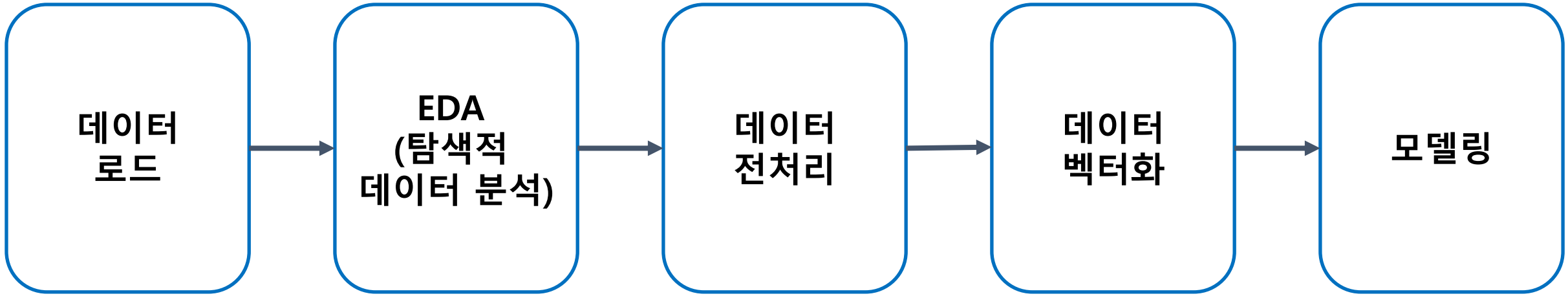
감성분석(Sentiment Analysis)

