제3장 텍스트 데이터 불러오기

1. 비정형 데이터의 현황

전체 데이터 중 비정형 데이터 비중이 급증(80% 이상)

2. 텍스트 데이터의 이해

2.1 텍스트 데이터의 사례

UCI Machine Learning Repository(https://archive.ics.uci.edu/ml/datasets.php)

→ 머신러닝 기법의 연구와 발전을 위하여 필요한 다양한 데이터 수록

2.2 텍스트 데이터의 특징

텍스트 데이터는 전통적인 의미의 문헌자료 뿐만 아니라 신문, 잡지, 연구자료, 보고서, SNS 데이터 등을 포함

오랜 기간을 거쳐 축적된 문헌도 현대의 IT기술로 디지털화되면서 양적인 면에서 텍스트 데이터 분석 대상이 되는 문헌자료가 급격히 증가

텍스트 데이터의 급증을 이룬 또 하나의 이유는 소셜 네트워크 서비스 사용의 확대

텍스트 마이닝: 텍스트 데이터로부터 유용한 정보를 추출해내는 방법

텍스트 마이닝은 통계학, 데이터 마이닝, 데이터베이스, 문헌정보학, 컴퓨터 언어학 및 인공지능 기법들이 종합된 것

3. 텍스트 데이터의 수집 방법

3.1 데이터 저장소를 통한 텍스트 데이터 수집

간단한 실습용 데이터 수집에 적합

3.2 API를 통한 텍스트 데이터 수집

API: 프로그램과 프로그램 간의 연결고리 / 운영체제나 프로그래밍 언어가 제공하는 기능을 제어하는 인터페이스

IT 관련 기업들은 API를 통하여 일부 데이터를 수집할 수 있도록 공개 API를 제공

3.3 웹문서 데이터의 수집

웹 스크래핑(web scraping): 웹문서에서 데이터를 추출하는 기술

웹 크롤링(web crawling): 웹문서의 데이터를 긁어오는 기술

웹 크롤러(web crawler): 조직적이고 자동화된 방법으로 월드 와이드 웹을 탐색하는 컴퓨터 프로그램

4. 텍스트 데이터 수집 사례

< R 코드 참조 >

5. 유용한 R 패키지

5.1 textstem 패키지

5.2 stopwords 패키지

5.3 tidytext 패키지

5.4 wordcloud 패키지