NLP vs 텍스트 분석

- NLP
  - 머신이 인간의 언어를 이해하고 해석하는 데 중점을 두고 기술이 발전
  - 기계번역, 질의응답 시스템
- **텍스트 분석**
  - 비정형 텍스트에서 의미 있는 정보를 추출하는 데 중점을 두고 기술이 발전
  - 텍스트 분류, 감정 분석, 텍스트 요약, 텍스트 군집화와 유사도 측정

텍스트 분석 이해

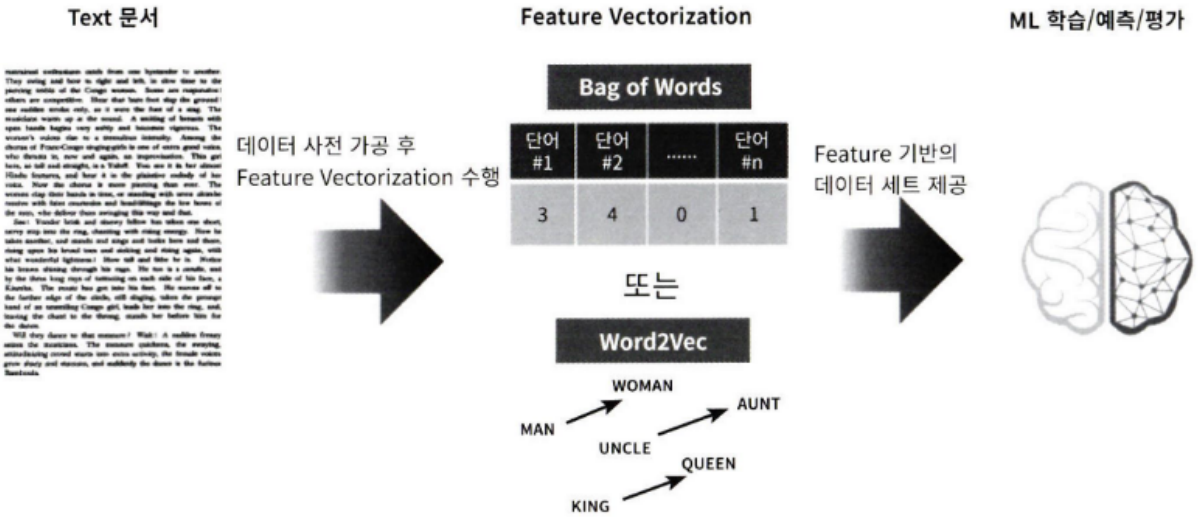
- **텍스트 분석** : *비정형 데이터*인 텍스트를 분석하는 것
- 비정형 텍스트 데이터를 *어떻게 피처 형태로 추출*하고 *추출한 피처에 의미 있는 값을 부여* 하는 것이 중요
- 텍스트를 단어의 조합인 벡터값으로 표현하는 것 : **피처 벡터와, 피처 추출**
  - BOW, Word2Vec

텍스트 분석 수행 프로세스

- 다음과 같은 프로세스 순으로 수행

1. **텍스트 사전 준비작업(텍스트 전처리)** : 피처화 이전 클렌징 작업, 토큰화 작업, 의미 없는 단어 제거 작업, 어근 추출 등의 텍스트 정규화 작업을 통칭
2. **피처 벡터화/추출** : 사전 준비 작업으로 가공된 텍스트에서 피처를 추출하고 벡터값을 할당.

3. ML 모델 수립 및 학습/예측/평가 : 피처 벡터화된 데이터 세트에 ML 모델을 적용해 학습/예측 및 평가를 수행



〈 Text Analytics 수행 프로세스 〉

파이썬 기반의 NLP, 텍스트 분석 패키지

- NLTK : 가장 대표적인 NLP 패키지, 수행 속도 측면에서 아쉬운 부분이 있음
- Gensim : 토픽 모델링 분야에서 가장 두각을 나타내는 패키지. SpaCy와 함께 가장 많이 사용됨.
- SpaCy : 뛰어난 수행 성능으로 최근 가장 주목을 받고 있음.

텍스트 사전 준비 작업(텍스트 전처리) - 텍스트 정규화

- 텍스트 정규화 : 텍스트를 머신러닝 알고리즘이나 NLP 애플리케이션에 입력 데이터로 사용하기 위해 다양한 텍스트 데이터의 사전 작업을 수행하는 것

클렌징

텍스트에서 오히려 분석에 방해가 되는 문자, 기후 등을 사전에 제거하는 작업

텍스트 토큰화

- 문장 토큰화 : 문서에서 문장을 분리
- 단어 토큰화 : 문장에서 단어를 토큰으로 분리
  - n-gram : 연속된 n개의 단어를 하나의 토큰화 단위로 분리해 내는 것

