NLP(National Language Processing)
: 언어를 해석하기 위한 기계번역, 자동으로 질문을 해석하고 답을 해주는 질의응답 시스템 등
텍스트 분석
: 머신러닝, 언어 이해, 통계 등을 활용해 모델을 수립하고 정보를 추출해 비즈니스 인텔레전스나 예측 분석 등의 분석 작업을 주로 수행
: 텍스트 분류, 감성 분석, 텍스트 요약, 텍스트 군집화

01 텍스트 분석
- 비정형 데이터인 텍스트 분석
- 비정형 텍스트 데이터를 어떻게 피처 형태로 추출하고 추출된 피처에 의미 있는 값을 부여하는가
- 피처 벡터화, 피처 추출 : 텍스트를 word 기반의 다수의 피처로 추출, 이 피처에 단어 빈도수와 같은 숫자값을 부여하여 텍스트를 단어의 조합인 벡터값으로 표현 ex. BOW, Word2Vec
- 텍스트 분석 수행 프로세스 : 텍스트 사전 준비작업(텍스트 전처리) -> 피처 벡터화/추출 -> ML 모델 수립 및 학습/예측/평가

02 텍스트 사전 준비 작업(텍스트 전처리) - 텍스트 정규화
- 클렌징(Cleansing) : 텍스트 분석에 방해가 되는 불필요한 문자, 기호 등 사전에 제거
- 토큰화(Tokenization) : 문서에서 문장을 분리하는 문장 토큰화, 문장에서 단어를 토큰으로 분리하는 단어 토큰화
    - 문장 토큰화 : 문장의 마침표, 개행 문자 등 문장의 마지막을 뜻하는 기호에 따라 분리
    - 단어 토큰화 : 공백, 콤마, 마침표 등으로 단어 분리
- 필터링/스톱 워드 제거/철자 수정 : 스톱 워드(Stop Word) 분석에 큰 의미가 없는 단어를 지칭, 의미 없는 단어 제거
- Stemming : 원형 단어로 변환 시 일반적인 방법을 적용하거나 더 단순화된 방법을 적용해 원래 단어에서 일부 철자가 훼손된 어근 단어를 추출하는 경향
- Lemmatization : 품사와 같은 문법적인 요소와 더 의미적인 부분을 감안해 정확한 철자로 된 어근 단어를 찾아줌, 변환에 더 오랜 시간

03 BOW(Bag of Words)
- 문서가 가지는 모든 단어를 문맥이나 순서를 무시하고 일괄적으로 단어에 대해 빈도 값을 부여해 피처 값을 추출하는 모델
- 단점 : 문맥 의미 반영 부족, 희소 행렬 문제
    - 희소 행렬 COO 형식 : 0이 아닌 데이터만 별도의 데이터 배열에 저장하고 그 데이터가 가리키는 행과 열의 위치를 별도의 배열로 저장하는 방식
    - 희소 행렬 CSR 형식 : COO 형식이 행과 열의 위치를 나타내기 위해서 반복적인 위치 데이터를 사용해야 하는 문제점 해결한 방식

04 텍스트 분류 실습

05 감성 분석
- 문서의 주관적인 감성/의견/감정/기분 등을 파악하기 위한 방법
- 문서 내 텍스트가 나타내는 여러가지 주관적인 단어와 문맥을 기반으로 감성 수치를 계산하는 방식 이용

06 토픽 모델링

- 문서 집합에 숨어 있는 주제를 찾아내는 것