박경규

감성 분석 절차



데이터셋

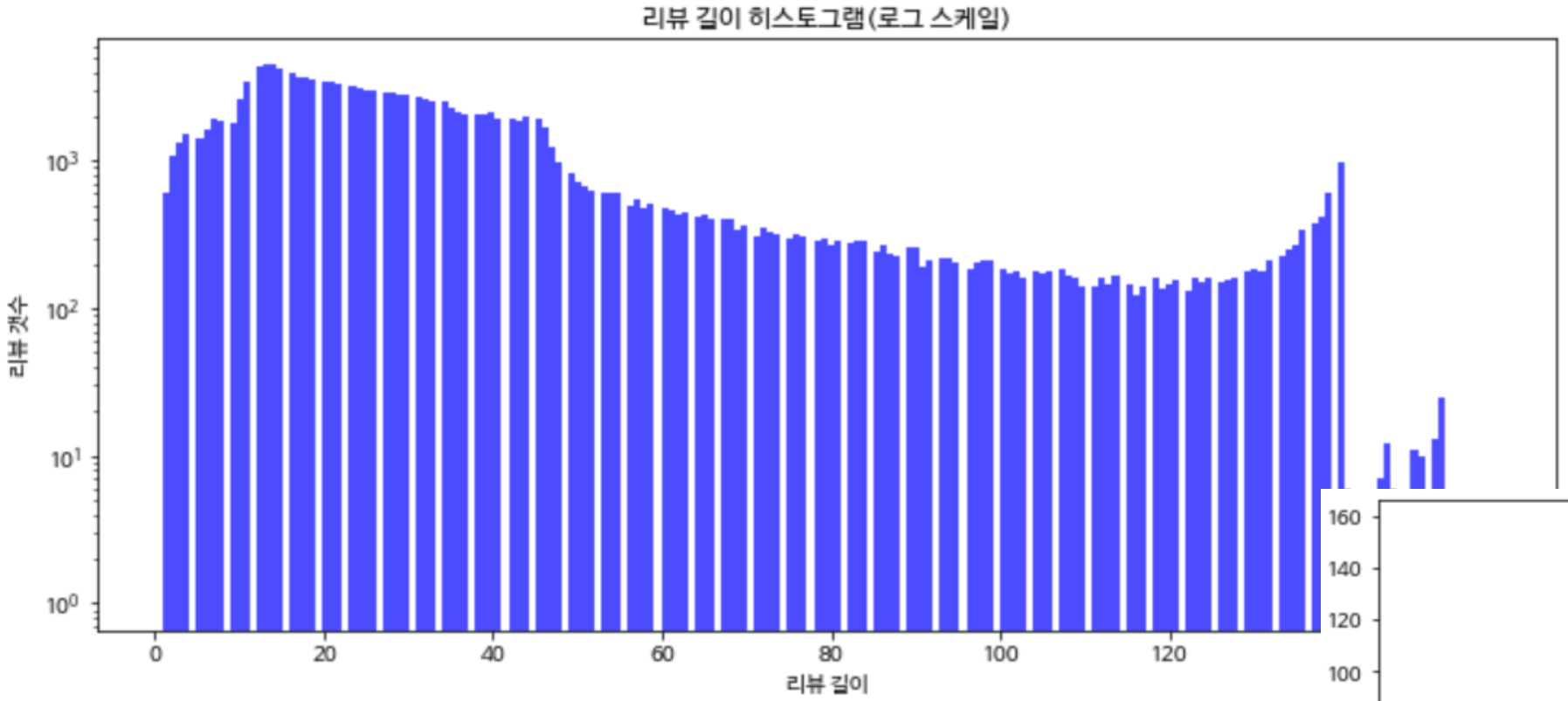
■ https://github.com/e9t/nsmc/

- id: The review id, provieded by Naver
- document: The actual review
- label: The sentiment class of the review. (0: negative, 1: positive)

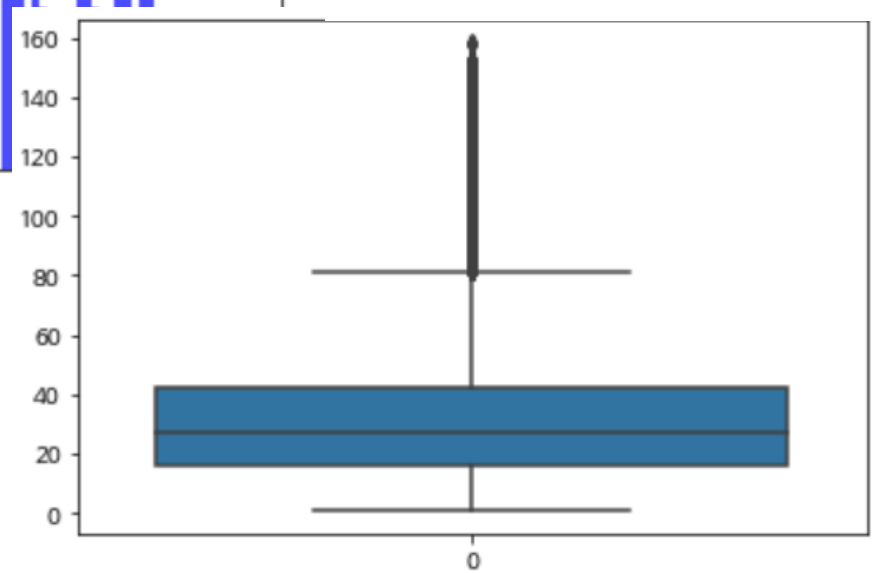
	id	document	label
0	9976970	아 더빙.. 진짜 짜증나네요 목소리	0
1	3819312	흠...포스터보고 초딩영화줄....오버연기조차 가볍지 않구나	1
2	10265843	너무재밋었다그래서보는것을추천한다	0
3	9045019	교도소 이야기구먼 ..솔직히 재미는 없다..평점 조정	0
4	6483659	사이몬페그의 익살스런 연기가 돋보였던 영화!스파이더맨에서 늙어보이기만 했던 커스틴 ...	1

데이터 분석

■ 리뷰의 문자 길이 분포 분석



Train 데이터 개수: 150,000
리뷰 길이 최대 값: 158
리뷰 길이 최소 값: 1
리뷰 길이 평균 값: 35.24
리뷰 길이 표준편차: 29.58
리뷰 길이 중간값: 27.0
리뷰 길이 제 1 사분위: 16.0
리뷰 길이 제 3 사분위: 42.0

