# 텍스트 마이닝과 데이터 마이닝

# Part 07. 데이터 마이닝

정 정 민

programmers

# Chapter 19. 데이터 마이닝 절차

1. 데이터 마이닝 프로세스

# 데이터 마이닝 프로세스

# 데이터 수집 및 통합

- 목적으로 하는 문제를 풀기 위한 다양한 데이터를 수집
  - 소셜 미디어, 고객 거래 기록, 센서 데이터 등
  - 다양한 분석을 위해 서로 상이한 종류의 데이터를 모으는 것도 좋은 접근

- 데이터 통합 과정
  - 만약 같은 종류의 데이터라면, 일관된 형식으로 만드는 과정 필요
    - 크롤링 과정으로 생성된 DOM 구조 제거
    - 이미지 데이터의 경우 크기 조절
    - 등등

- 데이터 품질 관리
  - 데이터 검증 및 정화 (오류, 중복을 수정 및 제거)
  - 완결성 검사 (누락 데이터 서칭 및 핸들링, 제거 혹은 가상의 값으로 대체)
  - 모니터링 (품질을 지속적으로 모니터링, 업데이트로 인한 버전 관리)

# 데이터 전처리

- 머신러닝 강의 혹은 이전 텍스트 마이닝 과정에서 진행한 것과 비슷!
- 데이터를 분석하기 위한 가장 초기 과정이며 중요한 첫 단추

- 노이즈 및 오류 제거
  - 노이즈로 인한 이상치 데이터를 확인 (IQR, 이상치 알고리즘 결과 등)
  - 수집 과정에서의 이상 상태로 인한 오류 데이터 존재 가능
  - 식별된 이상치 혹은 오류 데이터는 제거 혹은 수정

- 데이터 정규화
  - 데이터의 스케일을 일치시키는 과정
  - 서로 다른 데이터 사이의 일치 뿐 아니라
  - 같은 데이터 내에서도 통일성을 위해 정규화를 진행
    - 예를 들어, 너무 긴 문장을 자르기

- 등등, 모델 혹은 분석 방법에 맞는 전처리 과정 진행

# 데이터 마이닝 기법 적용

- 데이터 마이닝은 하나의 데이터에만 타겟팅 한 주제가 아님
  - 숫자 데이터, 텍스트, 이미지, 시계열 데이터 등등
  - 응용 분야로도.. 비지니스, 마케팅, 고객 관리, 공공 분야 등등

- 수집한 데이터에 특화된 데이터 분석 방법론을 적용
- 유의미한 패턴과 관계, 통찰을 도출하는 방법을 사용

- 가장 좋은 접근 방법으로는
- 비슷한 데이터를 분석한 사례를 확인
  - Kaggle 페이지에서 비슷한 데이터를 찾아보면 좋습니다!
  - 그리고 그런 데이터를 바탕으로 분석한 선례를 찾아보는 것도 좋아요.

- 큰 흐름으로 보는 주요 마이닝 기법으로는 아래와 같음
  - 분류(Classification), 클러스터링(Clustering), 예측(Prediction), 잠재적 의미 표면화 (Latent Representation)
  - 등등

# 데이터 마이닝 결과 분석

- 마이닝의 기본 의미에 맞춰
- 넓고 많은 데이터에서 인사이트를 얻고
- 이를 바탕으로 의사 결정과 같은 과정에 사용

- 이 과정에서 주의할 점은
- **모델 평가 과정이 존재**한다면,
  - 모델을 평가하는 평가 수치가 의사 결정에 도움이 되는 평가인지를 판단
  - 평가한 데이터가 의미 있는 데이터인지 확인 필요

- 평가 과정 없이 **사람의 직관과 판단**이 들어가야 한다면,
  - 원본 데이터에 특이성과 같은 편향에서 자유로운지
  - 그 직관에 위험성은 없는지 등이 필요

# E.O.D