ADVANCED TEXT MINING

by FINGEREDMAN (fingeredman@gmail.com)

Review

Business Understanding

START

Check 텍스트 마이닝을 통해 해결하고자 하는 문제가 명확한가?

Yes / No

Check 해결하고자 하는 비즈니스에 대한 충분한 이해를 가지고 있는가?

Yes / No

비즈니스 문제정의

Data Preparation

Check 문제해결을 위해 분석할 텍스트 데이터를 보유하고 있는가?

Yes / No

Check 해당 데이터에 접근할 권한을 가지고 있는가?

Yes / No

Check 문제해결을 위해 필요한 데이터가 온라인상에 존재하는가?

Yes / No

Check 해당 온라인 데이터 소스에서 원하는 정보를 가져오기 위한 API를 제공하는가?

Yes / No

- API 호출을 통한 데이터 수집
- 웹 스크래핑을 통한 데이터 수집
- 오프라인 데이터 수집

Data Pre-processing

Check 필요한 텍스트가 일정한 형태로 정규화되어 있는가?

Yes / No

데이터 정규화

Check 텍스트 데이터를 특성에 따라 분리하여 분석할 필요가 있는가?

Yes / No

데이터 분리

Check 분석할 텍스트가 영문이거나 영어단어가 다수 포함되어 있는가?

Yes / No

형태소분석

개체명인식

원형복원

불용어제거

단어빈도분석

Check 불용어 및 빈출어가 충분히 제거되었는가?

Yes / No

Check 필요한 단어들이 올바르게 추출되었는가?

Yes / No

- 동시출현분석
- 문서분류
- 문서요약
- 키워드 추출
- 클러스터링

- 임베딩
- 알고리즘 선정
- 학습
- 평가
- 배포

시각화

분석결과 비교/해석

분석결과 리포팅

COMPLETE

Data Analysis, Visualization, and Reporting

* references
** references
*** references

WEEK 07

Text Data Embedding

문서를 단어 가중치로 표현하는 방법

Review

문서 내 단어의 빈도 계산하기

OhmyNews

"벗꽃 상춘객 올까봐 불도 꺼... 올해는 제발 참아달라"

기사입력 2020.03.28. 오후 8:30 기사원문 스크랩 본문듣기 • 설정

27 14 요약봇 Beta 가

[사진] 경주 벚꽃터널, 주정차 금지... 상춘객 몰릴 것에 대비해 야간경관조명도 꺼

[오마이뉴스 한정환 기자]

▲ 28일 주말 오전, 조금은 한산한 경주 흥무로 벚꽃길 모습
© 한정환

천년고도 경주에 벚꽃이 활짝 피었다. 해마다 이맘때쯤이면 경주는 벚꽃 관광객 맞이로 분주했다. 그러나 코로나19 확산으로, 올해는 사정이 달라졌다.

28일 오전 주말을 맞아 벚꽃길로 유명한 흥무로 벚꽃터널을 찾아보았다. 예년에 비해 벚꽃 상춘객들은 많이 줄었다. 코로나19 여파로 많은 관광객이 경주 방문을 자제하고 있는 듯 보인다. 그러나 벚꽃 상춘객이 몰릴 것에 대비하여 벚꽃터널에는 도로 양방향 100m 간격으로 경찰관들이 배치되어 불법 주정차 단속을 하고 있다. 사진을 찍기 위해 차에서 잠시 내리는 것도 안 된다.

관광도시 경주의 특성상 다른 도시처럼 벚꽃 명소를 통제할 수가 없다. 도시 전체가 관광지이고, 대부분 가로수에 벚꽃이 심어져 있어 통제를 하게 되면, 도시 전체가 봉쇄가 되어야 하기 때문에 현수막을 걸어 두고 홍보만 하고 있다. 사진으로나마 경주 벚꽃 소식을 전하기로 했던 '경주시 벚꽃 알리미'도 잠정 중단된 상태이다.

경주시 관계자는 "지난 22일 일제히 불을 밝힌 야간경관조명도 27일 저녁부터 벚꽃 상춘객이 몰릴 것에 대비하여 일제히 불을 껐다. 벚꽃은 해마다 피니, 올해는 집단 감염이 우려되는 벚꽃 나들이를 내년으로 미루고, 코로나19 확산 방지를 위해 모두가 힘을 합쳐달라"라고 말했다.

경주시는 지난 2월 22일 첫 코로나19 확진자가 발생한 이래 현재까지 총 40명이 양성 판정을 받았으며, 이 중에서 사망 1명, 완치 10명을 제외한 29명이 현재 자가격리 및 생활치료센터에 입소하여 치료를 받고 있다.

* 한정환(오마이뉴스), "벚꽃 상춘객 올까봐 불도 꺼... 올해는 제발 참아달라", 2020.3.28., https://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=103&oid=047&aid=0002263496/.
** references
*** references

Review

문서를 단어 가중치로 표현하는 방법

형태소 분석을 통한 단어주머니 생성

구분	문장
문장 01	천년고도 경주에 벚꽃이 활짝 피었다.
문장 02	해마다 이맘때쯤이면 경주는 벚꽃 관광객 맞이로 분주했다.
문장 03	그러나 코로나19 확산으로, 올해는 사정이 달라졌다.
문장 04	8일 오전 주말을 맞아 벚꽃길로 유명한 흥무로 벚꽃터널을 찾아보았다.
문장 05	예년에 비해 벚꽃 상춘객들은 많이 줄었다.
문장 06	코로나19 여파로 많은 관광객이 경주 방문을 자제하고 있는 듯 보인다.
문장 07	그러나 벚꽃 상춘객이 몰릴 것에 대비하여 벚꽃터널에는 도로 양방향 100m 간격으로 경찰관들이 배치되어 불법 주정차 단속을 하고 있다.
문장 08	사진을 찍기 위해 차에서 잠시 내리는 것도 안 된다.
문장 09	관광도시 경주의 특성상 다른 도시처럼 벚꽃 명소를 통제할 수가 없다.
문장 10	도시 전체가 관광지이고, 대부분 가로수에 벚꽃이 심어져 있어 통제를 하게 되면, 도시 전체가 봉쇄가 되어야 하기 때문에 현수막을 걸어 두고 홍보만 하고 있다.
문장 11	사진으로나마 경주 벚꽃 소식을 전하기로 했던 '경주시 벚꽃 알리미'도 잠정 중단된 상태이다.
문장 12	경주시 관계자는 "지난 22일 일제히 불을 밝힌 야간경관조명도 27일 저녁부터 벚꽃 상춘객이 몰릴 것에 대비하여 일제히 불을 껐다.
문장 13	벚꽃은 해마다 피니, 올해는 집단 감염이 우려되는 벚꽃 나들이를 내년으로 미루고, 코로나19 확산 방지를 위해 모두가 힘을 합쳐달라"라고 말했다.
문장 14	경주시는 지난 2월 22일 첫 코로나19 확진자가 발생한 이래 현재까지 총 40명이 양성 판정을 받았으며, 이 중에서 사망 1명, 완치 10명을 제외한 29명이 현재 자가격리 및 생활치료센터에 입소하여 치료를 받고 있다.

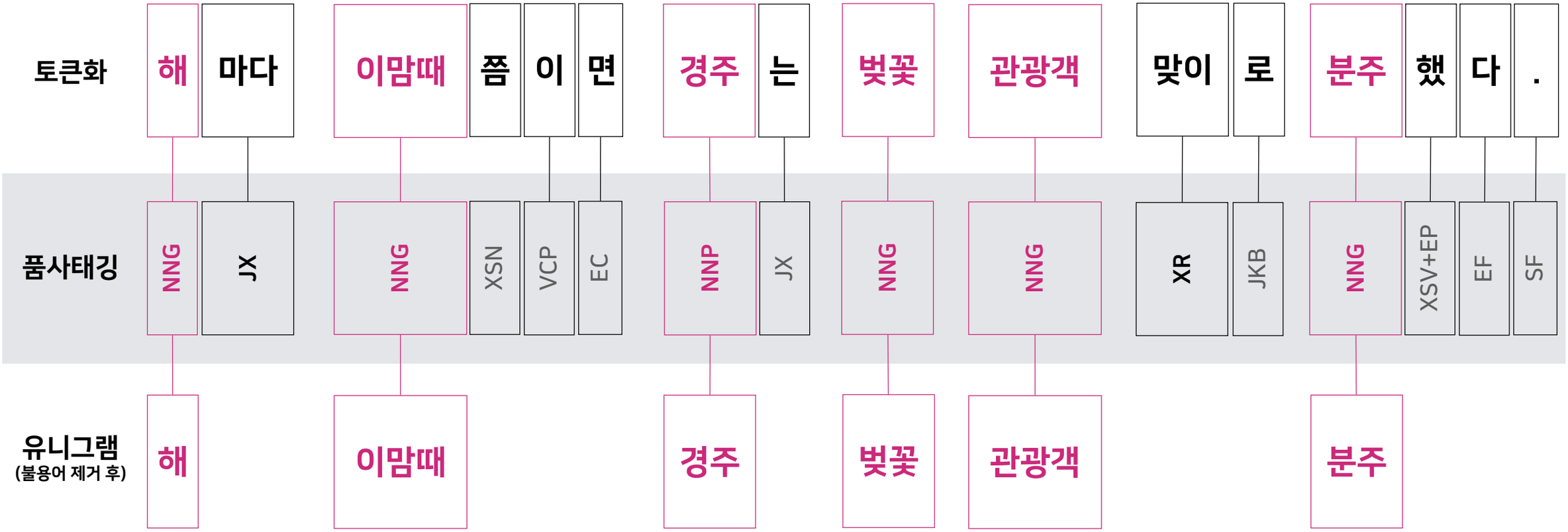
* references
** references
*** references

문서를 단어 가중치로 표현하는 방법

Review

형태소 분석을 통한 단어주머니 생성

구분	내용
원문	해 마 다 이 맘 때 쯤 이 면 경 주 는 벚 꽃 관 광 객 맞 이 로 분 주 했 다 .



