01. NLTK 자연어 처리 패키지 #src: 9_자연어처리/ch01_NLTK자연어처리패키지 참고

1) 텍스트 마이닝(Text Mining); 데이터 마이닝의 한 분야, 자연어에서 의미있는 정보를 찾는 것

- 비정형 문서 데이터로부터 문서별 단어의 행렬을 만든 후, 여러가지 분석기법과 데이터 마이닝 기법을 사용하여 Insight를 얻거나 의사결정을 지원하기 위해 사용
- 비정형 데이터(문서) -> 말뭉치(Corpus) -> 구조화된 문서 -> 분석(분류, 군집, 연관, 감성 등)

2) NLTK(Natural Language Toolkit); 말뭉치, 토큰 생성, 형태소 분석, 품사 태깅 등 기능 제공

(1) 토큰 생성

① sent_tokenize(); 문자열을 입력받아 문장 단위의 토큰 리스트 출력

② word_tokenize(); 문자열을 입력받아 단어 단위의 토큰 리스트 출력

③ RegexpTokenizer 클래스 이용시 정규표현식을 이용한 토큰화 가능

(2) 형태소(의미가 있는 가장 작은 말의 단위) **분석** ┌#품사 태깅(POS Tagging)

① 어간 추출(Stemming), ② 원형 복원(Lemmatizing), ③ 품사 부착(Part-Of-Speech Tagging)

(3) 기타 클래스: Text 클래스, FreqDist 클래스 등 #사용 단어 빈도 확인 용이