텍스트 마이닝과 데이터 마이닝

Part 02. 텍스트 마이닝 개요

정 정 민

programmers

Chapter 03. 텍스트 마이닝이란?

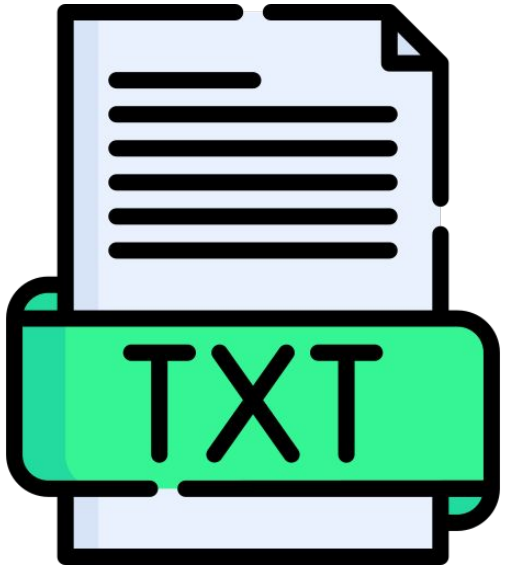
1. 텍스트와 텍스트의 특징
2. 텍스트 마이닝이란

텍스트와 텍스트의 특징

텍스트 데이터 (Text data)

- 텍스트 데이터란,
  - 문자, 단어, 문장으로 구성된 데이터
  - 웹사이트, SNS, 책, 학술 정보, 이메일 등 다양한 출처에서 발생
  - 대규모 데이터
    - 매일 약 8200만 Tb의 텍스트가 생성 (출처)
    - 전체 생성 데이터 중 text 부분만 추출

- 활용
  - 시장 변화를 파악하고 대응할 수 있는 확인 창구
  - 고객의 요구 사항과 피드백을 파악할 수 있음
  - 더 좋은 텍스트 이해를 위한 연구 도구
  - 등등



텍스트 데이터의 특징

- 텍스트 데이터를 구성하는 요소를 기준으로 (단어라고 가정하면)
- 단어는 주변의 단어들과 연관성이 존재

- 이 연관성을 이해하는 방향으로 텍스트 데이터를 처리해야 함

- 텍스트 처리 과정에서 아래의 이유로 어려움이 있음
  - 비구조적(비정형) 데이터
  - 다양성 : 같은 의미라도 여러 표현이 있을 수 있음 (ex. 맛있다, 맛이 좋다 등)
  - 다의성 : 같은 표현이 다른 의미로 사용될 수 있음 (ex. 잘한다 등)
  - 문맥 정보 포함
  - 언어별로 고유한 특징 (문법, 어휘, 발음 등)

텍스트 마이닝이란

마이닝(Mining, 채굴)

- 광물에서 귀중한 금속이나 광석을 채굴하는 작업

- 대량의 광물에서 소량의 귀중한 자원을 발견하고 추출하기 위해서는
- 광물을 면밀히 살펴봐야 함

- 비슷하게, 대량의 데이터(광물)에서
- 유용한 정보와 패턴(자원)을 찾기 위해 채굴이 필요
- 이렇게 채굴된 정보와 패턴으로 통찰력을 얻고, 의사 결정을 진행

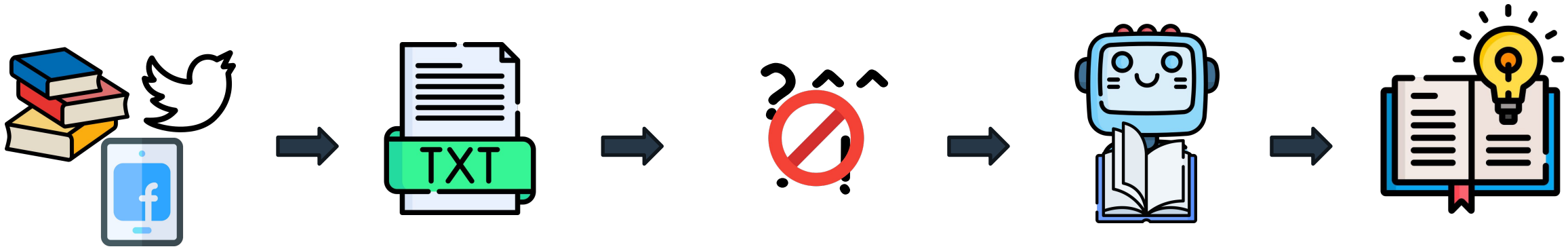
- 정보를 얻고자 하는 원천 데이터로

- 텍스트 데이터가 주어지면 > **텍스트 마이닝**
- 이미지 데이터가 주어지면 > **이미지 마이닝**
- 특정되지 않은 일반적인 데이터라면 > **데이터 마이닝**!

텍스트 마이닝이란

- 텍스트로 구성된 데이터를 바탕으로
- 대용량의 텍스트 안에 존재하는 관계, 패턴, 규칙을 탐색
- 이로부터 지식과 인사이트를 추출해
- 의사결정에 활용하는 일련의 과정을 의미

- 진행 과정은 아래와 같음



텍스트 추출 | 텍스트 정리 (전처리) | 텍스트 마이닝 | 해석 및 평가

자연어 처리 (Natural Language Processing, NLP)

- **자연어 처리**란,
  - 컴퓨터가 인간의 언어를 이해하고 해석하는데 사용되는 분야로
  - 컴퓨터 과학, 인공 지능, 언어학의 개념이 사용됨

- NLP의 목적은 인간 언어의 구조와 의미 이해를 바탕으로
  - 글을 활용한 문제를 해결하고
  - 향상된 사용자 경험을 제공하고자 함
    - chatGPT와 같은 사용 경험이 해당하겠죠?

- 대규모 텍스트 데이터 내의 존재하는 패턴, 관계, 정보를 발견하고 분석하는 텍스트 마이닝과 거리가 있음
- 두 개념의 목표 차이는
  - NLP : 언어의 이해
  - TM : 언어 속 내포된 정보 파악

텍스트 마이닝에 사용되는 패키지

- 텍스트 마이닝을 위해 활용할 수 있는 여러 언어와 패키지 존재
  - 특히 Python, R
  - 그 중 Python을 사용!

- 텍스트 데이터 수집, 처리, 분석, 시각화 등 다양한 작업을 지원

- 아래와 같은 주요 패키지가 존재
  - **Pandas** : 텍스트 데이터 조작과 처리에 용이
  - **Gensim** : 전처리 과정인 임배딩 과정을 지원
  - **nltk** : 자연어를 다루는 과정에서 유용한 tool kit 를 제공
  - 등등

- 실제 활용 사례는 수업에서 살펴보도록 하죠!

E.O.D