# Natural Language Processing & Deep Learning

지승훈

# 차례

- 자연어처리(Natural Language Processing)
- 기계 학습(Machine Learning)과 딥 러닝(Deep Learning)

# 자연어처리(Natural Language Processing)

- 말 그대로 자연어(인간의 언어)를 처리하는 것을 연구하는 학문 분야
- 위키피디아는 컴퓨터과학, 정보공학, 인공지능의 분야로 취급하고 있음.
- 대표적인 세부 분야는 기계번역과 음성인식(Speech Language Processing)

- 자연어는 방대하고, 예측불가능하며, definitive하지 않기 때문에 정량적 처리를 위해서는 모델링이 필요하다. -> Language Model

# 자연어처리의 기타 세부 분야

- 표제어 추출(Lemmatization)과 어간 추출(Stemming)
- 형태소 태깅(Part-of-speech tagging)
- 파싱(구문 분석, Parsing)
- 광학 문자 인식(Optical character recognition)
- 감정 분석(Sentiment analysis)
- 자동 요약(Automatic summarization)
- 음성 합성(Text-to-speech)

# 자연어처리 연구의 흐름: Rule-based NLP

- 인간이 직접 상정한 규칙들을 이용해 처리함
- 촘스키 언어학과 닮은 점이 있는 듯.

- 인간이 만드는 규칙이기 때문에 규칙 자체에 오류 가능성이 있으며, 규칙에 없는 것은 처리하기 어려움

# Rule-based NLP Example: nltk.RegexpParser

```
from nltk import RegexpParser
sentence = [(“John", "NNP"),
("is", "VB"), ("a", "DT"), ("boy",
"NN")]
grammar = “””
        NP: {<DT>?<NN.*>}
        VP: {<VB><NP>}
        S: {<NP><VP>}
“””
parser = RegexpParser(grammar)
print(parser.parse(sentence))
```

```
(S
        (NP John/NNP)
        (VP is/VB
            (NP a/DT boy/NN)
        )
)
```

# 자연어처리 연구의 흐름: Statistical NLP

- 1980년대 기계학습이 도입되고, 통계학적 발전이 크게 이루어졌으며, 코퍼스(corpus/corpora)의 구축에 따라 자연어도 통계학적으로 접근 시작
- 대량의 코퍼스 데이터를 학습하고 그를 바탕으로 자연어 데이터(처음 보는 것까지도)를 처리

# Statistical NLP Example: Markov Chain으로 단어 예측

학습 데이터에서 bigram 등장 확률을 학습하고, 주어진 데이터 이후 등장할 단어를 예측한다.

P(‘국민’|’대한민국’) = 0.23
P(‘대통령’|’대한민국’) = 0.41
P(‘헌법’|’대한민국’) = 0.19

-> ‘대한민국’ 다음에는 ‘대통령’이 등장할 확률이 가장 높다.

