Chapter. 05

자연어 예측

# 자연어 모델링 III

FASTCAMPUS

ONLINE

금융공학/퀀트 I

강사. 장순용

# 키포인트

- BOW와 TF IDF 모형의 문제점.
- Embedding 모형.

# I BOW와 TF IDF

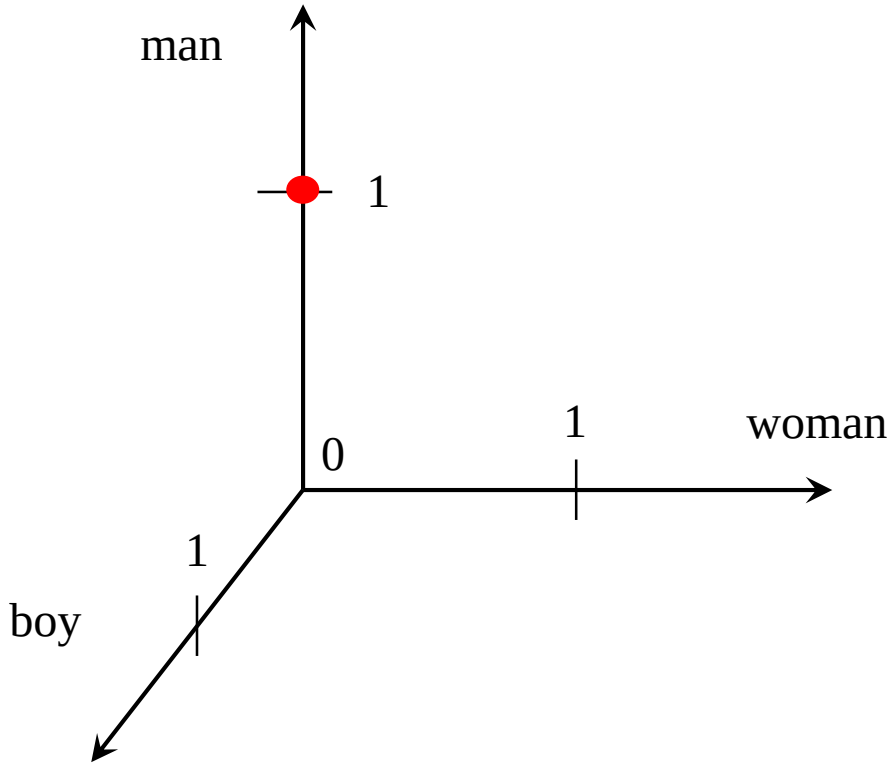
- BOW와 TF IDF의 문제점.

  ⇒ Feature (단어)의 가지 수 만큼 차원이 폭발적으로 증가한다.

  ⇒ Feature (단어)에 내재된 관계가 반영되지 않았다.

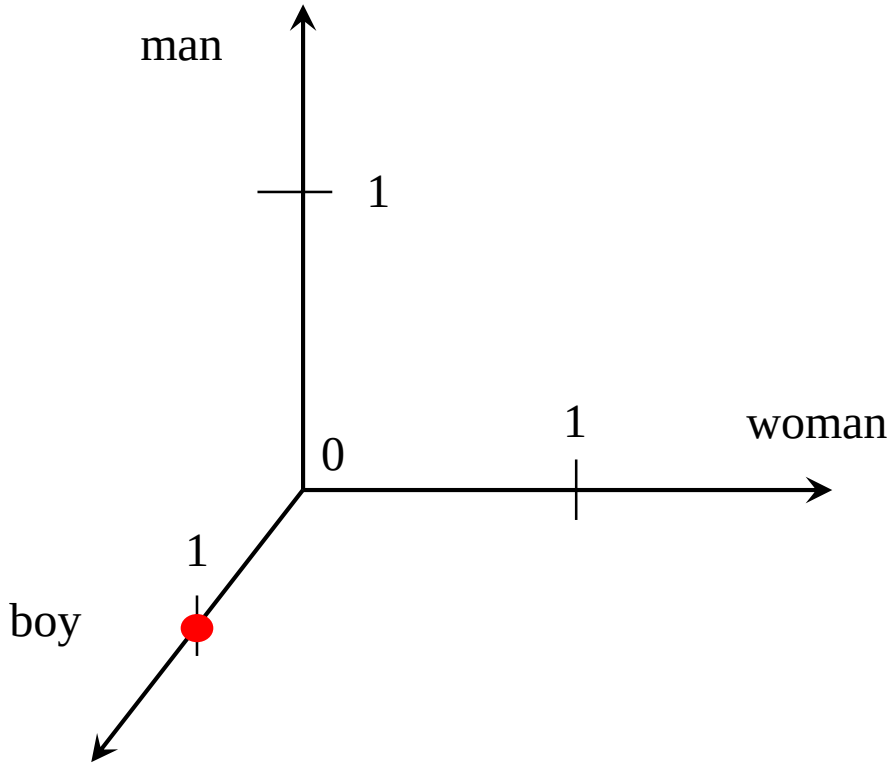
# I BOW와 TF IDF

- BOW의 경우: 1 또는 0.



