미래창조과학부 NIA 한국정보화진흥원 NATIONAL INFORMATION SOCIETY AGENCY

빅데이터 분석전문가 양성을 위한

빅데이터 분석 실습 교육
- 통계적 자료 분석의 기초

투이컨설팅

미래창조과학부 NIA 한국정보화진흥원 NATIONAL INFORMATION SOCIETY AGENCY

Chapter Ⅰ. 데이터 수집단계 가이드

데이터 수집 단계 가이드

- 데이터 안에 숨겨진 정보를 파악하여 의미있는 스토리를 제공하기 위해서는 올바른 데이터의 수집이 필수적이다.

  특정한 분석을 위해 스스로 데이터를 수집할 수도 있지만 이에는 한계가 있고, 이미 많은 양의 데이터가 오픈 API를 통해 제공되고 있기 때문에 이를 이용하는 것도 한가지 방편이 될 것이다.

  다만, 공개되어 있는 데이터를 바로 이용하기에는 어려움이 있고, 자신의 분석 목적에 맞게 테이블을 변경하고 데이터를 가공해야 데이터의 새로운 가치를 부여할 수 있다.

  특히, 의미있는 스토리를 만들어 내기 위해서는 다른 데이터의 정보도 접목시켜 이용해야 하는 경우가 많이 있기 때문에 올바른 데이터의 구성은 중요하다고 할 것이다.

- 빅데이터를 수집하는 기술로는 SNS, 뉴스 등의 웹정보를 인터넷에서 수집하는 크롤링(crawling), 각종 센서를 이용해 수집하는 센싱(sensing), 분산 시스템에서 데이터베이스 관리 시스템인 카산드라(Cassandra), 운영체제와 응요프로그램 간의 통신에 사용되는 메시지 형식의 개방된 오픈 API 등이 있다.

  이들 기술에 대해 여기서 깊이 다루기는 한계가 있으므로 생략하기로 하고, 회귀분석, 기계학습 등과 같이 특정한 분석 목적에 부합하는 맞는 형태의 데이터를 구성하는 것은 이후 자료분석의 각 절에서 다루어질 것이다.

Chapter II. 데이터 전처리 및 저장단계 가이드

데이터 전처리 및 저장 단계 가이드

- 본 장에서는 데이터를 효과적으로 저장하기 위한 수단인 data.table 을 이용하는 방법에 대해 숙지하는 것을 목표로 함

- Issue

> data.frame 과 data.table 의 처리속도 차이를 비교해 보고 효과적인 이유에 대하여 체감

> 조건을 이용한 데이터 선택방법과 간단한 Grouping 연산방법을 예제를 통해 확인

Chapter III. 데이터 시각화 단계 가이드

데이터 시각화 단계 가이드

- 본 장에서는 수집된 자료를 효과적으로 정리, 요약하기 위해 도표, 그래프등을 사용하여 시각화 하는 기법을 다룸.

- Issue

> 자료들의 분포형태를 시각화 하여 대칭 혹은 비대칭의 정도, 이상점의 유무, 봉우리의 위치등을 파악
> 이어서 모집단, 표본, 표본분포의 개념을 이해

> ggplot2 그래픽 라이브러리를 이용한 고급 시각화

Chapter IV. 프로세스별 데이터셋 구성 및 분석단계 가이드

프로세스별 데이터셋 구성 및 분석단계 가이드

- 통계적 추론 결과를 예측(Prediction)에 이용하는 분석 방법 중 가장 기본이 되는 회귀분석, 로지스틱 회귀분석 등을 익히고, 더 나아가 통계적 기계학습에 이해한다.

- Issue

> 서로 다른 변수 간 함수관계를 규명하는 회귀분석의 큰 틀을 이해한 후 다중회귀분석, 비모수적 회귀분석, 로지스틱 회귀분석을 이해

> 위의 내용을 근간으로 지도학습과 자율학습의 의미를 생각

> 지도학습에 의한 분류방법, 자율학습에 의한 분류방법 각각을 예제를 통해 확인

Q&A

미래창조과학부

NIA 한국정보화진흥원
NATIONAL INFORMATION SOCIETY AGENCY

투이컨설팅