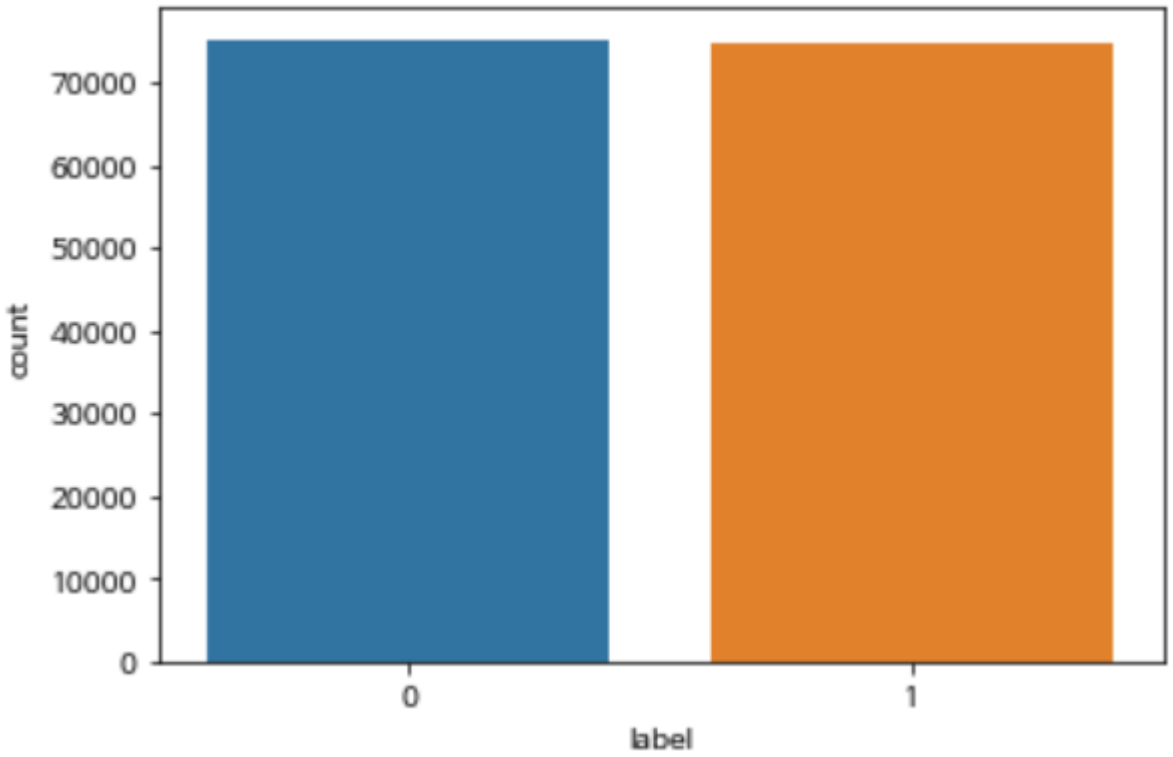


데이터 분석

■ 많이 사용된 단어 – Word Cloud

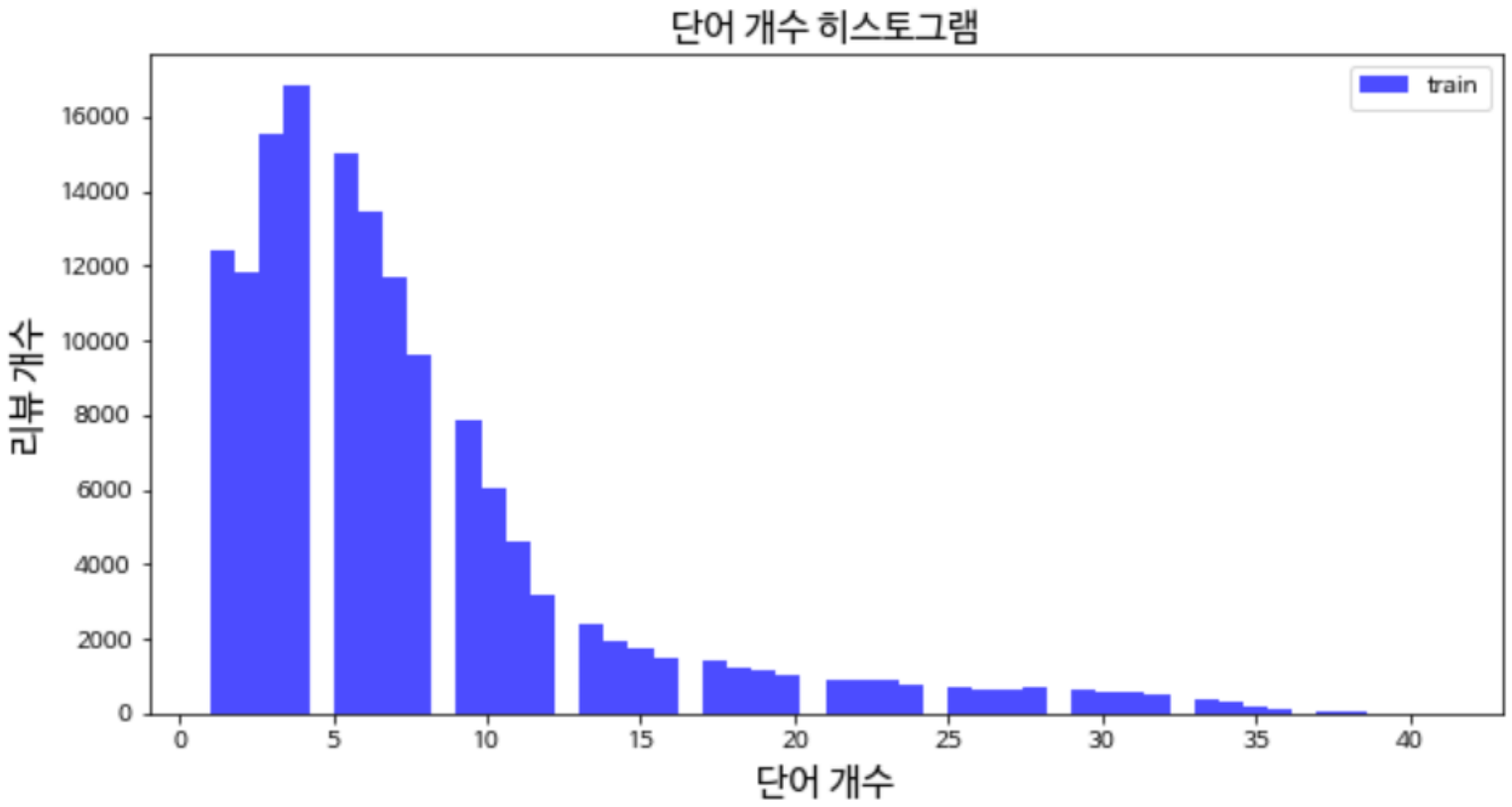
데이터 분석

■ 긍정, 부정 데이터(lable) 분포 확인



데이터 분석