* references
** references
*** references

문서를 단어 가중치로 표현하는 방법

Review

형태소 분석을 통한 단어주머니 생성

구분	유니그램 (품사=NNG|NNP, 불용어 제거)
문장 01	고도, 경주, 벚꽃
문장 02	이맘때, 경주, 벚꽃, 관광객, 분주
문장 03	코로나, 확산, 올해, 사정
문장 04	오전, 주말, 벚꽃, 길, 유명, 흥무, 벚꽃, 터널
문장 05	예년, 벚꽃, 상춘객
문장 06	코로나, 여파, 관광객, 경주, 방문, 자제
문장 07	벚꽃, 상춘객, 대비, 벚꽃, 터널, 도로, 양방향, 간격, 경찰관, 배치, 불법, 주정차, 단속
문장 08	사진
문장 09	관광, 도시, 경주, 특성, 도시, 벚꽃, 명소, 통제
문장 10	도시, 전체, 관광지, 대부분, 가로수, 벚꽃, 통제, 전체, 봉쇄, 현수막, 홍보
문장 11	사진, 경주, 벚꽃, 소식, 경주시, 벚꽃, 알리, 잠정, 중단, 상태
문장 12	경주시, 관계자, 야간, 경관, 조명, 저녁, 벚꽃, 상춘객, 대비
문장 13	벚꽃, 올해, 집단, 감염, 우려, 벚꽃, 나들이, 내년, 코로나, 확산, 방시, 위해, 모두
문장 14	경주시, 지난, 코로나, 확진, 발생, 현재, 양성, 판정, 사망, 완치, 제외, 격리, 생활, 치료, 센터, 입소, 치료

단어주머니 (74 단어)

벚꽃	도로	제외
경주	대부분	저녁
도시	단속	잠정
코로나	내년	자제
상춘객	관광지	자가
경주시	관광	입소
전체	관계자	이맘때
사진	고도	유명
치료	경찰관	우려
터널	경관	완치
통제	격리	오전
올해	감염	예년
대비	간격	여파
확산	나들이	양성
관광객	센터	양방향
발생	상태	야간
모두	생활	알리
방문	홍보	소식
봉쇄	현재	확진
방지	현수막	가로수
배치	판정	
분주	특성	
불법	집단	
사망	중단	
사정	주정차	
명소	주말	
흥무	조명	

* references
** references
*** references

문서를 단어 가중치로 표현하는 방법

Review

형태소 분석을 통한 단어주머니 생성

구분	벚꽃	경주	도시	코로나	상춘객	경주시	전체	사진	치료	터널	통제	올해	대비	확산	관광객	발생	모두	방문	봉쇄	방지	배치	분주	불법	사망	사정	명소	흥무
문장 01	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
문장 02	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
문장 03	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0
문장 04	2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
문장 05	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
문장 06	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
문장 07	2	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0
문장 08	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
문장 09	1	1	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
문장 10	1	0	2	0	0	0	2	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
문장 11	2	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
문장 12	1	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
문장 13	2	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0
문장 14	0	0	0	1	0	1	0	0	2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0

* references
** references
*** references

HOW?

책 한권은 몇개의 단어로 표현할 수 있을까?

* references
** references
*** references

단어를 표현하는 방법

원-핫 인코딩 (One-hot Encoding/Representation)

- 문서에 포함된 N개의 단어를 N차원 벡터로 표현하는 방법
  1) 한국어의 토큰(단어) 수는 50만개 이상이며, 모든 단어의 표현을 위해 50만개 이상의 차원이 필요함
  2) 영어의 토큰(단어) 수는 1,300만개 이상이며, 모든 단어의 표현을 위해 1,300만개 이상의 차원이 필요함

구분	문장
문장 A	안녕하세요. 저는 워너원 멤버 강다니엘 입니다.
문장 B	안녕하세요. 얘는 트와이스 멤버 정연 입니다.
문장 C	안녕하세요. 저분은 워너원 멤버 황민현 입니다.

구분	유니그램 (품사=NNG|NNP, 불용어 제거)
문장 A	안녕, 워너원, 멤버, 강다니엘
문장 B	안녕, 트와이스, 멤버, 정연
문장 C	안녕, 워너원, 멤버, 황민현

구분	안녕	워너원	트와이스	멤버	정연	강다니엘	황민현
안녕	1	0	0	0	0	0	0
워너원	0	1	0	0	0	0	0
트와이스	0	0	1	0	0	0	0
멤버	0	0	0	1	0	0	0
정연	0	0	0	0	1	0	0
강다니엘	0	0	0	0	0	1	0
황민현	0	0	0	0	0	0	1

[문제점]

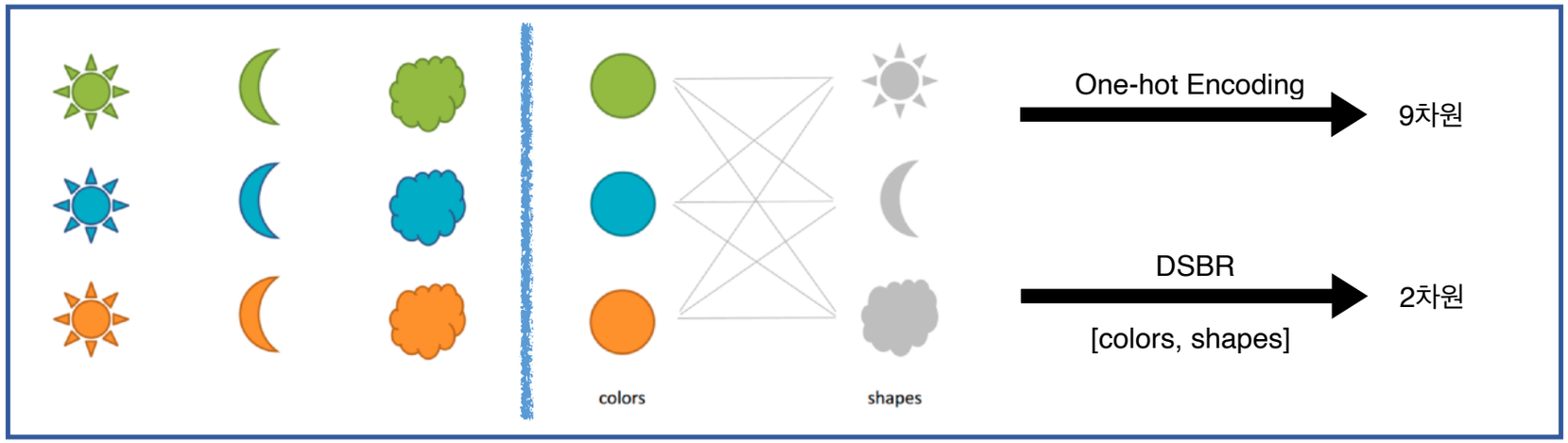
- **차원의 저주(curse of dimensionality)** : 하나의 단어를 표현하기 위해 문서에 존재하는 단어의 수만큼 차원을 만들어야함
  (구어체: 20K, 일반문서(PTB): 50K, 형태소분석: 500K, Google: 13M/1T)
- 각 단어는 서로 관계를 가질 수 없기 때문에($VS_T = 0$), 유의어, 반의어와 같은 관계를 표현할 수 없고 서로 독립적으로만 존재함

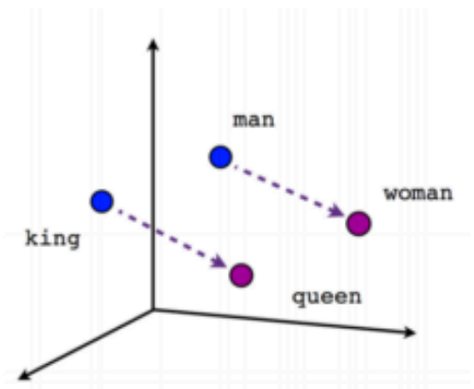
* references
** references
*** references

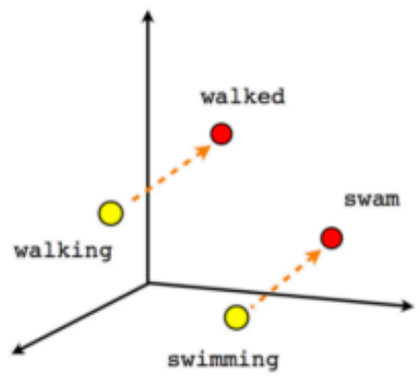
단어를 표현하는 방법

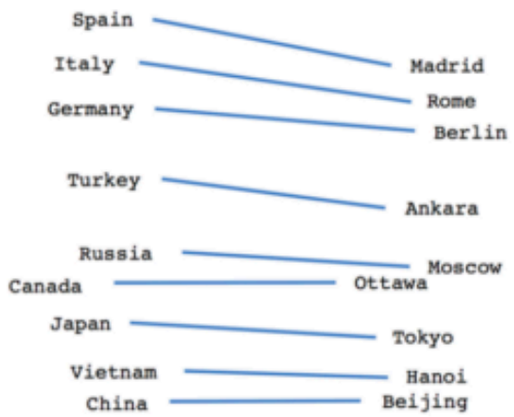
분산표상 (Distributed Similarity Based Representation, DSBR)

- 단어의 의미를 그 단어 주변에 출현하는 단어의 분포로 표현하는 방법
- 원-핫 인코딩이 단어와 단어 사이의 관계를 설명할 수 없는 단점을 해결하기 위해, N차원 단어 벡터를 그보다 훨씬 적은 n차원 벡터로 표현하는 방법









* Tensor Flow, Vector Representations of Words, https://www.tensorflow.org/tutorials/representation/word2vec/.
** references
*** references

단어 가중치: 네트워크 중심성

Review

동시출현 분석 (Co-word Analysis)

- 문서에 서로다른 두 단어의 동시출현 횟수와 네트워크 중심성을 통해 단어의 특징을 표현하는 방법
- 두 단어 사이의 동시출현을 연관성의 척도로 취급하고, 그 관계를 네트워크 중심성으로 표현하여 가중치를 계산함
- **연관어** (공기어, Co-word) : 하나의 문서에서 함께 출현하여 서로 밀접한 의미관계를 가지는 단어

