9. 자연어 처리

9.1 NLTK(Natural Language Toolkit)

- 교육용으로 개발된 자연어 처리와 문서 분석용 파이썬 패키지

- NLTK 패키지 주요기능

 1) 말뭉치(corpus)

 2) 토큰 생성(tokenizing)

 3) 형태소 분석(morphological analysis)

 4) 품사 태깅(POS tagging)

9.2 한글 형태소 분석

용어	상세
형태소(Morpheme)	의미를 가진 최소 단위
용언	꾸미는 말(동사, 형용사) 용언은 어근 + 어미로 이루어짐
어근(Stem)	용언이 활용될 때, 원칙적으로 모양이 변하지 않는 부분
어미	용언이 활용될 때, 변하는 부분으로 문법적 기능을 수행 어미에는 연결 어미, 선어말 어미, 종결 어미가 있음
자모	문자 체계의 한 요소(자음, 모음)
품사	명사, 대명사, 수사, 동사, 형용사, 관형사, 부사, 감탄사, 조사가 있음
어절 분류	명사+주격 조사, 명사+목적격 조사, 명사+관형격 조사, 동사+연결 어미 또는 동사+선어말 어미+종결 어미 등으로 분류
불용어(Stopword)	검색 등에서 의미가 없어 무시되도록 설정된 단어
n-gram	문자의 빈도와 문자간 관계를 의미. "안녕하세요"를 2-gram으로 나누면, "안녕", "녕하", "하 세", "세요"로 나눌 수 있음

©2021.03.26 KwangyumPark all rights reserved.