NLTK (Natural Language Toolkit)

2018/07/18

- 주요기능 ① 말뭉치 (Corpus)
② 토큰 생성
③ 형태소 분석
④ 품사 태깅

① 말뭉치 (Corpus) - 자연어 분석 작업을 위해 만든 Sample 문서 집합.
↳ 단순히 문서들 모아 놓은 것도 있지만, "품사", "형태소" 등의 부가적 의미를 추가하고 자연어 분석을 위해 구조적인 형태로 정리해 놓은 것도 포함.

② 토큰생성 - 자연어 문서를 분석하기 위해서는 긴 문자열을 분석하기 위한 작은 단위로 나누어야 함.
이 문자열의 단위를 "토큰", 문자열을 토큰으로 나누는 작업을 "토큰생성"이라고 함.

- 말뭉치 => 문서 집합.
- 토큰 => 문자열의 단위
- 토큰생성 => 문자열을 토큰으로 나누는 작업.
- Tokenizer => 문자열을 토큰으로 분리하는 함수 (토큰 생성 함수)

③ 형태소 분석 (morpheme Analysis)

- 형태소(morpheme)란? - 일정한 의미가 있는 가장 작은 말의 단위
↳ 보통, 자연어 처리에서는 "토큰"을 "형태소"로 이용한다.

- 따라서, 형태소 분석이란? - 단어로 부터 명사, 동사, 조사, 형용사 등 다양한 언어적 속성을 파악하고 이를 이용하여 형태소를 찾아내거나 처리하는 작업.

↳ 형태소 분석의 예로는 다음과 같은 작업이 있다.
① 어간추출 (Stemming)
② 원형복원 (Lemmatizing)
③ 품사부착 (Part-of-Speech tagging)