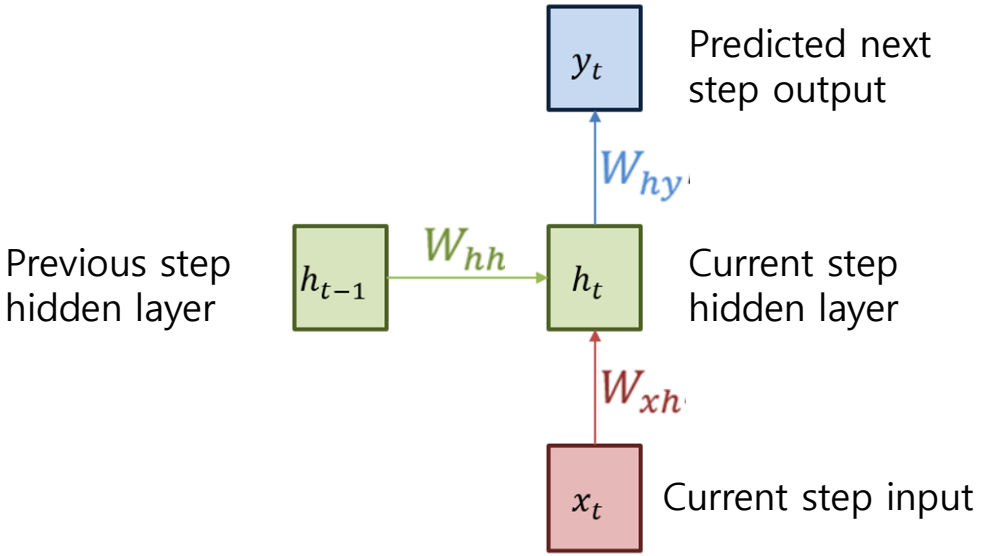
■ 리뷰의 문자 길이 분포 분석



리뷰 단어 개수 최대 값: 41
리뷰 단어 개수 최소 값: 1
리뷰 단어 개수 평균 값: 7.58
리뷰 단어 개수 표준편차: 6.51
리뷰 단어 개수 중간 값: 6.0
리뷰 단어 개수 제 1 사분위: 3.0
리뷰 단어 개수 제 3 사분위: 9.0