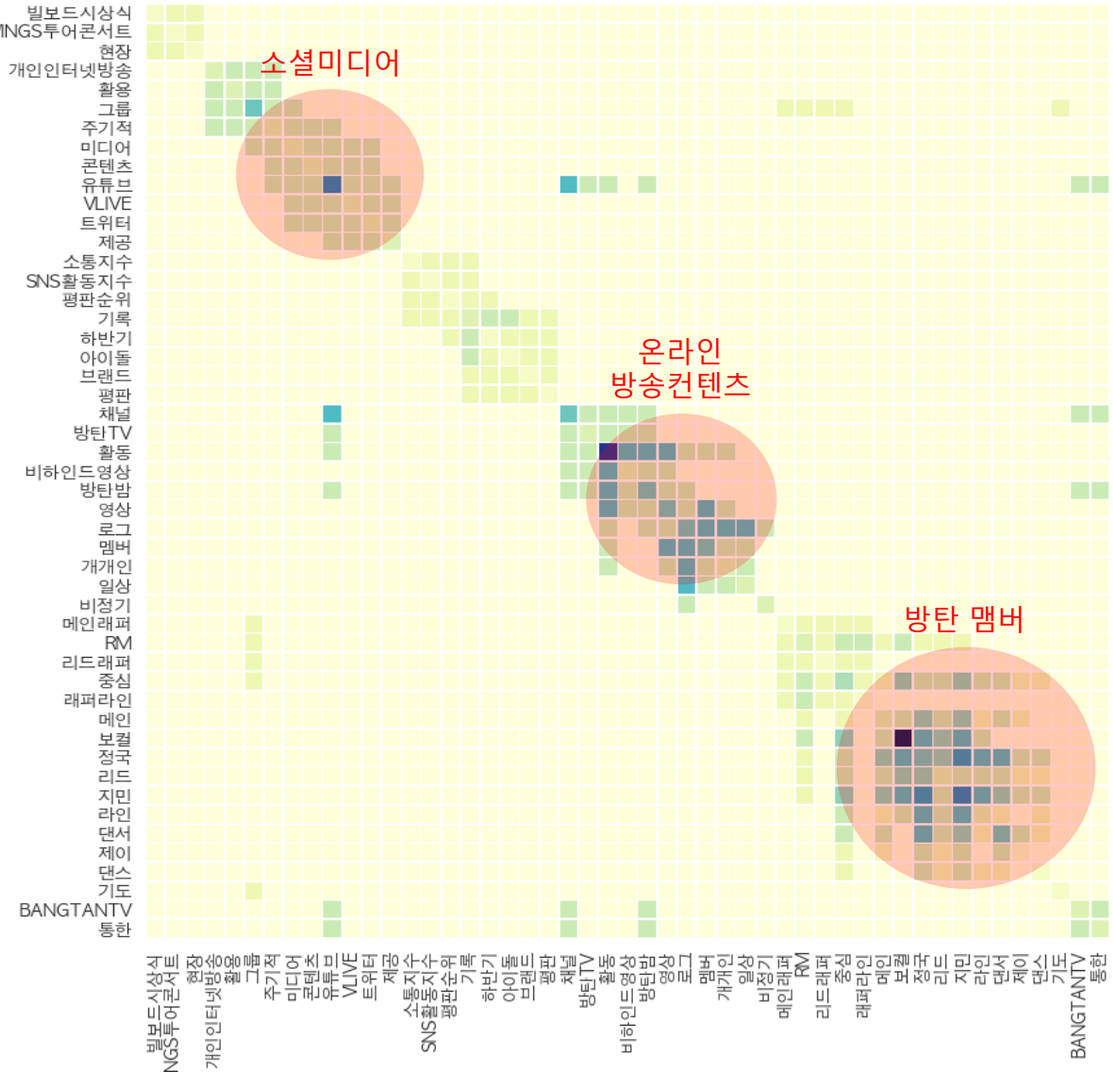


표1 '야쿠르트 아줌마' 연관어 변화

야쿠르트 아줌마는 여전히 '야쿠르트'와의 연관도가 가장 높지만 2016년 들어 '커피' 및 '크림치즈' 제품 연관어와 '10일'이라는 키워드가 등장. 야쿠르트 아줌마는 '배달하는' 역할에서 맛난 제품을 위해 '만나고' '찾고' '발견하는' 대상으로 변화중.

2013년			2014년			2015년			2016년		
No.	연관어	언급 비중	No.	연관어	언급 비중	No.	연관어	언급 비중	No.	연관어	언급 비중
1	야쿠르트	21.3%	1	야쿠르트	26.3%	1	야쿠르트	26.6%	1	야쿠르트	13.1%
2	먹다	4.9%	2	건강	4.5%	2	집	4.7%	2	콜드브루	8.2%
3	아침	4.4%	3	아침	4.0%	3	아침	4.4%	3	커피	7.4%
4	엄마	4.2%	4	집	3.6%	4	맛	3.9%	4	맛	6.6%
5	집	3.5%	5	제품	3.4%	5	먹다	3.4%	5	끼리	5.7%
6	오다	2.8%	6	엄마	3.3%	6	사다	2.8%	6	치즈	5.3%
7	사다	2.7%	7	맛	2.7%	7	주다	2.8%	7	과자	5.0%
8	주다	2.5%	8	같다	2.6%	8	다니다	2.7%	8	아메리카노	4.1%
9	구입하다	2.4%	9	우유	2.6%	9	엄마	2.6%	9	먹다	3.3%
10	아이	2.4%	10	주다	2.2%	10	우유	2.1%	10	크림치즈	3.1%
11	야쿠르트 주다	2.3%	11	먹다	2.2%	11	만나다	2.1%	11	라떼	2.8%
12	배달하다	2.3%	12	만나다	2.0%	12	제품	2.0%	12	만나다	2.7%
13	수입	2.3%	13	사다	1.9%	13	사진	2.0%	13	가격	2.4%
14	다니다	2.1%	14	알다	1.9%	14	나오다	2.0%	14	찾다	1.9%
15	얼려먹다	2.0%	15	배달하다	1.8%	15	팔다	1.9%	15	아침	1.8%
16	살다	2.0%	16	다니다	1.8%	16	지나가다	1.8%	16	10일	1.6%
17	제품	2.0%	17	하루야채	1.7%	17	하나	1.7%	17	엄마	1.5%
18	세븐	1.8%	18	나누다	1.7%	18	판매	1.7%	18	우유	1.4%
19	가다	1.8%	19	지나가다	1.6%	19	일하다	1.6%	19	팔다	1.3%
20	자녀	1.8%	20	세븐	1.5%	20	오다	1.6%	20	발견하다	1.3%
21	만나다	1.8%	21	수입	1.5%	21	찾다	1.6%	21	사다	1.2%
22	마시다	1.7%	22	찾다	2.3%	22	음료	1.5%	22	인기	1.2%
23	유산균	1.7%	23	노인	1.4%	23	마시다	1.4%	23	편의점	1.2%
24	일하다	1.7%	24	마시다	1.4%	24	길	1.4%	24	끼리딥앤크런치	1.1%
	...			...			...			...	
29	팔다	1.4%	29	묻다	1.3%	29	배달하다	1.3%	29	구입하다	1.0%

상승 키워드 하락 키워드 신규 키워드

* 전병진, 신한은행 파이썬으로 시작하는 데이터분석: 텍스트 마이닝 기초, 2018.12.12.
** 백경혜(DBR), "매력을 소비하는 나는 덕후! 즐거움을 위해 기꺼이 지갑을 연다", 2017.1., http://dbr.donga.com/article/view/1203/article_no/7935/.
*** references

단어를 표현하는 방법

동시출현 매트릭스 (Co-occurrence Matrix)

- 문서에 포함된 NxN개의 단어쌍과 단어 사이의 동시출현빈도를 NxN 벡터로 표현하는 방법
- 단어 간 동시출현은 하나의 단어를 기준으로 앞뒤에 존재하는 k개 단어로 정의하며, 문장 단위로 정의하기도 함

구분	문장
문장 A	안녕하세요. 저는 워너원 멤버 강다니엘 입니다.
문장 B	안녕하세요. 얘는 트와이스 멤버 정연 입니다.
문장 C	안녕하세요. 저분은 워너원 멤버 황민현 입니다.

구분	유니그램 (품사=NNG|NNP, 불용어 제거)
문장 A	안녕, 워너원, 멤버, 강다니엘
문장 B	안녕, 트와이스, 멤버, 정연
문장 C	안녕, 워너원, 멤버, 황민현

구분	안녕	워너원	트와이스	멤버	정연	강다니엘	황민현
안녕	3	2	1	3	1	1	1
워너원	2	2	0	2	0	1	1
트와이스	1	0	1	1	1	0	0
멤버	3	2	1	3	1	1	1
정연	1	0	1	1	1	0	0
강다니엘	1	1	0	1	0	1	0
황민현	1	1	0	1	0	0	1

[문제점]