# I BOW와 TF IDF

- BOW의 경우: 1 또는 0.

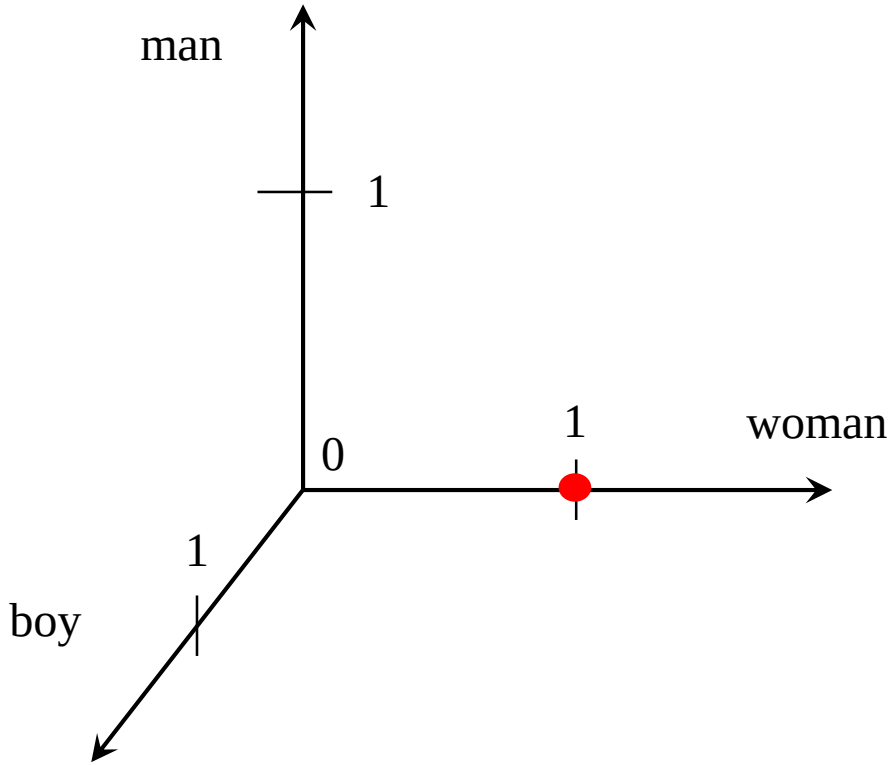





FAST CAMPUS
ONLINE
장순용 강사.