RNN 모델

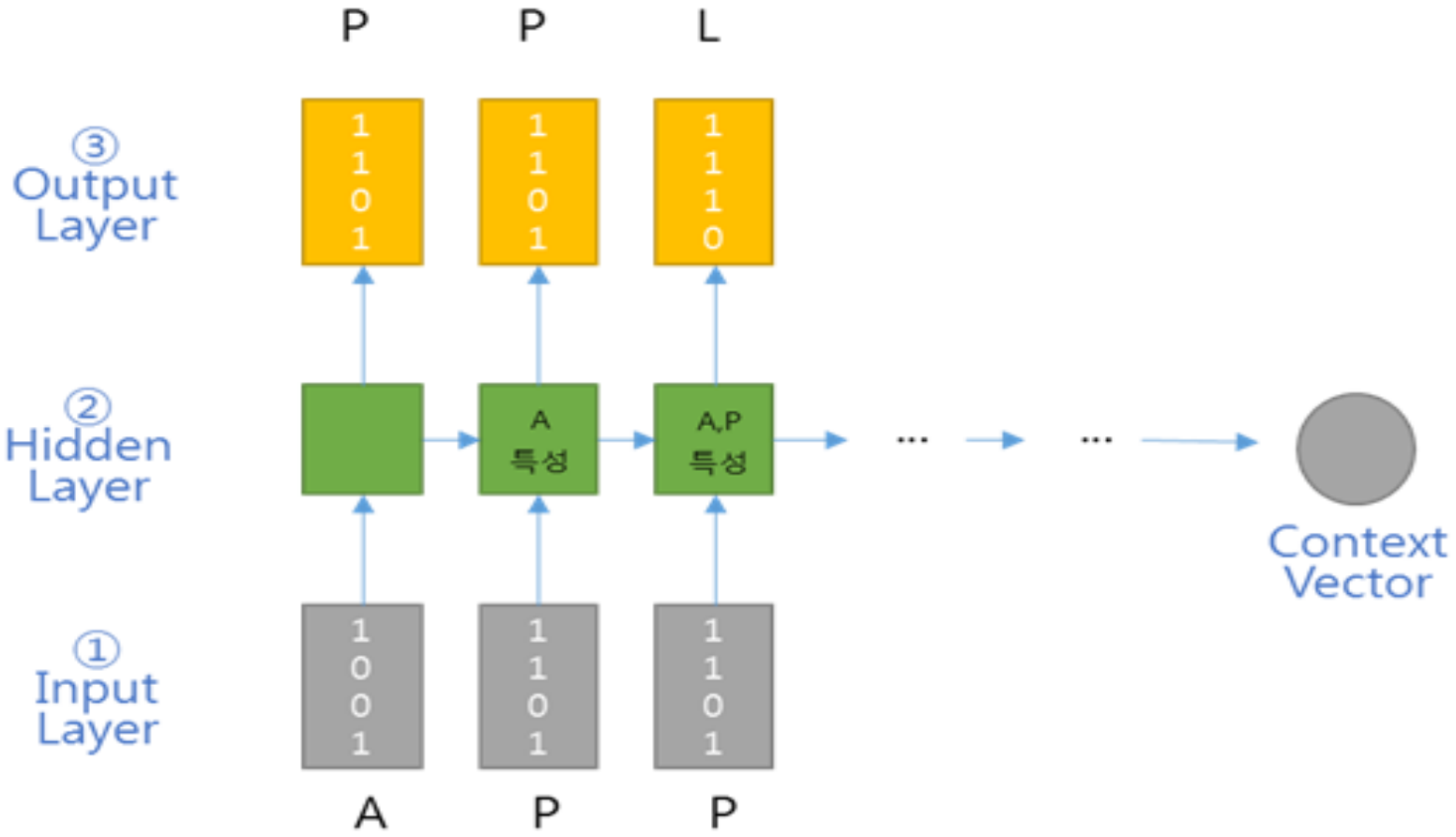
- RNN은 순환구조 인공신경망으로 이전 state 정보가 다음 state 를 예측하는데 사용되어 순차데이터 처리에 특화