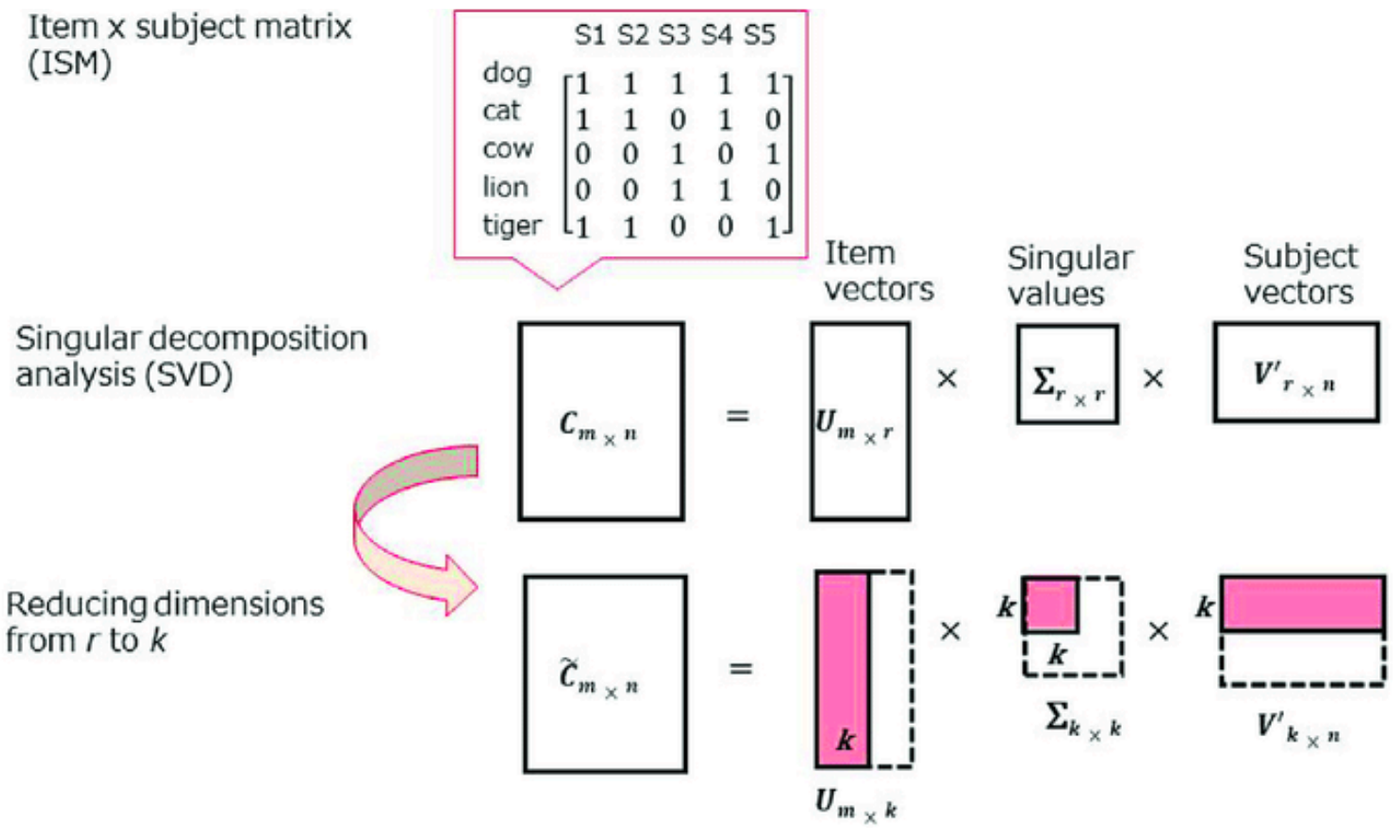
- **차원의 저주(curse of dimensionality)** : 동시출현 매트릭스로 표현하더라도 문서의 양이 많아질수록 차원은 늘어나고, 희소성(sparsity)이 늘어 의미없은 벡터값이 많이 존재함
- ~~각 단어는 서로 관계를 가질 수 없기 때문에($VS_T = 0$), 유의어, 반의어와 같은 관계를 표현할 수 없고 서로 독립적으로만 존재함~~

* references
** references
*** references

단어를 표현하는 방법

SVD (Singular Value Decomposition)

- Singular Value를 활용하여 벡터의 차원을 축하는 방법
- 동시출현 매트릭스에 대해 SVD를 수행하는 경우 유사한 단어끼리 유사한 벡터값을 가지게 됨



[문제점]

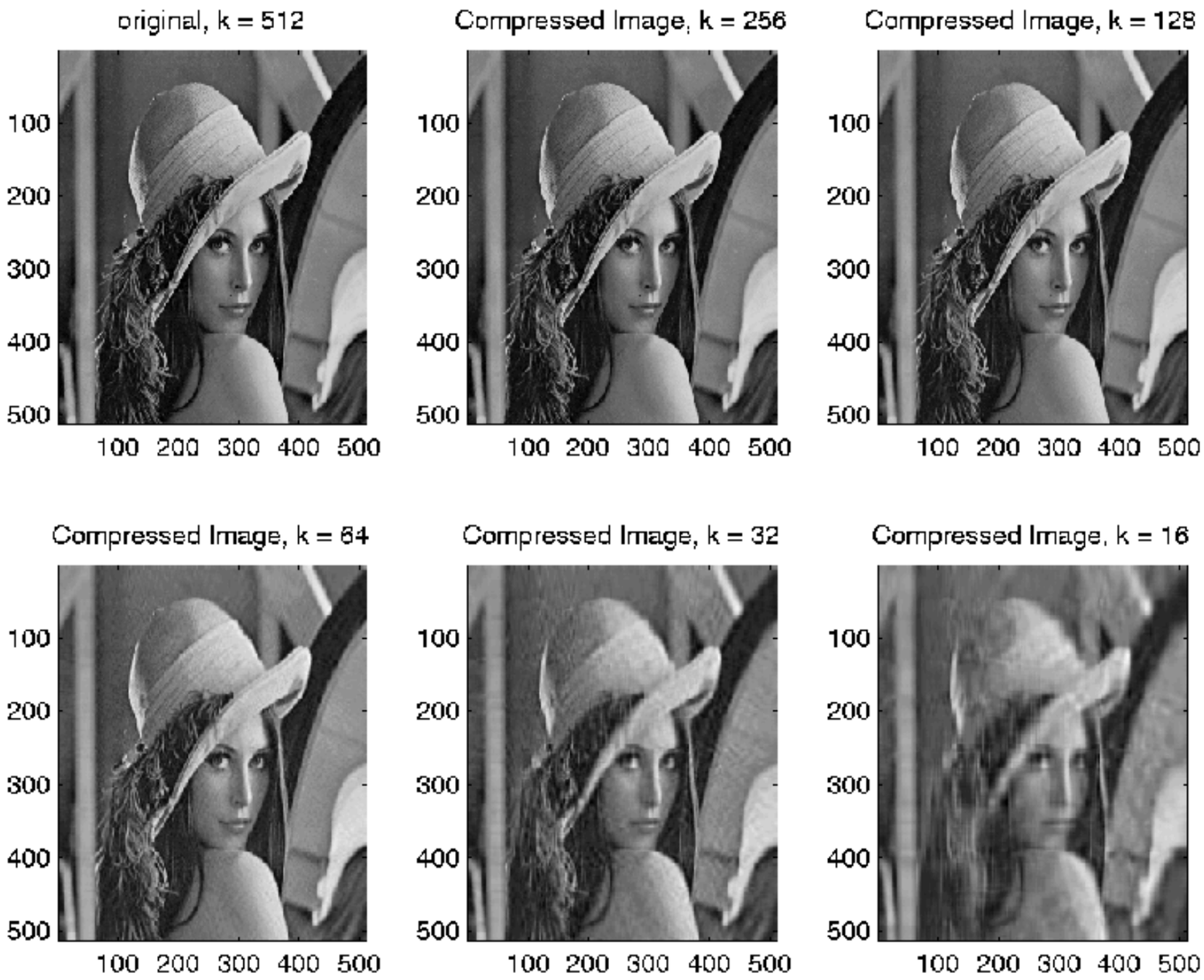
- 계산량이 매우 많이 소요됨

- 1) 사전에 생성한 동시출현 매트릭스에 새로운 단어가 추가되는 경우,
2) 문서의 특성이 바뀌어 동시출현 매트릭스를 다시 생성해야하는 경우 → 매번 다시 SVD를 수행해야함

* Chika Sumiyoshi, et al., 2018.3., Semantic Memory Organization in Japanese Patients With Schizophrenia Examined With Category Fluency, https://www.researchgate.net/publication/323907837_Semantic_Memory_Organization_in_Japanese_Patients_With_Schizophrenia_Examined_With_Category_Fluency/.
** references
*** references

단어를 표현하는 방법

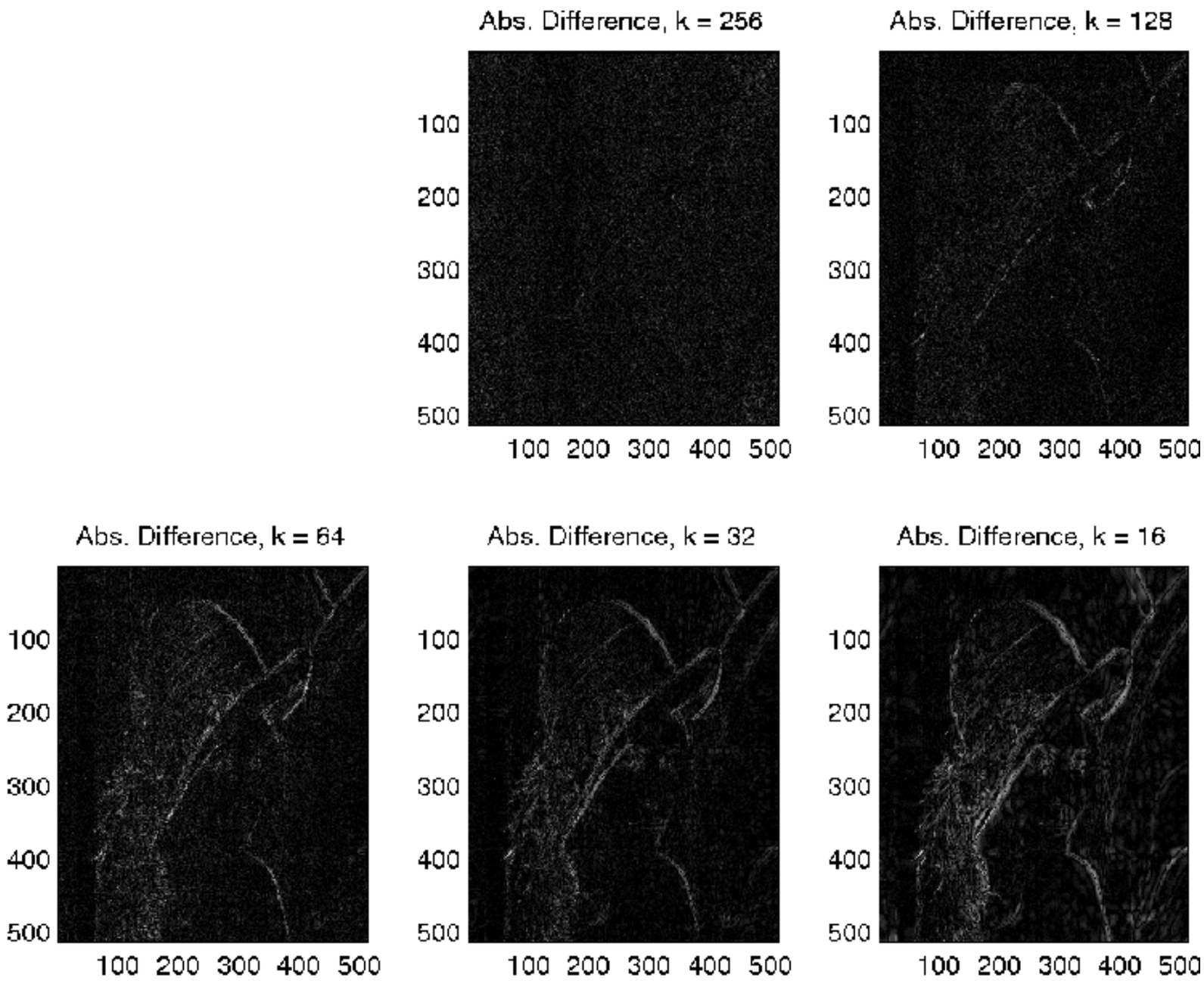
SVD (Singular Value Decomposition)



