텍스트 마이닝과 데이터 마이닝

Part 02. 텍스트 마이닝 개요

정 정 민

programmers

Chapter 04. 텍스트 마이닝 절차

1. 텍스트 마이닝 프로세스

텍스트 마이닝 프로세스

텍스트 수집 및 추출

- 텍스트 마이닝 프로세스의 첫 단계

- 다음과 같은 방법으로 데이터 수집 가능
  - **웹 크롤링**
    - 웹에 존재하는 텍스트 데이터를 수집 (Selenium, Beautiful Soup, Scrapy 등의 패키지 활용)
    - 단, 크롤링 가능한 사이트를 위주로 사용해야 함
  - **API 사용**
    - 대형 SNS 플랫폼, 뉴스 사이트, 온라인 포럼 등은 데이터 제공을 위한 API를 제공
    - API 사용 비용을 지불해야할 수 있음
  - **공개 데이터**
    - 연구 기관, 정부 기관, 기업 등에서 제공하는 공개 데이터를 활용

- 수집한 데이터는 기본적으로 원시의 데이터로 쓸 수 없는 형태의 데이터도 존재
  - 데이터의 질을 관리하기 위해
  - 목표 관련성이 높은 글, 다양성이 확보되는 글을 주기적으로 모니터링 해야 함

텍스트 전처리

- 수집된 데이터는 일반적으로 비구조화 데이터로 특정 분석을 하기에 불완전한 상태일 가능성 ↑
- 전처리 과정을 통해 데이터를 정제하고 분석이 가능한 형태로 변환해야 함

- 전처리는 풀어야 하는 문제에 따라 다양한 방법이 존재
- 일반적으로 아래의 방법들이 사용됨
  - **노이즈 값 제거**
    - 원시 데이터에는 이모티콘, 오타, 비속어 등 다양한 노이즈 값이 존재
    - 이러한 이상 데이터를 제거 혹은 수정
  - **분석에 최소 단위로 글을 분류**
    - 단어 기반 문제 풀이, 문장 기반 문제 풀이 등에 따라 사용하는 정보의 단위가 다름
    - 이러한 정보의 단위로 글을 분리해야 하고 (Tokenize, 추후에 다룹니다!)
    - 이것을 컴퓨터가 이해할 수 있는 형태로 변환해야 함 (Embedding, 추후에 다룹니다!)
  - **글 길이 조절**
    - 제한된 환경에서 작성된 글이 아닌 경우, 너무 길거나 짧은 글이 존재
    - 이를 통일된 형태로 변경해야 함 (길다면 자르거나, 짧다면 복제 혹은 다른 글과 통합 혹은 dummy 값 추가 등)

텍스트 마이닝 기법 적용

- 전처리 이후의 단계로, 데이터로부터 유의미한 정보를 추출하고 인사이트를 도출하는 과정

- 다양한 문제가 존재
  - **내용 파악 및 분석**
    - 자연어 이해 : 글에 존재하는 의미와 의도 파악
    - 요약 : 글의 내용을 요약 정리
    - 개체명 인식 : 글에서 인물, 장소, 기관 등의 특정 정보를 식별 & 분류
  - **숨겨진 의미 파악**
    - 토픽 모델링 : 글에 담겨있는 숨겨진 주제를 발견
    - 트렌드 분석 : 시간에 따른 데이터 변화를 분석, 패턴과 변화를 식별
    - 감정 분석 : 글에 존재하는 저자의 감정 상태를 파악
  - **관계 파악 및 구조화**
    - 군집화 : 비슷한 의미의 글을 그룹화 해서 문서간의 관계 파악
    - 글 분류 : 글을 특정 범주로 분류
  - 등등

텍스트 마이닝 결과 분석

- 분석된 결과를 활용해 정보 이해, 통찰 도출, 의사 결정 과정에서 사용

- **정보 이해**
  - 텍스트의 전반적인 내용을 파악
  - 타겟 그룹에서 생성된 글의 패턴과 흐름을 빠르게 확인
  - Ex) 제품을 사용하는 사용자 중 40대의 반응을 보고 제품 사용 관점에서 나오는 키워드를 파악

- **통찰 도출**
  - 데이터 안에 숨어있는 연결 정보를 추출
  - 숨은 정보를 추출하는 기술적인 모델이 필요
  - Ex) 제품에 대한 긍정적인 포인트와 부정적인 포인트를 이해

- **의사 결정**
  - 통찰을 바탕으로 비지니스 전략, 제품 개발, 마케팅 등의 의사 결정 과정에서 활용
  - Ex) 부정적인 부분을 개발하고, 이를 적극적으로 마케팅에 활용

E.O.D