$$y_t = W_{hy}h_t + b_y$$

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t + b_h)$$

RNN 모델



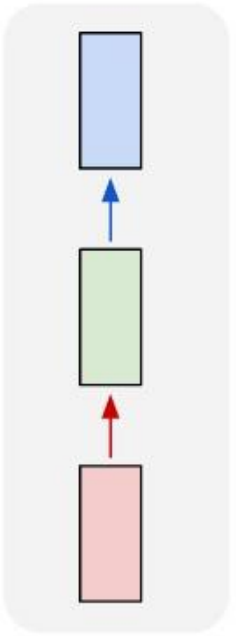
① Input Layer : 현재의 값
② Hidden Layer : 다음을 예측하기 위해 이전 값의 특성을 담는 곳
③ Output Layer : 현재를 기반으로 예측된 다음의 값

RNN 모델

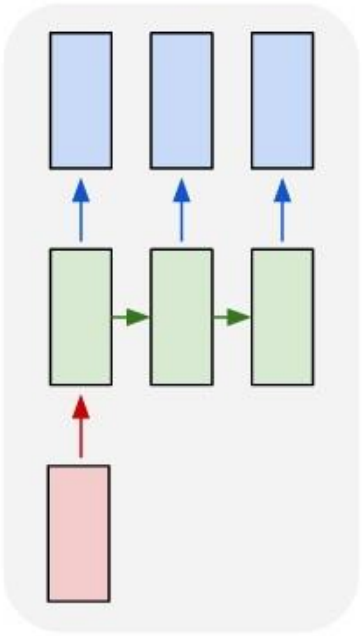
예. **감성 분류**
일련의 단어들 -> 감성

예. **기계 번역**
일련의 단어들 -> 일련의 단어들

one to one

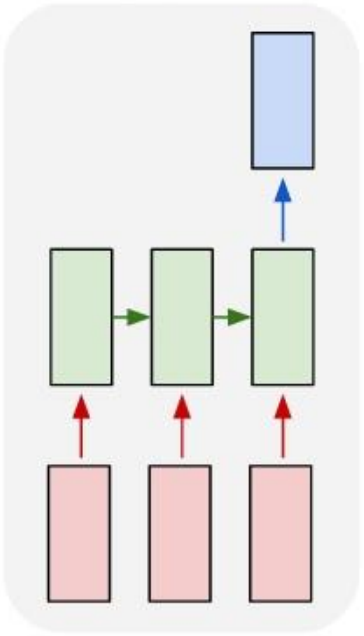


one to many



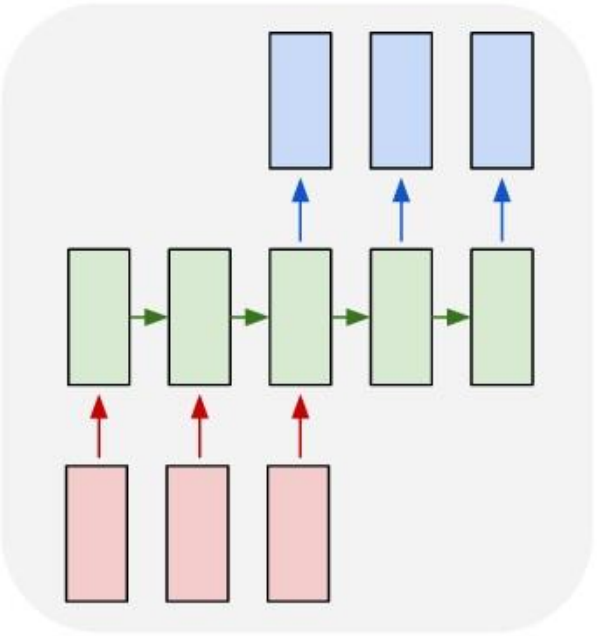
예) 이미지 Captioning
이미지 → 일련의 단어들

many to one