* references
** references
*** references

단어를 표현하는 방법

SVD (Singular Value Decomposition)



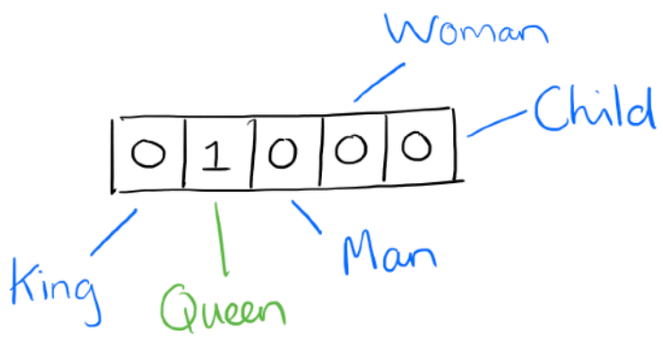
* James Chen, Image Compression with SVD, ECS 289K Scientific Computation.2000.12.13., http://fourier.eng.hmc.edu/e161/lectures/svdcompression.html/.
** references
*** references

단어를 표현하는 방법

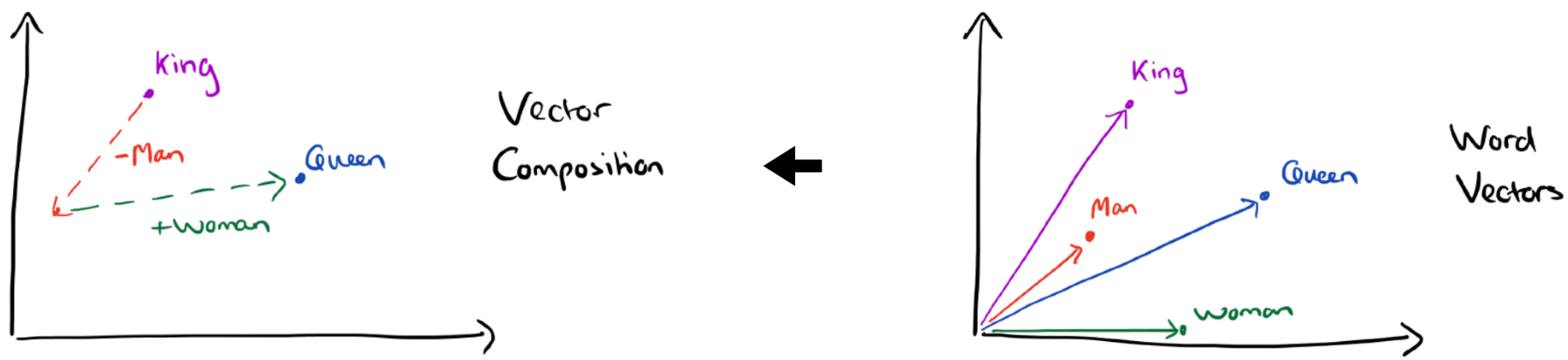
단어 임베딩 (Word Embedding)

- 고차원 상에 존재하는 단어 벡터를 저차원에 상의 실수 벡터로 표현하는 방법 (Word2Vec, Glove, …)
- 글 내부에서 가까이 위치해 있는 단어끼리는 유사한 의미를 지닌다는 가정(distributional hypothesis)을 기반으로,
  벡터 공간에서 각 단어들이 어떻게 분포해 있는지를 학습하는 방법

[One-hot encoding]



[Distributed Similarity-based Representation]

