i) 어간 추출과 원형 복원 (Stemming and lemmatizing)

- 어간추출 (Stemming) : 여러가지 어미로 변화된 단어의 접사나 어미를 제거하여 같은 의미를 가지는 형태소의 실제 형태를 동일하게 만드는 방법.

ex) words = ['lives', 'dies', 'flies', 'died']

⇓ Stemmer

['live', 'die', 'fli', 'die']

- 원형복원 (Lemmatizing) : 같은 의미를 가지는 여러 단어를 사전형으로 동일하는 작업.

ex) words = ['lives', 'dies', 'flies', 'died']

⇓ Lemmatizing

['life', 'dy', 'fly', 'died']

> - 어간추출 : 같은 의미를 가지는 형태소의 실제 형태를 동일하게 만드는 방법
> - 원형복원 : 같은 의미를 가지는 여러 단어를 사전형으로 동일하는 작업.

ii) pos tagging (품사 부착)

- 품사(pos)는 낱말을 문법적 기능이나, 형태, 등에 따라 구분한 것.
  - 명사, 동사, 인칭대명사, 서수, 관형사, ...

정리
① 말뭉치 : Sample 문서 집합.
② 토큰 : 문자열의 단위
③ 토큰생성 : 문자열을 토큰으로 나누는 작업.
④ tokenizer : 문자열을 토큰으로 분리하는 함수.
⑤ 형태소 : 일정한 의미가 있는 가장 작은 말의 단위 (토큰을 형태소로 이용)
⑥ 형태소 분석 : 단어 뒤의 다양한 언어적 속성을 파악하고 이를 이용해 형태소를 찾아내는 작업
  - i) stemming : 같은 의미를 가지는 형태소의 실제 형태를 동일하게 만드는 방법.
  - ii) Lemmatizing : 같은 의미를 가지는 여러 단어를 사전형으로 동일하는 작업.
  - iii) pos tagging : 품사 (명사, 동사, 서수)와 같이 구분한 것.