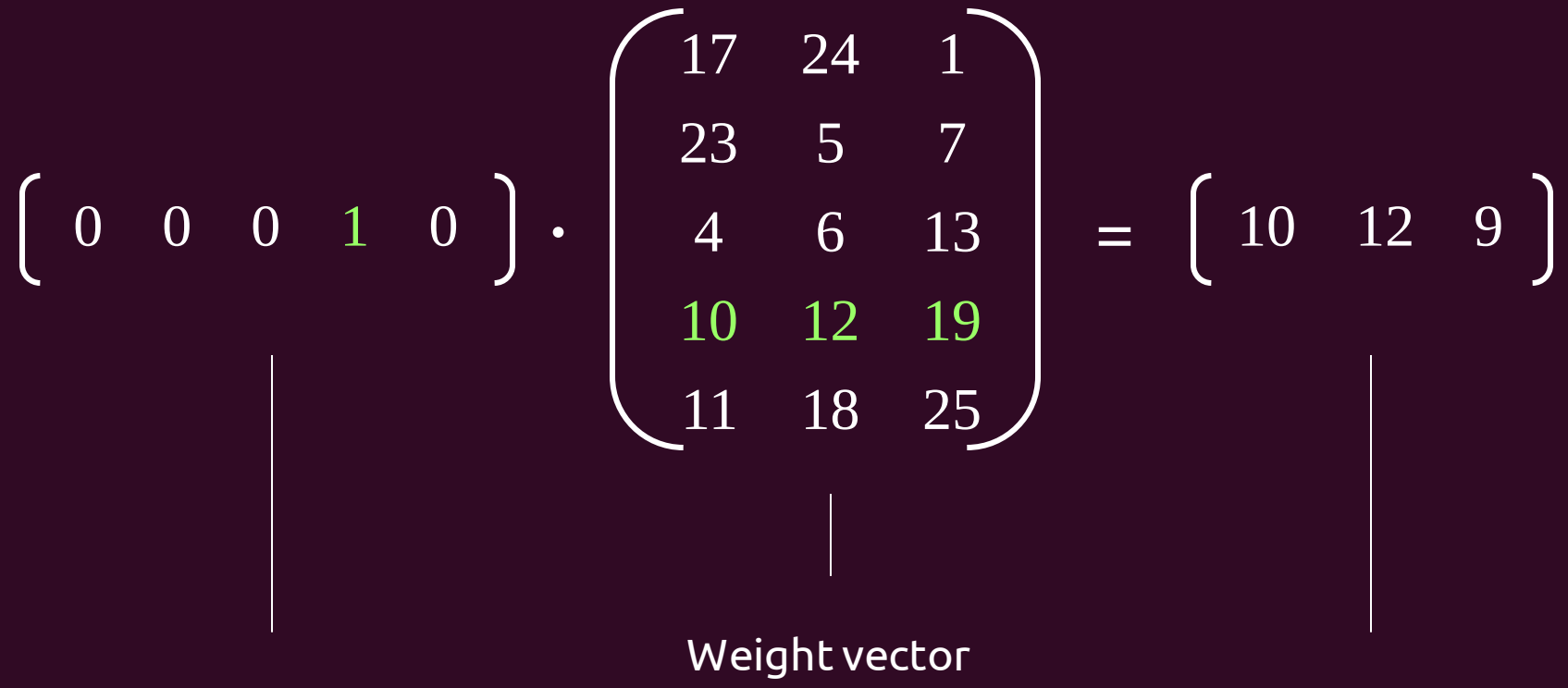
# 자연어처리 연구의 흐름: NLP with Deep Learning

- 2010년대에 대두되기 시작
- 보통 Word embedding을 사용해 많은 세부 분야에서 통계적 NLP의 performance를 뛰어넘고 있음
- Word embedding: “비슷한 위치에서 등장하는 단어들은 비슷한 의미를 가진다”

# Deep Learning NLP Example: Word2Vec

- Word Embedding: 단어별로 고유한 벡터값을 적절히 부여하여 정규화하는 기술
- Word2Vec: 구글에서 개발한 Word Embedding 모델로, 지금까지 나온 다른 Word Embedding 방식보다 월등한 performance를 보여 Word2Vec이라는 용어가 거의 Word Embedding을 지칭하는 보편적인 용어로 사용되고 있음
- 0과 1로 이루어진 'One-Hot-Encoding' 방식 벡터를 더 차원이 낮은 실수 벡터로 만드는 방식
- 단어 간의 유사도는 코사인 유사도를 통해 측정