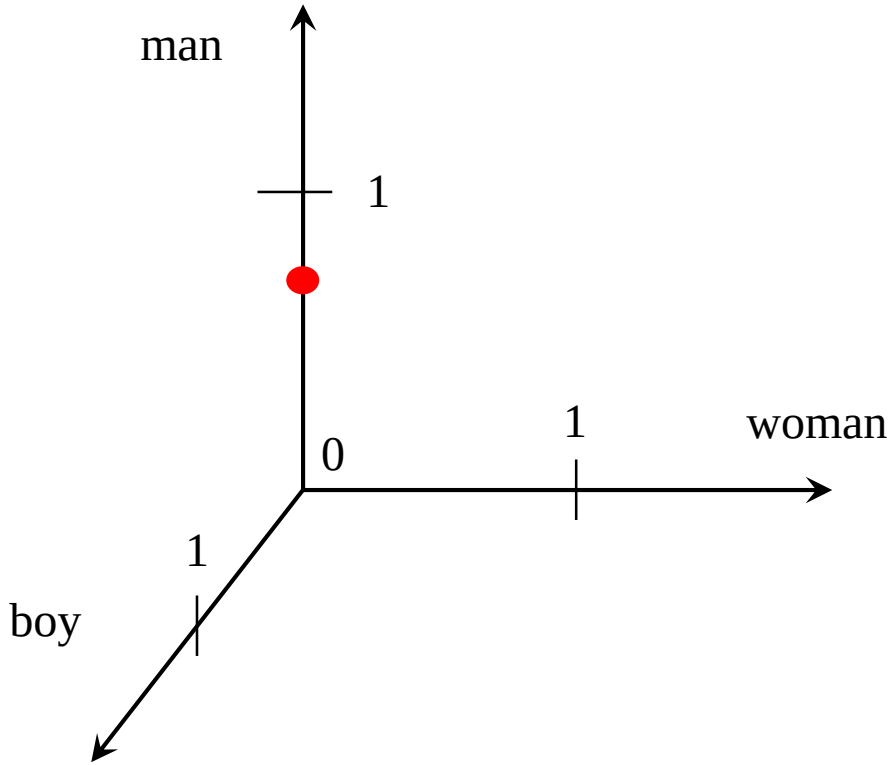
# I BOW와 TF IDF

- BOW의 경우: 1 또는 0.



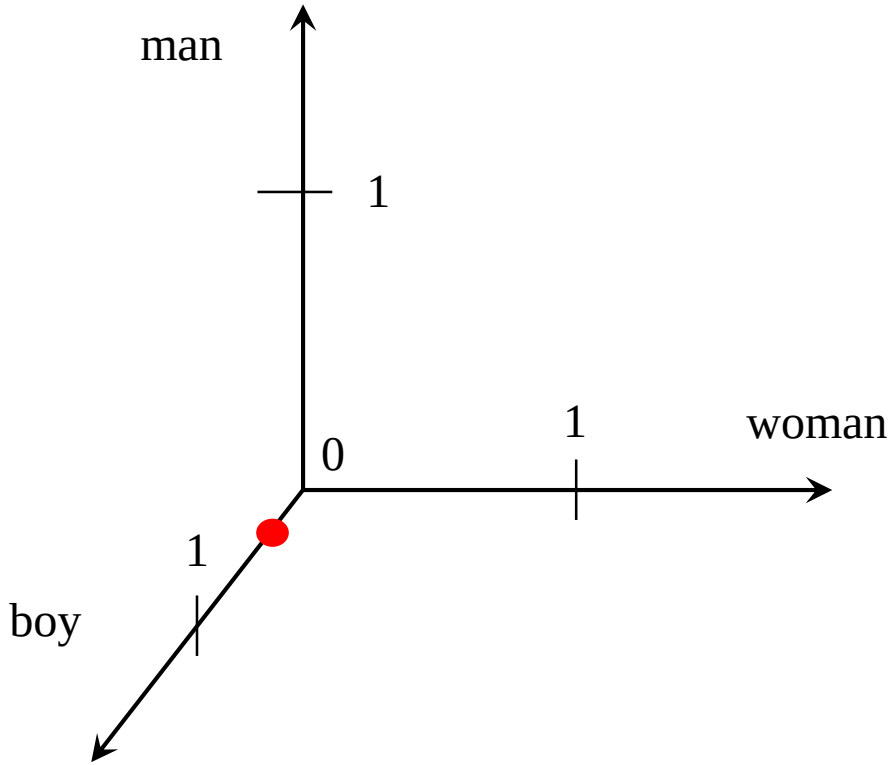
# I BOW와 TF IDF

- TF IDF의 경우: 1, 0 이외의 성분을 가질 수 있다.



