NLP 스터디 1주차-1

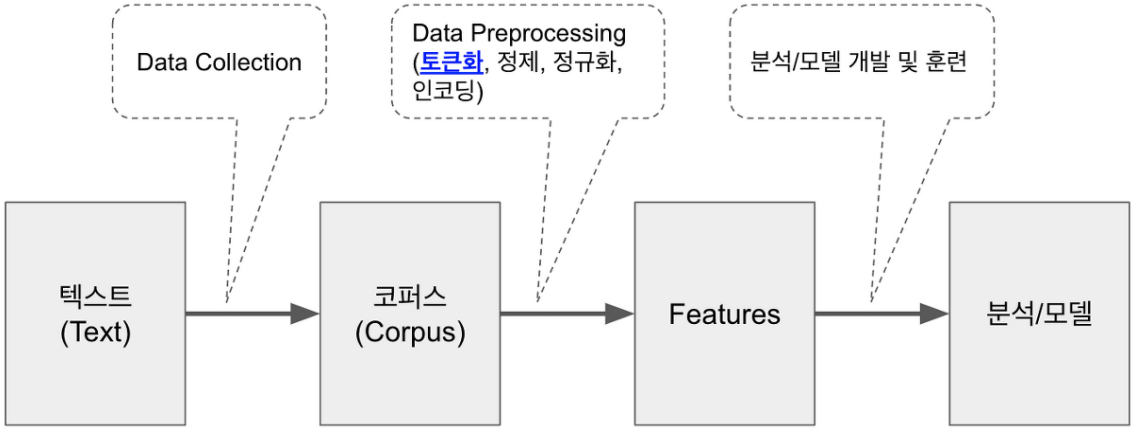
토큰화

> 문장을 분석하기 위해 더 이상 쪼개 지지 않을 때까지 않을때 까지 잘게 분해하는 작업(형태소 단위로 자르는 것)

I am a boy. → I, am, a, boy, .

토큰화는 왜 할까?

토큰화는 데이터 전처리의 한 단계이다. 토큰화, 정제, 불용어 처리, 인코딩 등 여러 단계를 거쳐 실제 모델의 입력데이터로 사용되는 피쳐가 만들어지게 된다.



자연어 처리 과정 중 토큰화의 위치

<이유>

1. 많은 모델에 토큰화 된 데이터를 필요로 하기 때문
2. 단어사전을 만들기 위해 (토큰의 출현 횟수를 이용해서 다양한 분석 방법을 이용) Bag of Words,나 TF-IDF가 그 예시

토큰화 종류

단어 토큰화

특정 구분기호를 가지고 텍스트를 나누는 방법

영어의 경우 기본적으로 공백을 구분자로 사용(Word2Vec, GloVe)

한글의 경우 교착어라는 특징이 있어 구분자나 공백으로 토큰화 하면 성ㅇ능이 좋지 않다. 한국어는 영어와 달리 조사가 큰 의미를 가지고 있기 때문

단점

1. OOV (Out of Vocabulary)의 약자 즉, 입력된 데이터가 이미 만들어져 있던 단어사전에 없는 경우

→ 훈련데이터에서 많이 출현하지 않은 단어/토큰을 새로운 사용자 사전에 등록하는것

2. 사전에 포함된 단어가 많을 수록 사전에서 단어를 찾고 표현을 찾는데 시간이 많이 걸린다. 모델의 응답시간에 문제가 생길 수도 있고, 고용량의 메모리 서버 컴퓨팅 자원이 필요하기도 한다.

문자 토큰화

문자 토큰화는 영어 같은 경우 알파벳 26개로 하나씩 분리하는 것이고, 한글 같은 경우 자음과 모음으로 분리하는것

ex) ‘Tokenization’ 을 T-o-k-e-n-i-z-a-t-i-o-n’으로 ‘토큰화’를 ㅌ-ㅗ-ㅋ-ㅡ-ㄴ-ㅎ-ㅗ-ㅏ로 분리

메모리 문제와 OOV 문제도 해결

그러나 한건의 입력 내용이 길어진다는것. 하나의 단어 인데도 이렇게 표현이 길어지는데 문장이면 더 길어 질 수 잇다.