# Deep Learning NLP Example: Word2Vec



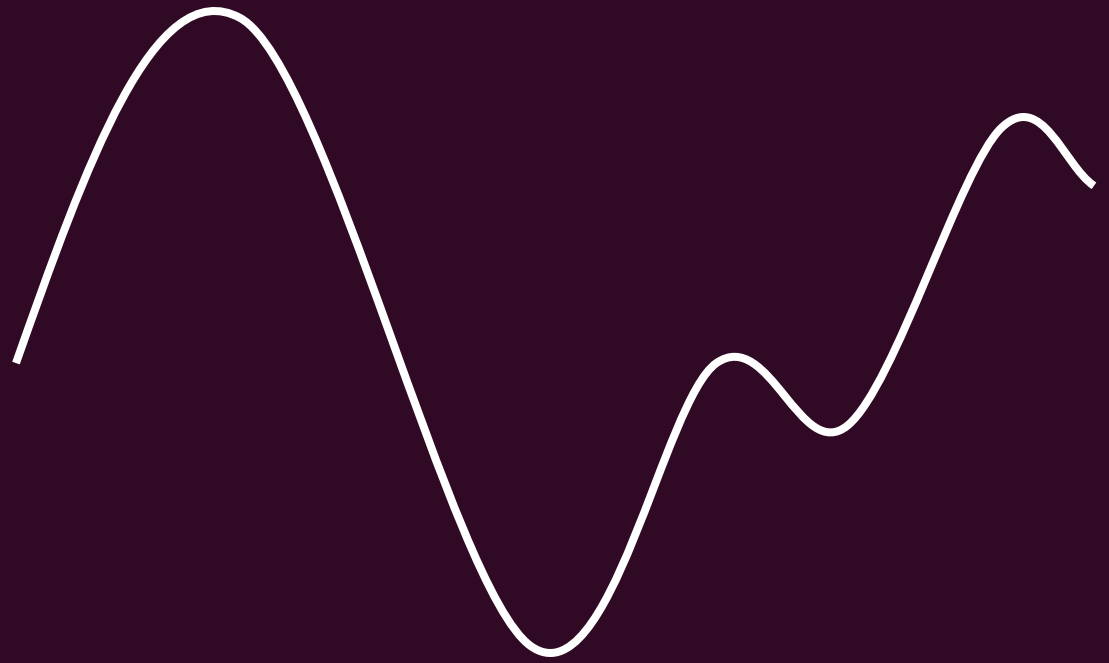
# Deep Learning NLP Example: Word2Vec

대한민국 헌법을 Kkma 형태소 분석기를 이용하고 gensim의 Word2Vec 함수를 이용하여 Word2Vec 모델로 모델링했을 때, '대통령'과 가장 유사한 단어는?

| | | |
|---|---|---|
| 0 | 법률 | 0.775513172... |
| 1 | 헌법 | 0.858850121... |
| 2 | 때 | 0.822009086... |
| 3 | 회의 | 0.795571088... |
| 4 | 국회 | 0.766394972... |
| 5 | 국가 | 0.760404527... |

# 기계 학습(Machine Learning)

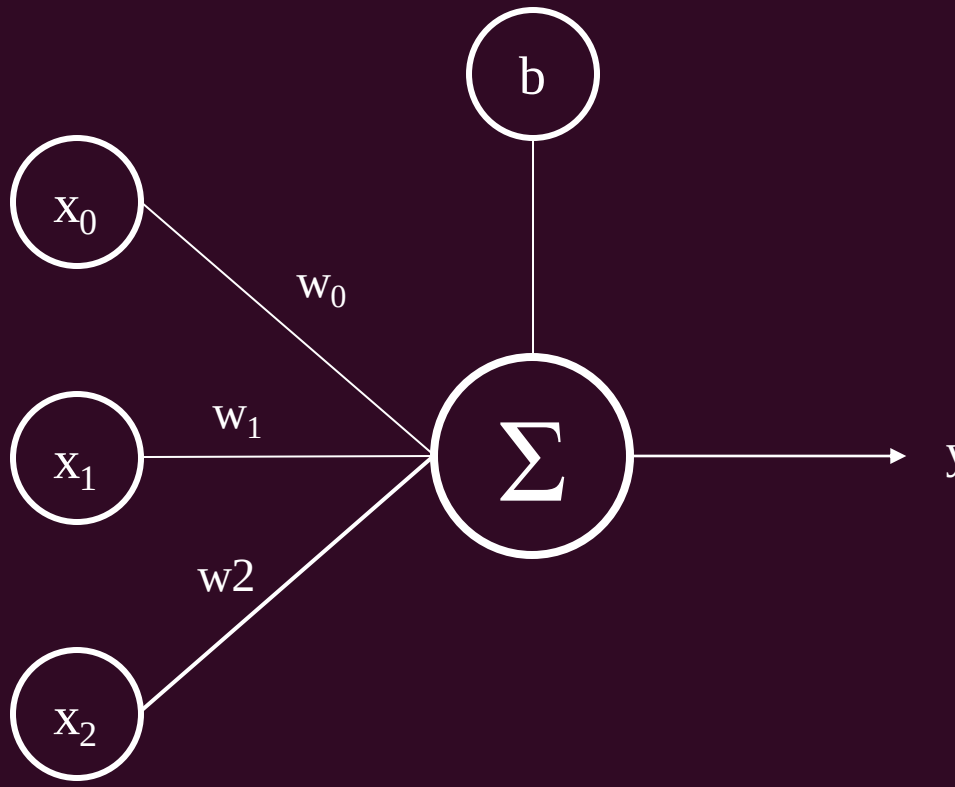
- 인공지능(Artificial Intelligence)의 방법 중 하나
- 지도 학습 – 비지도 학습 – 강화 학습
- 미분과 기울기의 역할이 큼