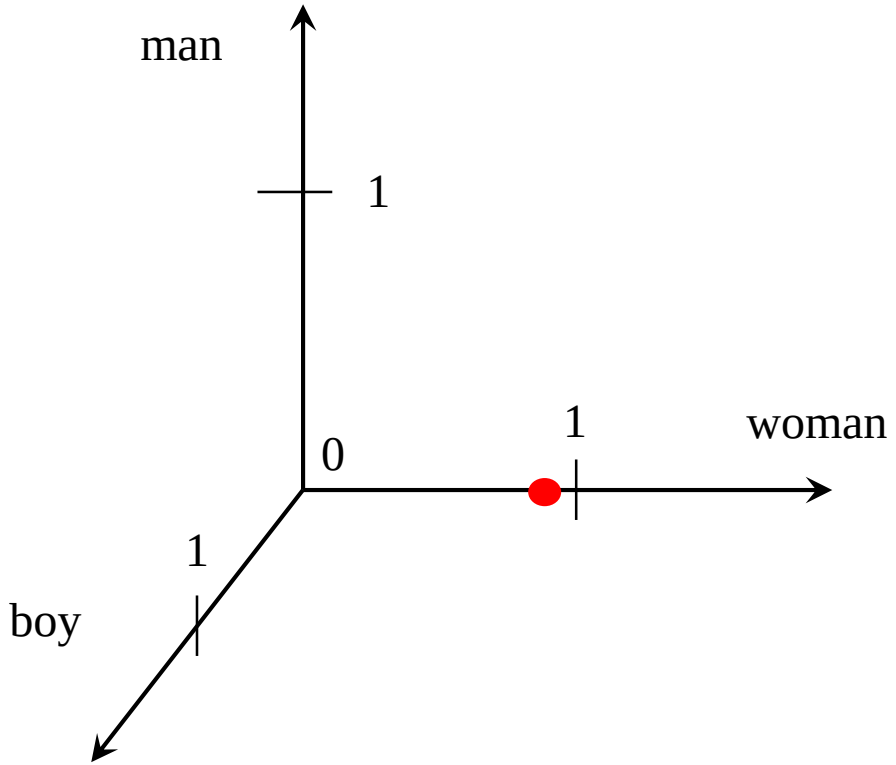
# I BOW와 TF IDF

- TF IDF의 경우: 1, 0 이외의 성분을 가질 수 있다.



# I BOW와 TF IDF

- TF IDF의 경우: 1, 0 이외의 성분을 가질 수 있다.






FAST CAMPUS
ONLINE
장순용 강사.

# Embedding 모형



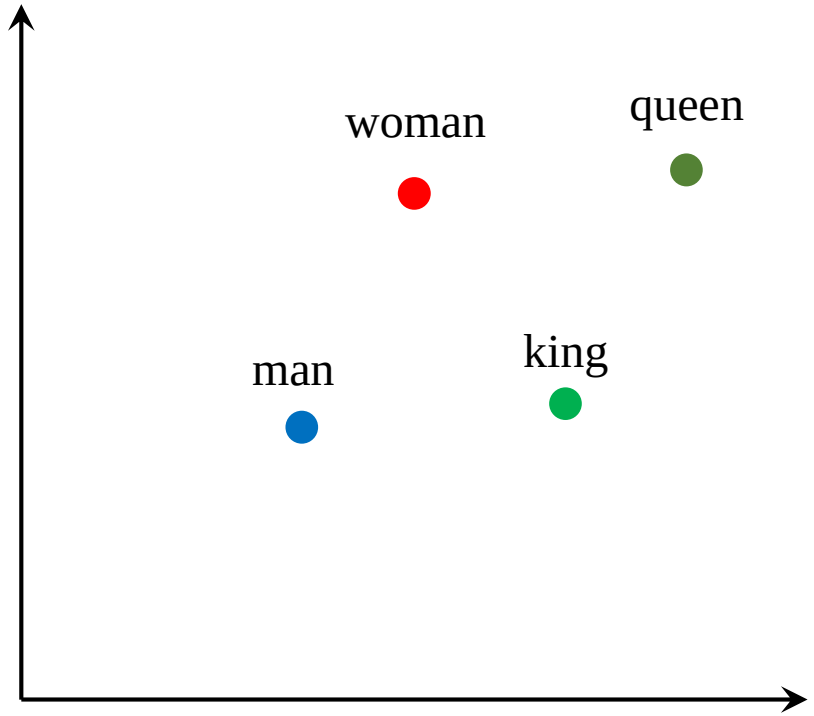
- 비교:

| BOW, TF-IDF | Embedding |
|---|---|
| Feature 공간의 차원이 크다. | Feature 공간의 차원수가 제한적이다. |
| Sparse 행렬 (벡터)로 표현된다. | Dense 행렬 (벡터)로 표현된다. |
| Feature의 의미 (관계)가 반영되지 않음. | Feature의 의미 (관계)가 반영됨. |

# Embedding 모형



- Embedding 모형은 컴팩트한 representation에 기반한다.