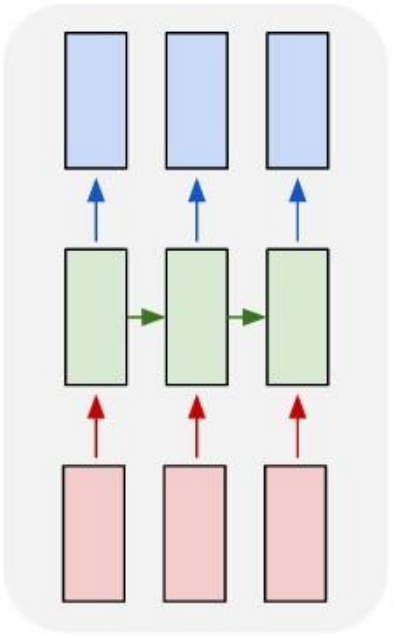
예) 감성 분류
일련의 단어들 → 감성

many to many



예) 기계 번역
일련의 단어들 → 일련의 단어들

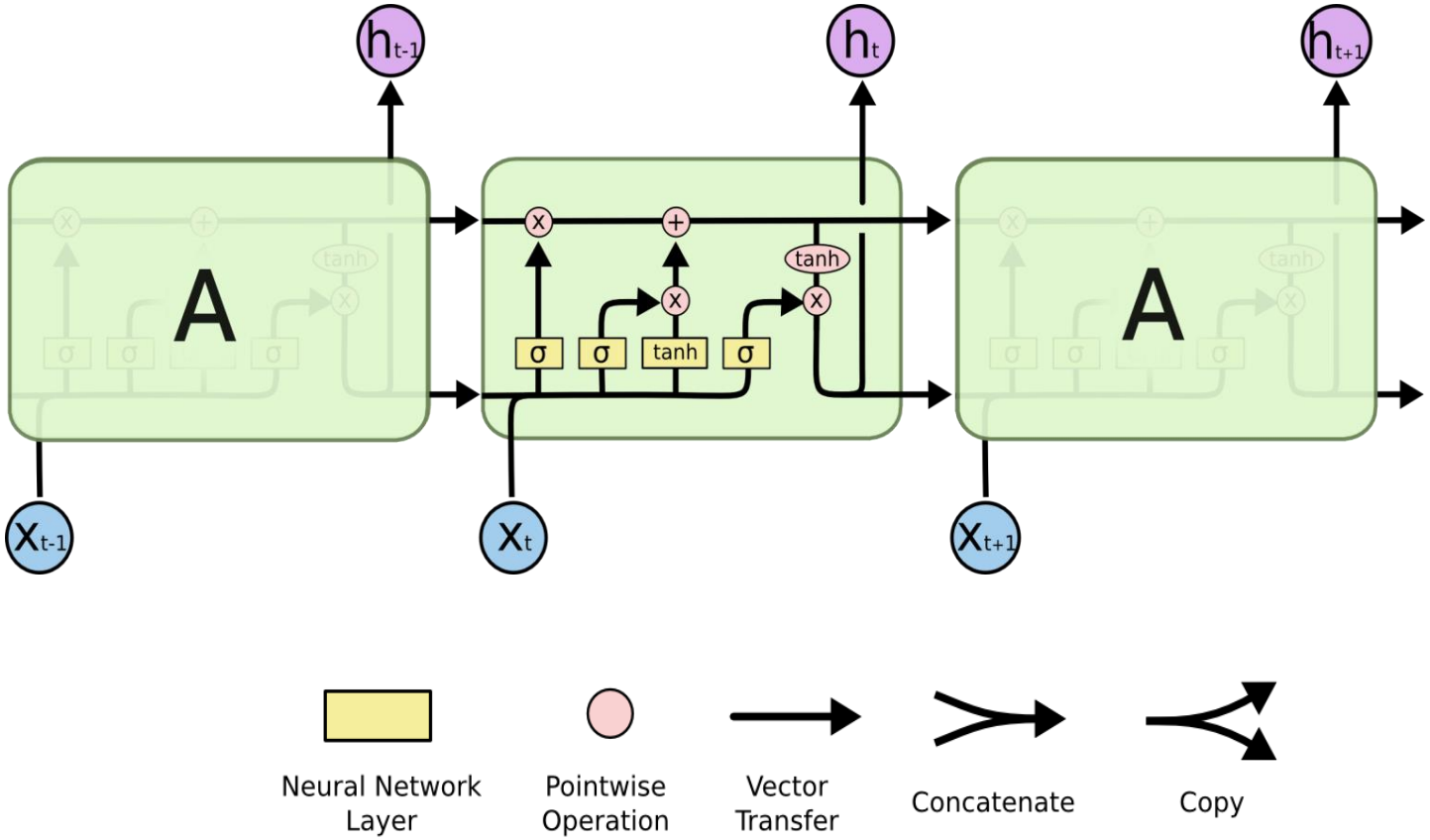
many to many



예) 프레임 수준에서 비디오 분류

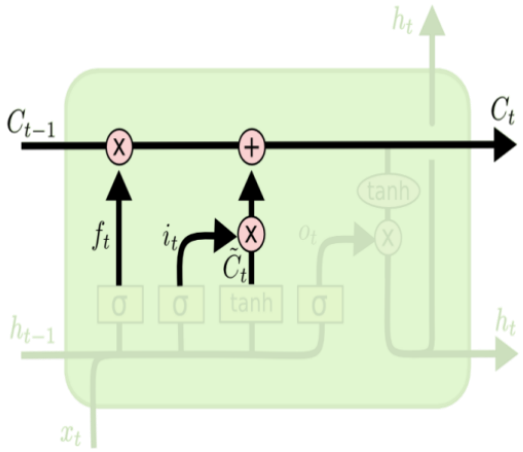
LSTM 모델

- LSTM(Long-Short Term Memory) 네트워크는 장기적인 종속성을 학습할 수 있는 특수한 종류의 RNN입니다.
- LSTM은 RNN과 동일하게 입력과 출력사이 신경망이 재귀하는 구조를 갖고 있습니다.
- RNN은 재귀를 통한 정보전이 및 전파가 하나의 레이어로 제어되는 반면,
  LSTM은 Forget gate, Input gate, Output gate를 통한 정보전이 및 전파를 제어합니다.



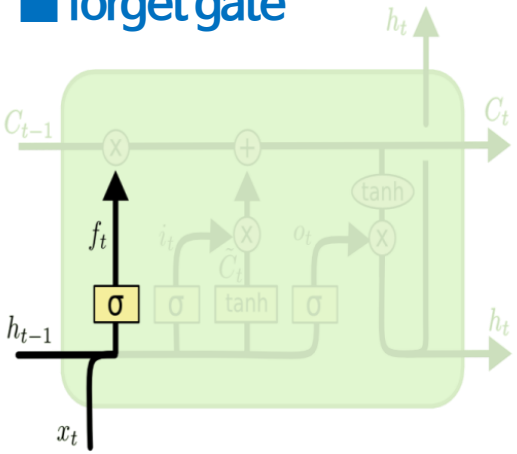