# 인공 신경망(Artificial Neural Network)

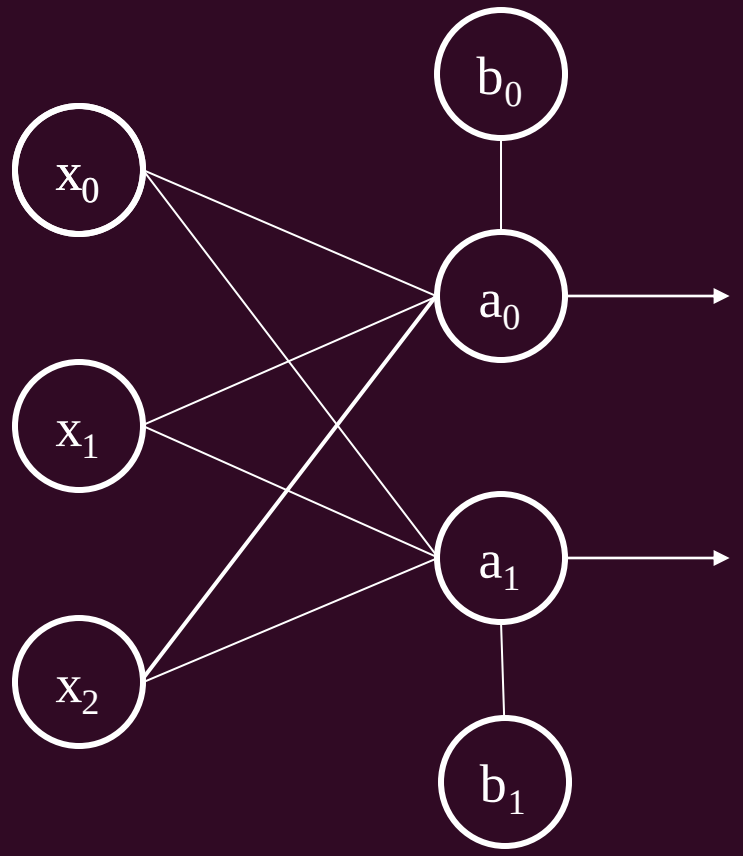
- 기계 학습의 모델 중 하나
- 입력값과 출력값이 있고, 하나의 뉴런(퍼셉트론) 안에서 가중치와 함께 정해진 계산을 진행하여 출력값을 내보내는 방식
- 입력값과 출력값을 계산하려면 행렬 모델링이 필요하다.

# Perceptron




$$y \Rightarrow \begin{pmatrix} x_0 & x_1 & x_2 \end{pmatrix} \cdot \begin{pmatrix} w_0 \\ w_1 \\ w_2 \end{pmatrix} + \begin{pmatrix} b \end{pmatrix} = \begin{pmatrix} \Sigma \end{pmatrix}$$

# Perceptrons




$$\begin{pmatrix} x_0 & x_1 & x_2 \end{pmatrix} \cdot \begin{pmatrix} w_{00} & w_{01} \\ w_{10} & w_{01} \\ w_{20} & w_{01} \end{pmatrix} + \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix}$$

# Deep Learning