스톱 워드 제거

- 스톱 워드 : 분석에 큰 의미가 없는 단어
  - 텍스트에 빈번하게 나타나므로 중요한 단어로 인지될 수 있음.
  - -> 제거하는 것이 중요

Stemming과 Lemmatization

- 문법적 또는 의미적으로 변화하는 단어의 원형을 찾는 것
- Lemmatization이 Stemming보다 정교하며 의미론적인 기반에서 단어의 원형을 찾음

Bag of Words - BOW

- Bag of Words : 문서가 가지는 모든 단어를 문맥이나 순서를 무시하고 일괄적으로 단어에 대해 *빈도값* 을 부여해 피처 값을 추출하는 모델
- 장점 : 쉽고 빠른 구축
- 단점 :
  - 문맥 의미 반영 부족
  - 희소 행렬 문제 - ML 알고리즘의 수행 시간과 예측 성능을 떨어뜨림

BOW 피처 벡터화

- **피처 벡터화** : 텍스트를 특정 의미를 가지는 숫자형 값인 벡터로 변환하는 것
- **BOW 모델에서의 피처 벡터화** : 모든 문서에서 모든 단어를 칼럼 형태로 나열하고 각 문서에서 해당 단어의 횟수나 정규화된 빈도를 값으로 부여하는 데이터 세트 모델로 변경하는 것
- 두 가지 방식의 피처 벡터화
  - **카운트 기반의 벡터화** : 해당 단어가 나타나는 횟수를 값으로 부여
    - 언어의 특성상 문장에서 자주 사용될 수밖에 없는 단어까지 높은 값으로 부여하는 문제를 가짐
  - TD-IDF : 개별 문서에서 자주 나타나는 단어에 높은 가중치, 전반적으로 자주 나타나는 단어에는 페널티
- 사이킷런 : CounterVectorizer, TfdifVectorizer로 구현

BOW 벡터화를 위한 희소 행렬

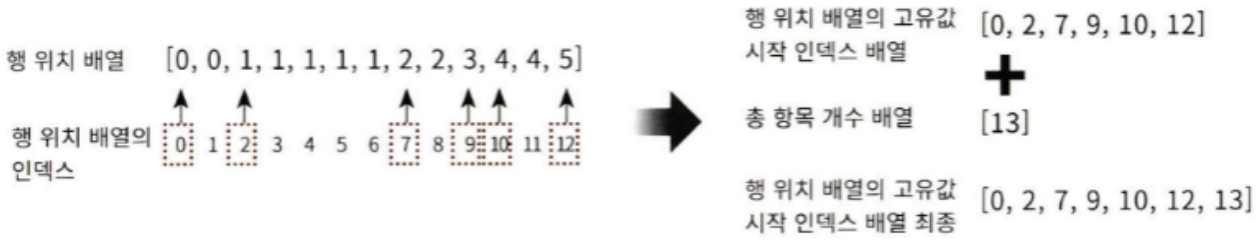
- **희소 행렬** : 대규모 행렬의 대부분의 값을 0이 차지하는 행렬
  - 메모리 공간이 많이 필요
  - 행렬의 크기가 크므로 연산 시 데이터 엑세스를 위한 시간이 많이 소모
- -> 물리적으로 적은 메모리 공간을 차지할 수 있도록 변환하는 방법 :
  - COO 형식
  - CSR 형식

희소 행렬 - COO 형식

- COO 형식 : 0이 아닌 데이터만 별도의 데이터 배열에 저장하고, 그 데이터가 가리키는 행과 열의 위치를 별도의 배열로 저장

희소 행렬 - CSR 형식

- 행 위치 배열 내에 있는 고유한 값의 시작 위치만 다시 별도의 위치 배열로 가지도록 변환



감정 분석

감정 분석 소개

- 감정 분석 : 문서의 주관적인 감성/의견/감정/기분 등을 파악하기 위한 방법
  - 문서 내 텍스트의 감성 수치를 계산하는 방법을 이용
- 지도학습과 비지도학습 방식으로 나눌 수 있음
  - 지도학습 : 학습 데이터와 타깃 레이블 값을 기반으로 감성 분석 학습을 수행한 뒤 이를 기반으로 다른 데이터의 감성 분석을 예측하는 방법
  - 비지도학습 : 'Lexicon'이라는 감성 어휘 사전을 이용

비지도학습 기반 감정 분석 소개

- Lexicon : 감성만을 분석하기 위해 지원하는 감성 어휘 사전
  - 대표격 : NLTK 패키지
- 감성 지수 : 긍정 감성 또는 부정 감성의 정도를 의미하는 수치
- Wordnet
  - 시맨틱 분석을 제공하는 어휘 사전 : 다양한 상황에서 같은 어휘라도 다르게 사용되는 시맨틱 정보를 제공
  - 개별 단어를 Synset이라는 개념을 이용해 표현
- 대표적인 감성 사전의 목록

- SentiWordNet
- VADER
- Pattern