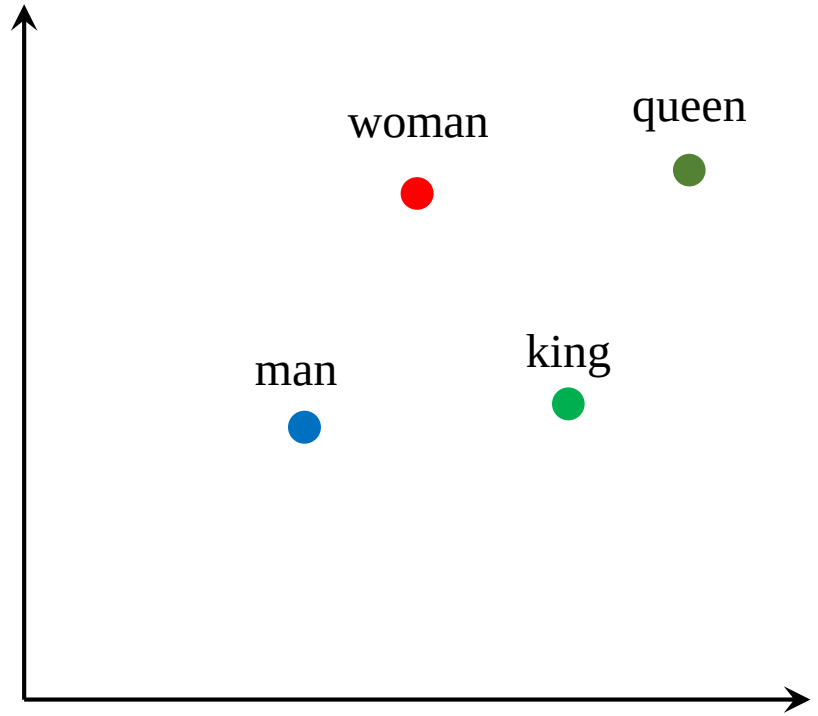


# Embedding 모형



- 벡터 사이에 다음과 같은 연산이 가능하다.

$$\text{Queen} - \text{Woman} + \text{Man} = \text{King}$$



# I Embedding 모형의 원리

- CBOW (Continuous Bag of Words)와 Skip-Gram이 있다.

⇒ 인공 신경망으로 문장 예측모형을 만들어서 학습 시킨다.

⇒ 단어는 one-hot-encoding된 벡터의 형태로 입출력 된다.

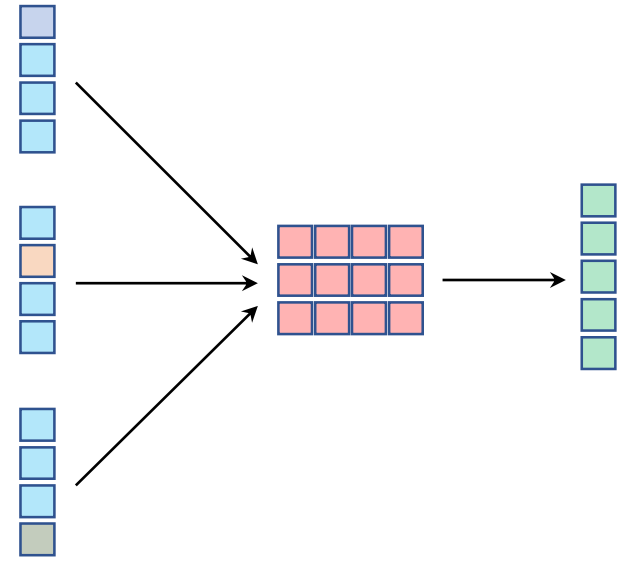
⇒ 학습된 가중치 행렬에서 Embedding 벡터를 추출한다.

# Embedding 모형의 원리: CBOW

- CBOW는 주변의 단어를 바탕으로 하나의 단어를 예측하는 모형이다.

나는 빨간 사과를 깨물어 먹었다

나는 빨간 _____ 깨물어 먹었다

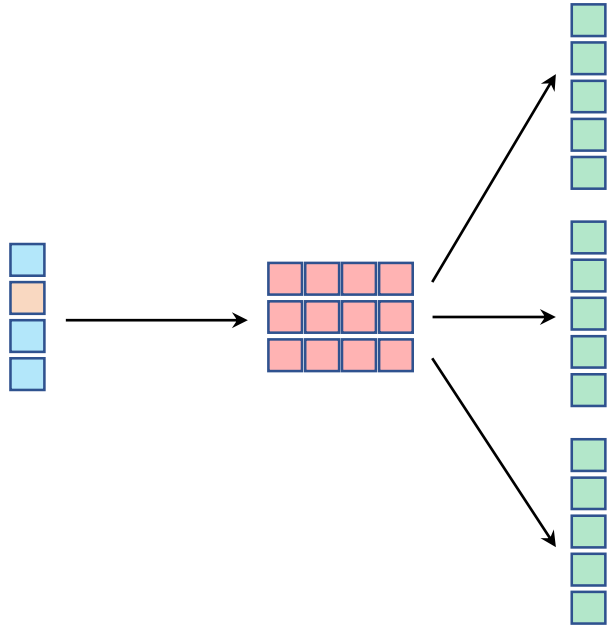


# I Embedding 모형의 원리: Skip-Gram

- Skip-Gram은 하나의 단어를 가지고 주변을 예측하는 모형이다.

나는 빨간 사과를 깨물어 먹었다

_____ _____ 사과를 _____ _____



끝.



# 감사합니다.

FAST CAMPUS
ONLINE
장순용 강사.