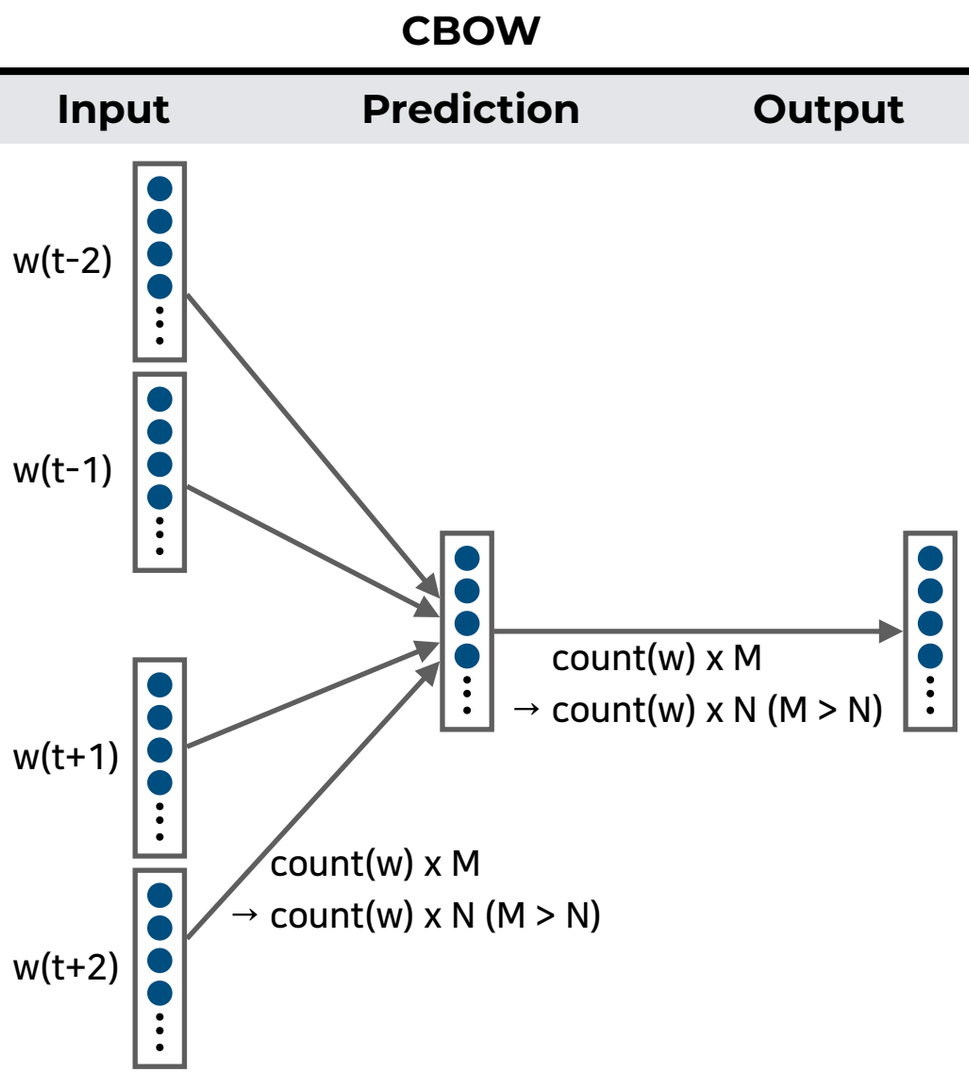


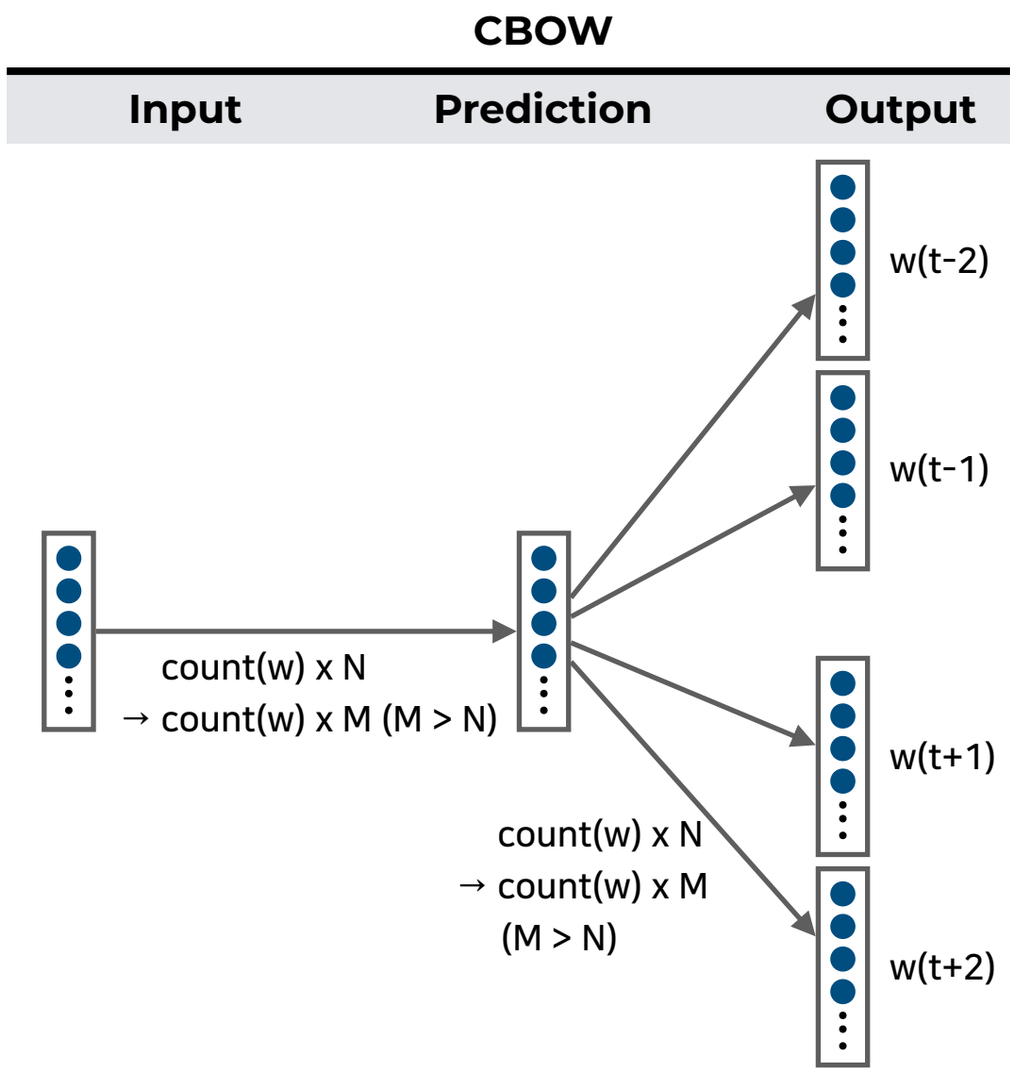
* The Morning Paper, The amazing power of word vectors, 2016.4.21., https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/.
** references
*** references

단어 임베딩 기법: Word2Vec

Word2Vec (Word to Vector, Mikolov et al., 2013)

- 2013년 구글에서 발표된 연구주제로 단어와 단어의 동시출현 단어 정보를 기반으로 단어를 벡터로 바꾸는 방법
- CBOW (Continuous Bag of Words) : 주변에 있는 단어 정보를 통해 중심에 있는 단어를 맞추는 방법
- Skip-gram : 중심에 있는 단어를 기반으로 주변에 있는 단어를 예측하는 방법



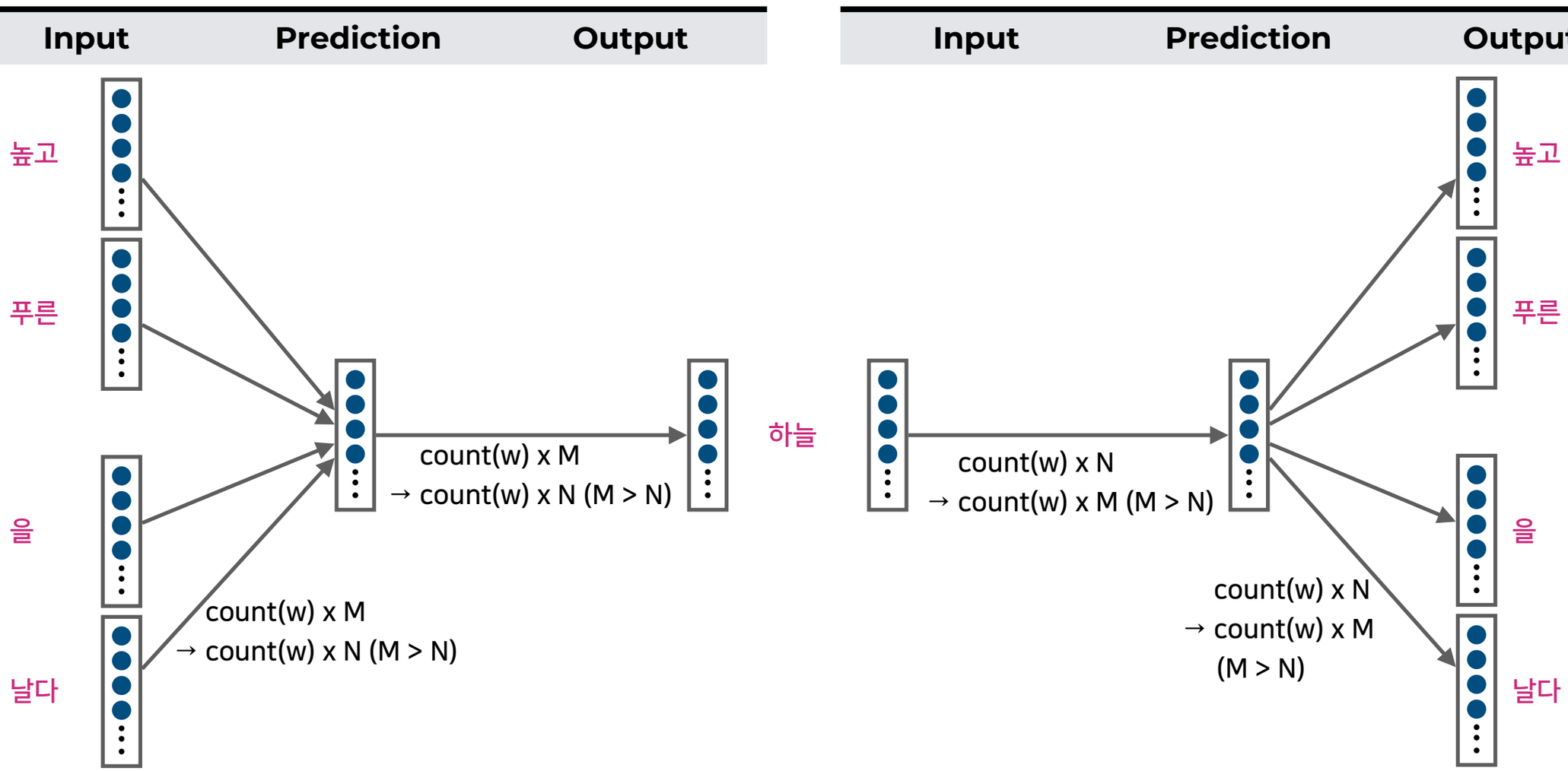


* Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).
** references
*** references

단어 임베딩 기법: Word2Vec

Word2Vec (Word to Vector, Mikolov et al., 2013)

- 2013년 구글에서 발표된 연구주제로 단어와 단어의 동시출현 단어 정보를 기반으로 단어를 벡터로 바꾸는 방법
- CBOW (Continuous Bag of Words) : 주변에 있는 단어 정보를 통해 중심에 있는 단어를 맞추는 방법
- Skip-gram : 중심에 있는 단어를 기반으로 주변에 있는 단어를 예측하는 방법

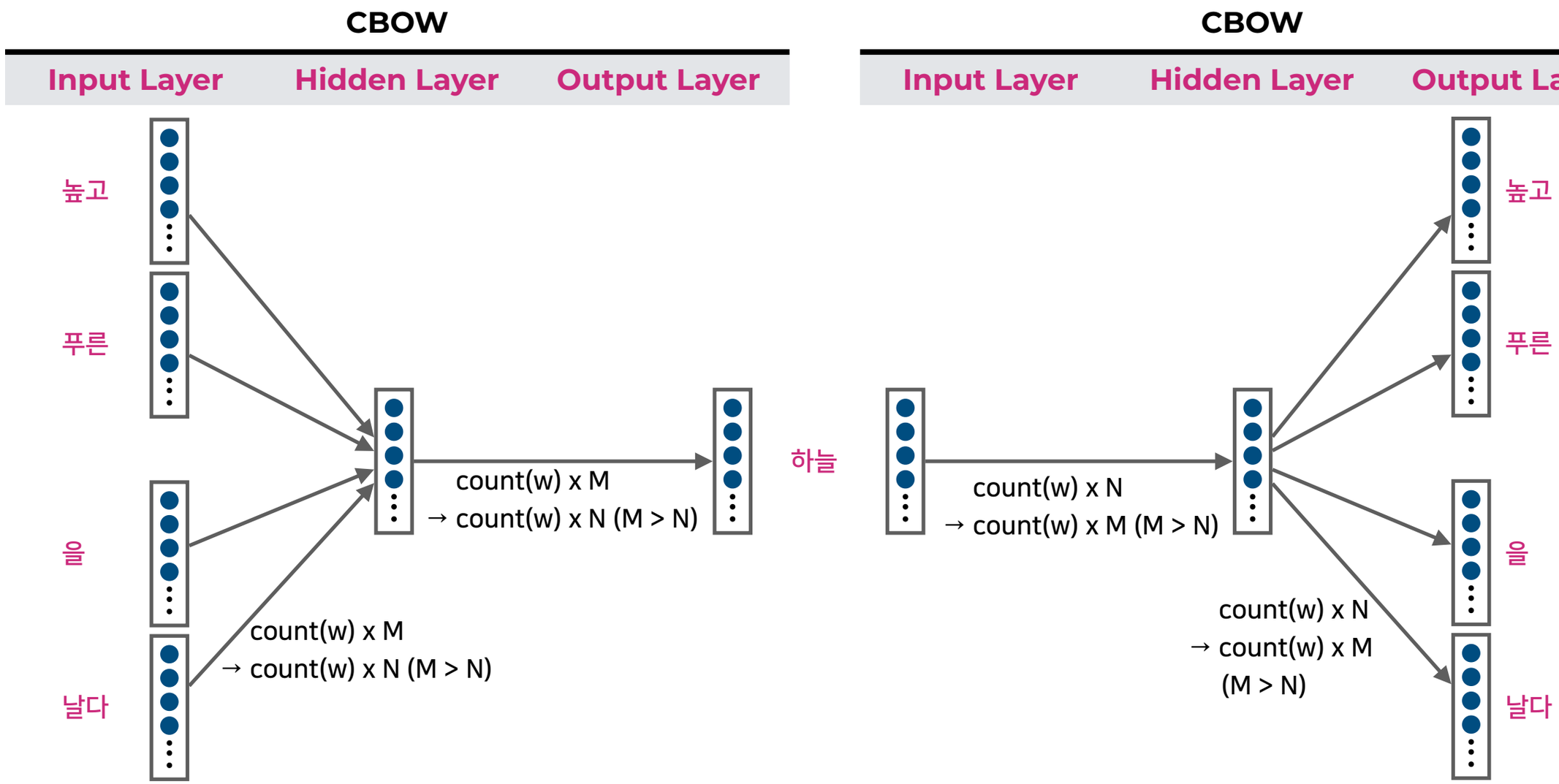


* Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).
** references
*** references

