NLP 스터디 1주차

1. 자연어 처리란?

자연어 : 사람들이 일상적으로 쓰는 언어를 인공적으로 만들어진 언어인 인공어와 구분하여 부르는 개념

자연어 처리 : 일상생활에서 사용하는 언어의 의미를 분석하여 컴퓨터가 처리할 수 있도록 하는 과정

자연어 처리 특징

- 딥러닝에 대한 이해 + 인간 언어에 대한 이해 필요
- 자연어 처리를 위해 사용되는 용어가 익숙하지 않음
- 언어 종류가 다르고 그 형태가 다양함
    - 예) 영어 : 띄어쓰기 有, 중국어 : 띄어쓰기 無 → 단어 단위의 임베딩이 어려움

1.1 자연어 전처리 관련 용어

① 말뭉치 (corpus)

자연어 처리에서 모델을 학습시키기 위한 데이터

② 토큰 (token)

자연어 처리시 텍스트를 작은 단위로 나누어야하는데 이때 텍스트를 나누는 단위를 의미

③ 토큰화 (tokenization)

텍스트를 문장이나 단어로 분리하는 것

토큰화 단계를 마치면 텍스트 ➡️ 단어 단위로 분리됨

```
"Is it possible distinguishing cats and dogs"
➞ ['Is', 'it', 'possible', 'distinguishing', 'cats', 'and', 'dogs']
```

문장 토큰화

④ 불용어 (stop words)

문장 내에서 많이 등장하는 단어

자주 등장해서 의미가 없는 단어는 자연어 처리의 효율성 감소, 처리 시간을 길어지게 함

➡️ 사전에 제거 必

불용어 예) a, the, she, he 등

⑤ 어간 추출 (stemming)

단어를 기본 형태로 만드는 작업

예) consign, consigned, consigning, consignment ➡️ consign

⑥ 품사 태깅 (part-of-speech tagging)

주어진 문장에서 품사를 식별하기 위해 붙여 주는 태그

예) A ➡️ Det, cat ➡️ Noun, is ➡️ Verb

⑦ 정규화 (normalization)

표현 방법이 다른 단어들을 통합시켜 같은 단어로 만들어줌

예) USA와 US ➡️ 같은 의미로 해석되도록 만들어줌