** (smaller better)

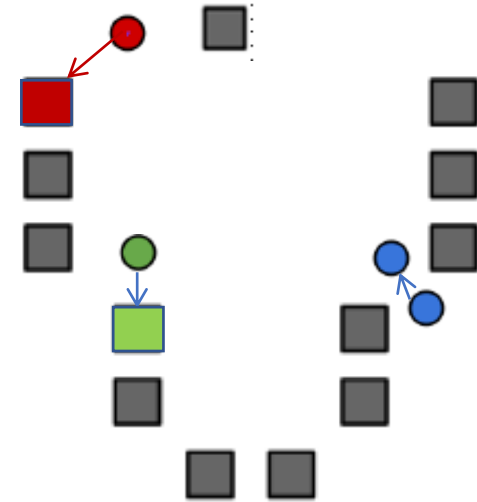
1. Initialize cluster centers
2. Assign observations to closet cluster center →

첫 단계는 모든 샘플을 가장 가까운 클러스터 중심에 할당

3. Revise cluster centers as mean of assigned observations

해당 중심점에서 각 점까지의 거리 값을 구한다.

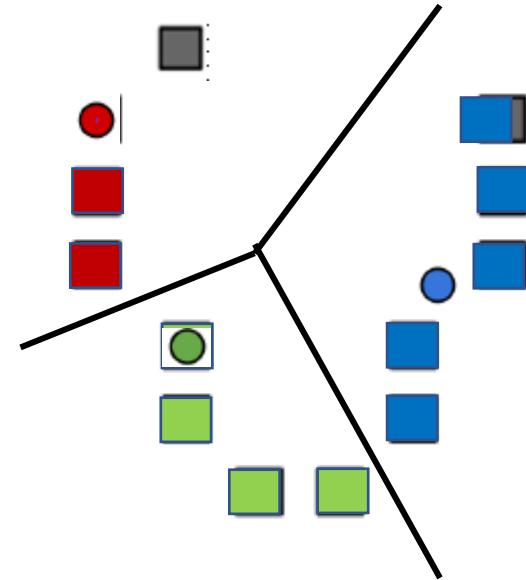
각 클러스터마다 중심점과 점들까지의 거리 값을 구한 후,
평균을 구한다.



예제로 살펴보기 : k-means (4) 중심점 업데이트하기

Similarity metric = Distance to Cluster center (smaller better)

1. Initialize cluster centers
2. Assign observations to closet cluster center →
첫 단계는 모든 샘플을 가장 가까운 클러스터 중심에 할당
3. Revise cluster centers as mean of assigned observations
해당 중심점에서 각 점까지의 거리 값을 구한다.
각 클러스터마다 중심점과 점들까지의 거리 값을 구한 후,
평균을 구한다.
4. Repeat 2.+3. until convergence