단어 임베딩 기법: Word2Vec

Word2Vec (Word to Vector, Mikolov et al., 2013)

- 2013년 구글에서 발표된 연구주제로 단어와 단어의 동시출현 단어 정보를 기반으로 단어를 벡터로 바꾸는 방법
- CBOW (Continuous Bag of Words) : 주변에 있는 단어 정보를 통해 중심에 있는 단어를 맞추는 방법
- Skip-gram : 중심에 있는 단어를 기반으로 주변에 있는 단어를 예측하는 방법



* Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).
** references
*** references

단어 임베딩 기법: Word2Vec

Word2Vec (Word to Vector, Mikolov et al., 2013)

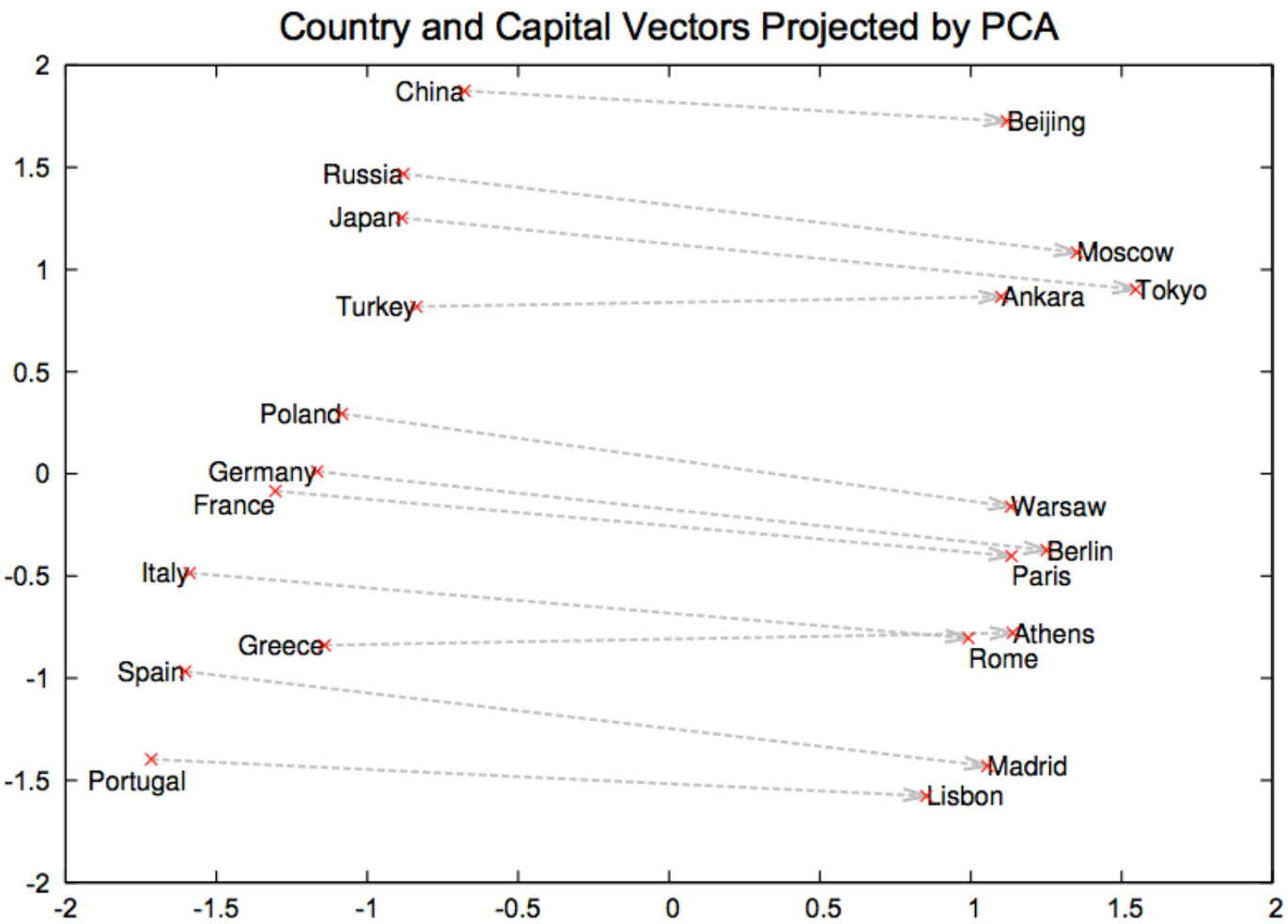


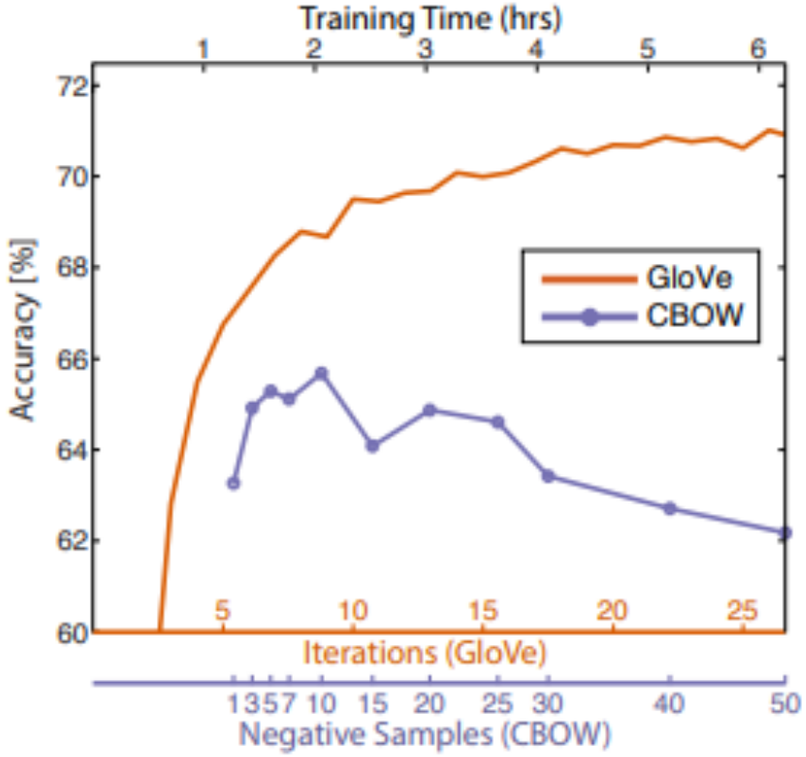
Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

* Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).
** references
*** references

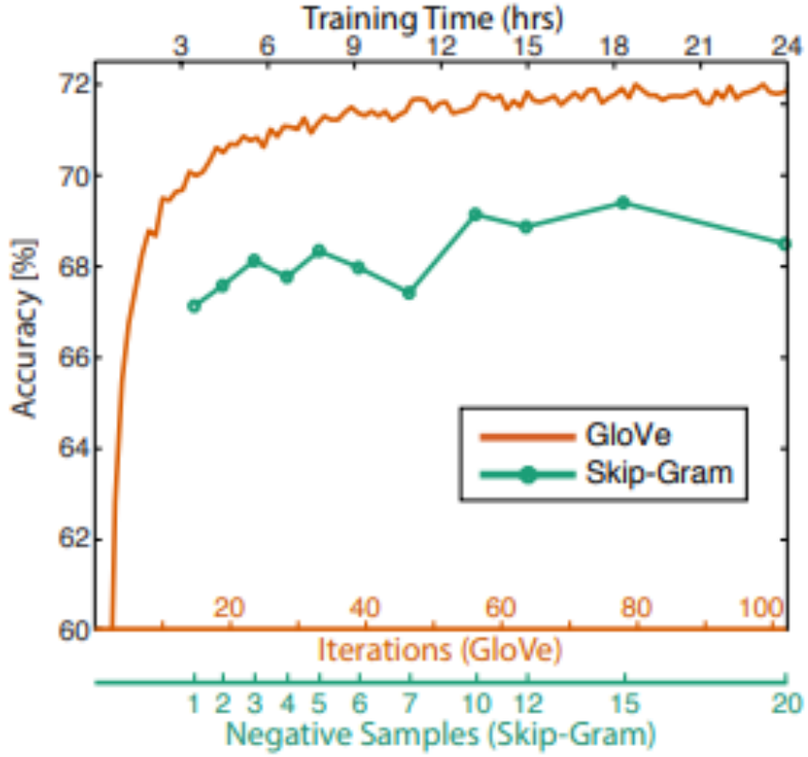
단어 임베딩 기법: GloVe

GloVe (Global Vectors for Word Representation,)

- 2014년 미국 스탠포드대학에서 발표한 연구주제로, 단어의 동시출현 확률을 기반으로 단어를 벡터로 바꾸는 방법
- Word2Vec 대비 장점
  1) 학습 시간이 비교적 빠르며, 매우 큰 문서에 대해서도 균일한 성능을 보이
  2) 적은 문서에 대해서도 좋은 성능을 보임 (벡터 크기가 작아도 좋은 성능을 보임)
  3) 빈도가 적은 단어를 비교적 잘 표현할 수 있음



(a) GloVe vs CBOW



(b) GloVe vs Skip-Gram

* Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).
** Jeffrey Pennington, Richard Socher, and Christopher D. Manning, GloVe: Global Vectors for Word Representation, https://nlp.stanford.edu/projects/glove/.
*** references

단어 임베딩 기법: GloVe

GloVe (Global Vectors for Word Representation)

