- NLP와 텍스트 분석 차이
    - NLP: 머신이 인간의 언어를 이해하고 해석하는 데에 중점. 텍스트 분석을 향상하게 하는 기반 기술.
    - 텍스트 분석(텍스트 마이닝): 비정형 텍스트에서 의미 있는 정보를 추출하는 것에 중점. 머신러닝, 언어 이해, 통계 등을 활용해 모델을 수립하고 정보를 추출해 비즈니스 인텔리전스나 예측 분석 등의 분석 작업을 주로 수행
- 텍스트 분석 기술 영역
    - 텍스트 분류
    - 감성분석
    - 텍스트요약
    - 텍스트 군집화

1. 텍스트 분석 이해

- 텍스트 분석: 비정형 데이터인 텍스트를 분석하는 것
- 비정형 텍스트 데이터를 어떻게 피처 형태로 추출하고 추출된 피처에 의미 있는 값을 부여하는가 하는 것이 매우 중요한 요소
- 피처 벡터화 or 피처 추출: 텍스트를 word 기반의 다수의 피처로 추출하고 이 피처에 단어 빈도수와 같은 숫자 값을 부여해서 텍스트를 단어의 조합인 벡터값으로 표현

2. 텍스트 사전 준비 작업(텍스트 전처리) – 텍스트 정규화

- 텍스트 정규화: 텍스트를 머신러닝알고리즘이나 NLP 애플리케이션에 입력 데이터로 사용

하기 위해 클렌징, 정제, 토근화, 어근화 등의 다양한 텍스트 데이터의 사전 작업을 수행하는 것

- 클렌징: 분석에 방해가 되는 불필요한 문자, 기호 등을 사전에 제거
- 텍스트 토큰화: 문서에서 문장을 분리하거나(문장 토큰화) 문장에서 단어를 토큰으로 분리(단어 토큰화)
- 스톱 워드 제거: 분석에 큰 의미가 없는 스톱 워드를 제거 (is, the, a, will 등)
- Stemming 과 Lemmatization: 문법적 또는 의미적으로 변화하는 단어의 원형을 찾는 것

3. Bag of Words – BOW

- Bag of Words: 문서가 가지는 모든 단어를 문맥이나 순서를 무시하고 일괄적으로 단어에 대해 빈도 값을 부여해 피처 값을 추출하는 모델 (문서 내 모든 단어를 한꺼번에 봉투 안에 넣은 뒤에 흔들어서 섞는다는 의미)
- BOW 장점, 단점
  - 장점: 쉽고 빠른 구축
  - 단점: 문맥 의미 반영 부족, 희소 행렬 문제
- BOW 피처 벡터화
  - 카운트 기반의 벡터화: 단어 피처에 값을 부여할 때 각 문서에서 해당 단어가 나타나는 횟수를 부여하는 경우
  - TF-IDF: 개별 문서에서 자주 나타나는 단어에 높은 가중치를 주되, 모든 문서에서 전반적으로 자주 나타나는 단어에 대해서는 페널티를 주는 방식
- 사이킷런의 countervectorizer 클래스는 카운트 기반의 벡터화를 구현한 클래스로, 소문

자 일괄 변환, 토큰화, 스톱 위드 필터링 등의 텍스트 전처리도 함께 수행

- 희소 행렬 문제: 모든 단어를 중복을 제거하고 피처로 만들면 많은 행렬이 생성되는데, 레코드의 각 문서가 가지는 단어의 수는 제한적이기 때문에 이 행렬의 값은 대부분 0이 차지할 수밖에 없다. 그러면 메모리 공간이 많이 필요하고 연산 시에도 데이터 액세스를 위한 시간이 많이 소모된다.
    - COO, CSR 형식으로 희소 행렬을 물리적으로 적은 메모리 공간을 차지할 수 있도록 변환할 수 있다.
- COO: 0이 아닌 데이터만 별도의 데이터 배열에 저장하고 그 데이터가 가리키는 행과 열의 위치를 별도의 배열로 저장하는 방식
- CSR: COO형식이 행과 열의 위치를 나타내기 위해서 반복적인 위치 데이터를 사용해야 하는 문제점을 해결한 방식.

4. 텍스트 분류 실습 – 20 뉴스그룹 분류

5. 감성 분석

- 감성 분석: 문서의 주관적인 감성, 의견, 감정, 기분 등을 파악하기 위한 방법으로 소셜 미디어, 여론조사, 온라인 리뷰, 피드백 등 다양한 분야에서 활용
- 주관적인 단어와 문맥을 기반으로 감성 수치를 계산
- 긍정 감성 지수, 부정 감성 지수

6. 토픽 모델링(Topic Modeling) – 20 뉴스그룹

- 토픽 모델링: 문서 집합에 숨어있는 주제를 찾아내는 것. 숨겨진 주제를 효과적으로 표현

할 수 있는 중심 단어를 함축적으로 추출한다.