예제로 살펴보기 : 주제별 그룹핑



SPORTS



WORLD NEWS

예제로 살펴보기 : 이미지 그룹핑

Ocean



Pink flower



Dog



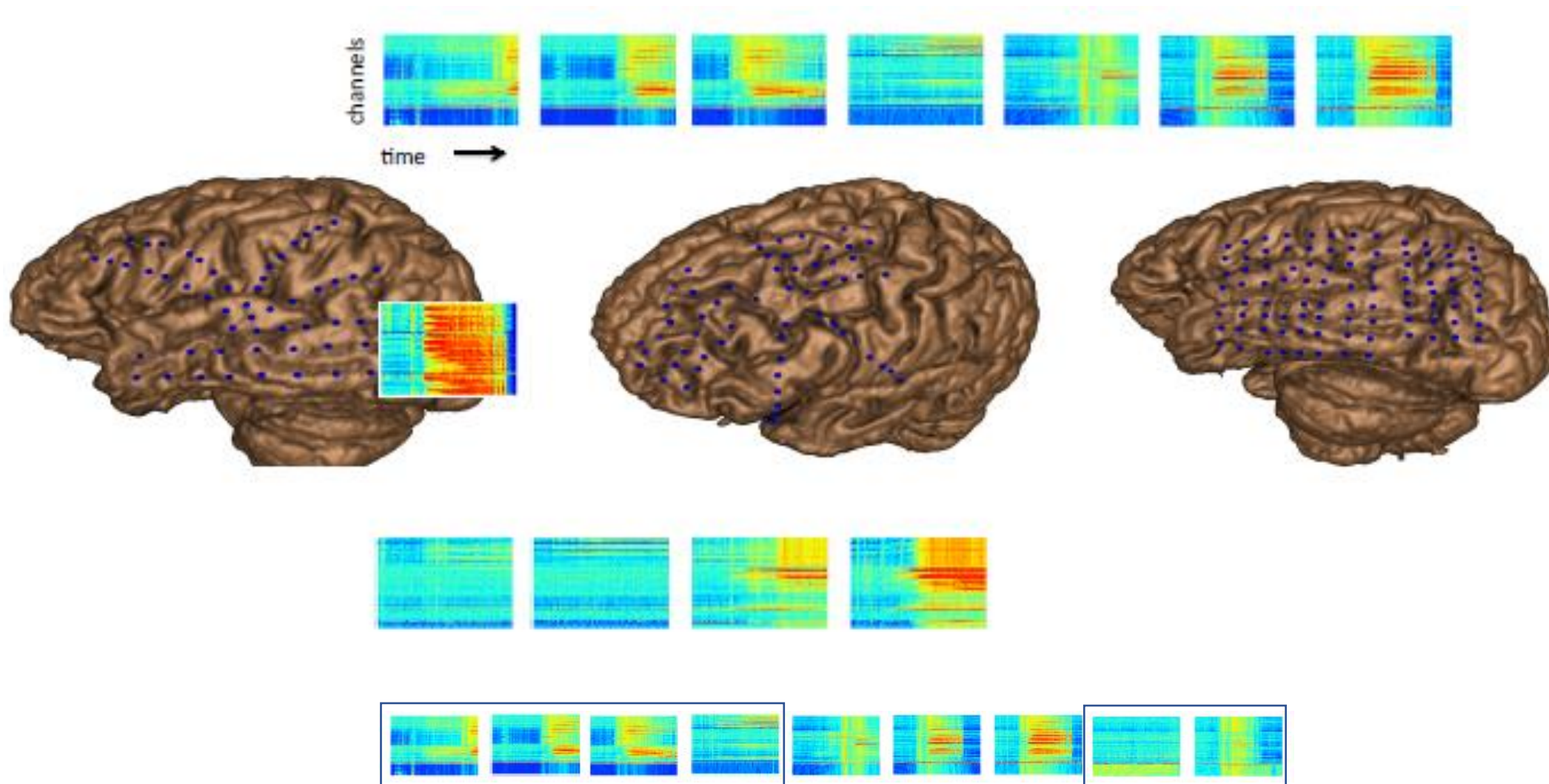
Sunset



Clouds

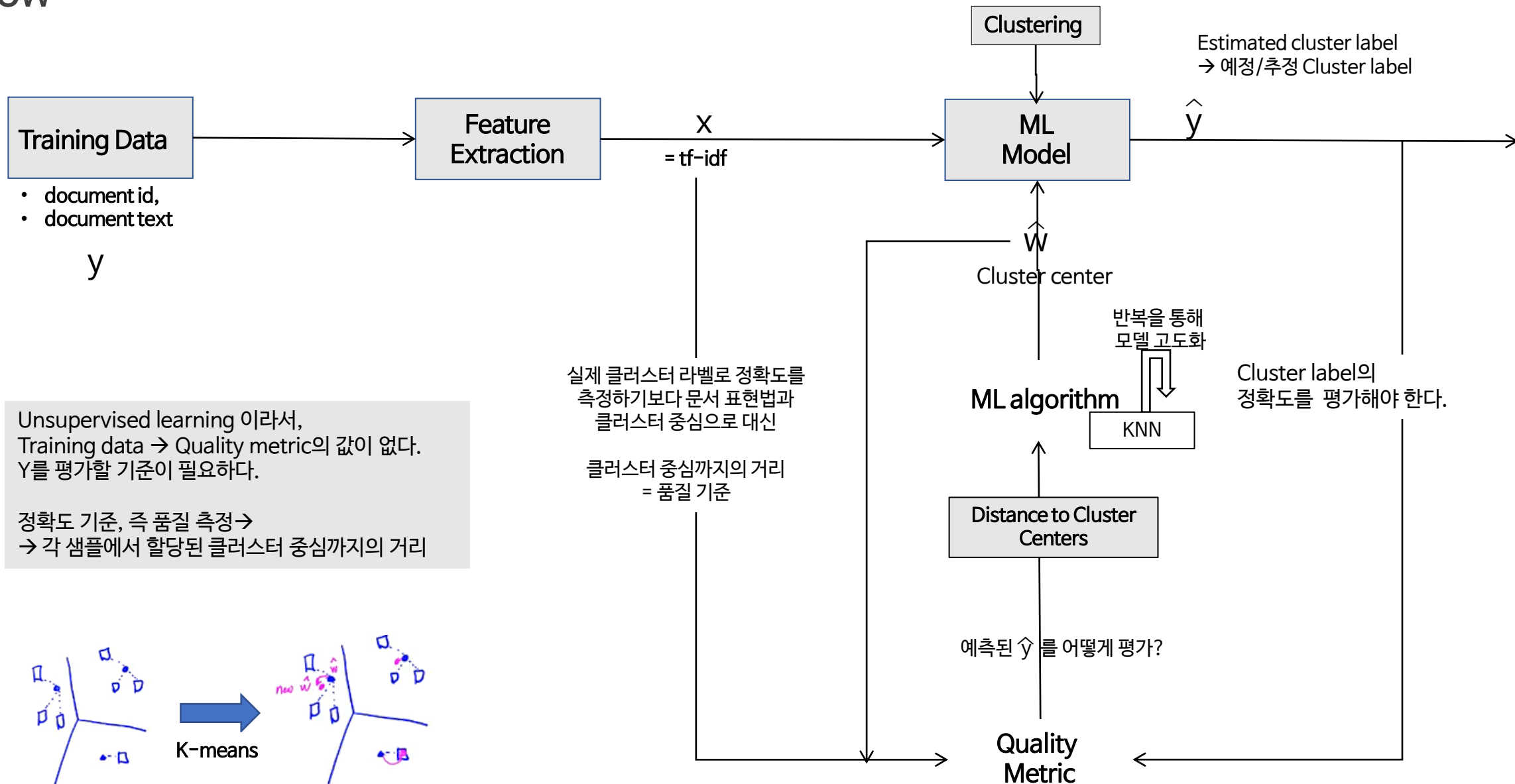


예제로 살펴보기 : 환자 CT 판독





요약하기 : 한 눈에 보기



	내 용	비 고
1주 (8/29)	강의소개	오프라인 수업
2주 (9/5)	Overview	실습준비 자기 소개 section, 조편성 (3인 1조)
3주(9/12)	Regression 이해하기	
4주(9/19)	Regression 실습 및 사례로 배우기	
5주(9/26)	Regression 이해를 점검하기 주제를 선택해서, Regression으로 분석하기 - 주제 선정 - 주제선택이유 (회귀분석이 가능한 이유, 목적 명시) - 분석 - 결과 및 인사이트	Take Home : 중간고사 오프라인으로 카운팅
6주(10/3)	개천절	휴강
7주(10/10)	Midterm Recital	Peer-Review
8주(10/17)	Supervised Learning 이해하기	
9주(10/24)	Supervised Learning / 실습 및 사례로 배우기 Unsupervised Learning 이해하기	
10주(10/31)	Unsupervised Learning 실습하기	
11주 (11/7)	Recommender System 이해하기 / 실습 및 사례로 배우기	
12주(11/14)	프로젝트 설명 및 Proposal 준비시간	프로젝트 설명 및 이전 예제
13주(11/21)	Final Project Proposal	조별 제출 및 발표
14주(11/28)	프로젝트 준비 시간 및 질의/응답시간	
15주(12/5)	Final Presentation	기말고사: 오프라인 수업



실습 with Python