- 단어와 단어 사이의 동시출현 확률의 비율이 커지는 방향으로 학습함

(Long-bilinear model)

$$w_1 \cdot w_2 = \log P(i \mid j)$$

(Vector differences)

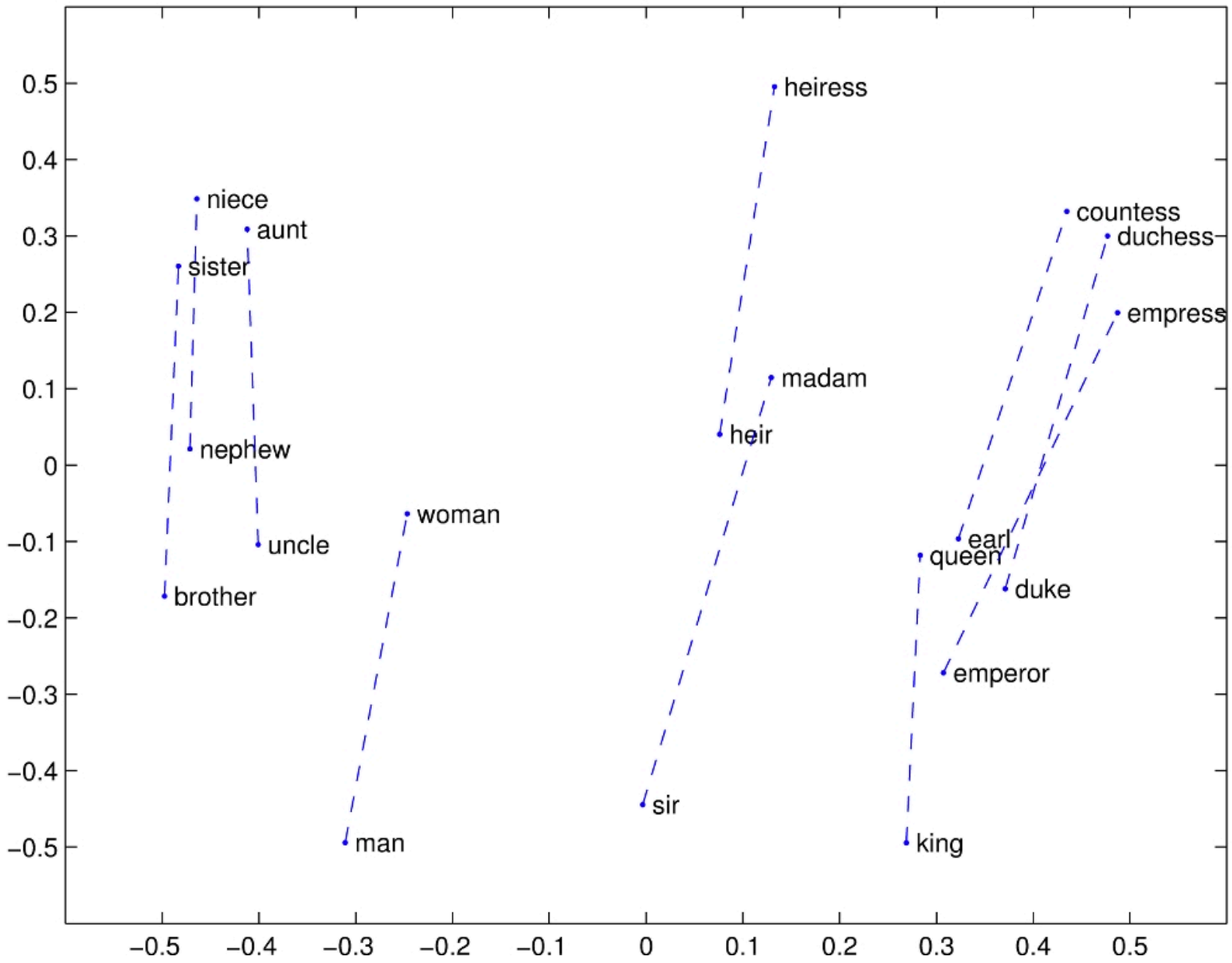
$$w_1 \cdot (w_a - w_b) = \log \frac{P(x|a)}{P(a|b)}$$

구분	패딩	가디건	계절	Random
P(단어|겨울)	0.8	0.2	0.9	0.1
P(단어|가을)	0.2	0.8	0.9	0.1
$\frac{P(단어\|겨울)}{P(단어\|가을)}$	4	0.25	1	1

* Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).
** Jeffrey Pennington, Richard Socher, and Christopher D. Manning, GloVe: Global Vectors for Word Representation, https://nlp.stanford.edu/projects/glove/.
*** references

단어 임베딩 기법: GloVe

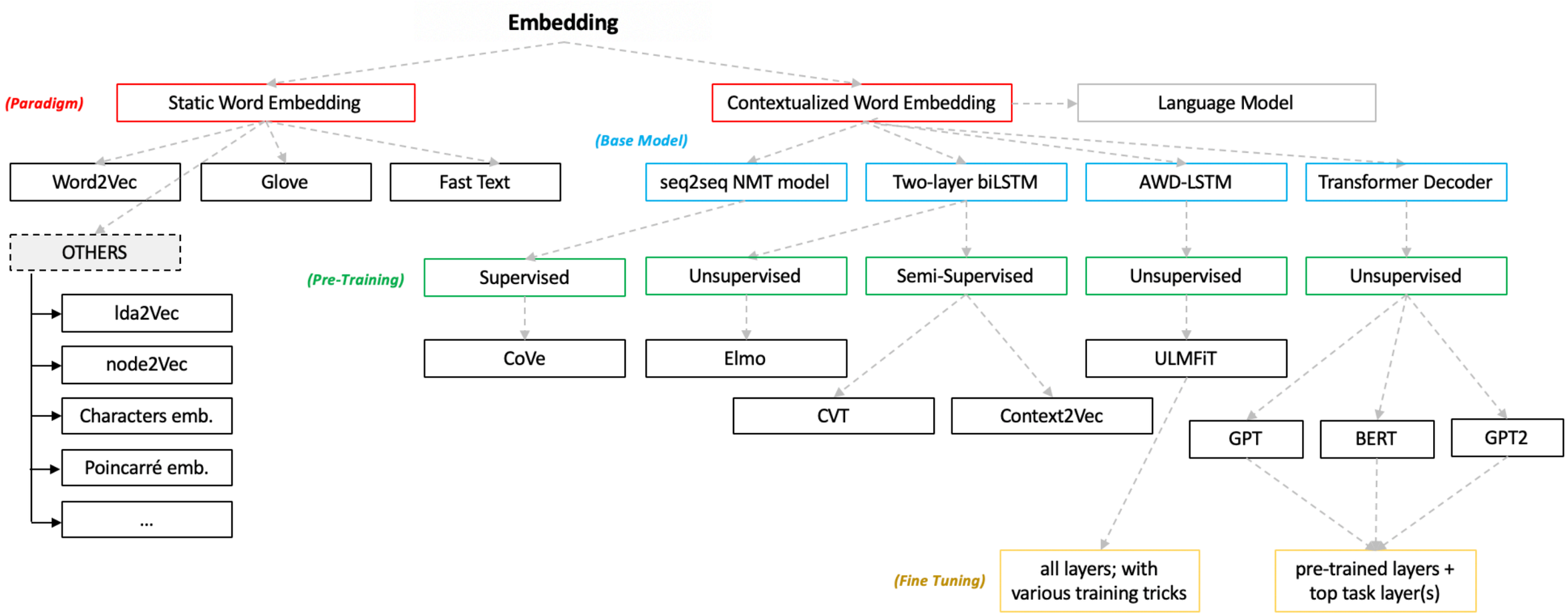
GloVe (Global Vectors for Word Representation)

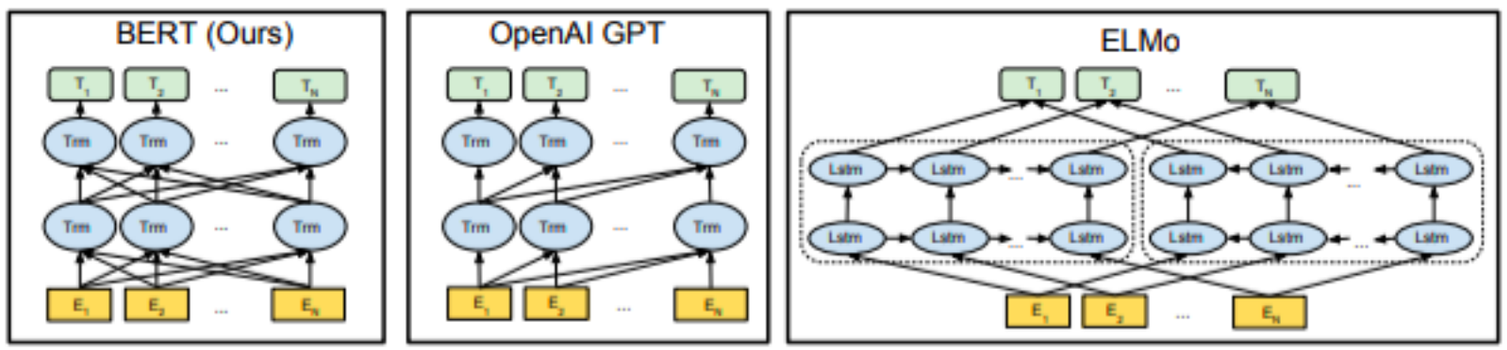


* Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).
** Jeffrey Pennington, Richard Socher, and Christopher D. Manning, GloVe: Global Vectors for Word Representation, https://nlp.stanford.edu/projects/glove/.
*** references

Too Many Embedding...

©AdrienSIEG





* Adrien Sieg, FROM Pre-trained Word Embeddings TO Pre-trained Language Models — Focus on BERT, 2019.8.30., https://towardsdatascience.com/from-pre-trained-word-embeddings-to-pre-trained-language-models-focus-on-bert-343815627598/.
** references
*** references

E.O.D

Contact
http://www.teanaps.com
fingeredman@gmail.com