딥러닝을 이용한 자연어 처리

https://www.edwith.org/deepnlp/lecture/29206/
https://www.edwith.org/deepnlp/lecture/29207/

텍스트 분류(Text classification)

- 문장, 문단 또는 글을 어떤 카테고리에 있는지 분류하는 작업
- 텍스트 분류는 지도학습
- output : 문장이나 글이 어떤 카테고리에 속하는지 결과

예시) 감정분석, 카테고리 분류, 의도 분류

How to represent sentence & token

-문장은 토큰으로 구성되어 있다. 텍스트 토큰은 주관적, 임의적(arbitrary)인 성격을 가지고 있다.
-토큰은 공백, 형태소(Morphs), 어절, 비트숫자로 나눌 수 있다.
-토큰은 integer "index"인데 토큰은 arbitrary한 성격을 가지고 있기 때문에 encoding 할 때 반영해야 한다. (one – hot coding 사용)
-one-hot coding은 모든 토큰 간에 거리가 같아지게 한다. 하지만 모든 단어의 뜻이 같지 않기에 거리가 달라져야한다.
-이를 해결하기 위해 토큰마다 연속 벡터 공간(W, Continuous vector space를 준다.)
-Table Look UP : 각 one hot encoding된 토큰에게 벡터를 부여하는 과정, 실질적으로 one hot encoding 벡터(x)와 연속 벡터 공간(W)을 내적 한 것이다. 토큰에 대한 의미를 찾는 과정
-CBoW
-문장에 대한 의미는 어떤 형식으로 찾을 것 인가?