LSTM 모델

■ Cell State(장기 상태)



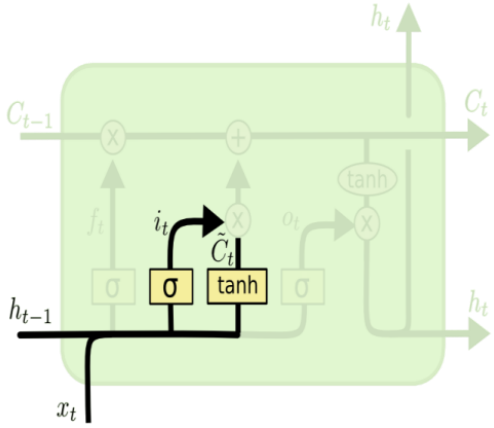
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

■ forget gate



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

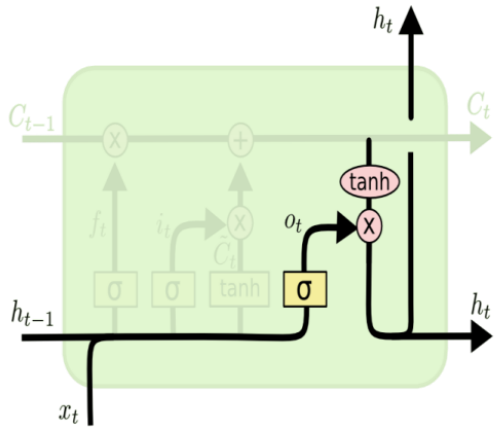
■ input gate



$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

■ output gate, hidden state(단기 상태)



$$o_t = \sigma\left(W_o \left[h_{t-1}, x_t\right] \; + \; b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

NSMC 감성 분석 모델링 실습

nsmc_lstm.ipynb