**학습계획서**

|  |  |  |  |
| --- | --- | --- | --- |
| 팀 | 준비데이터 | 구성원 | 김호영, 안상현 |

|  |  |  |  |
| --- | --- | --- | --- |
| 일정 | 발제자 | 주제 | 주요내용 |
| 1일차  ( 5 / 27 ) | 안상현 | -빅데이터 이해  -관련 개념 | -빅데이터 개념 정의  -데이터베이스, NoSQL, SQL |
| 2일차  ( 5 / 28 ) | 안상현 | -빅데이터 기획 | -기획 배경 및 필요성  -다양한 기획 방법(top-down, bottom-up) |
| 3일차  ( 5 / 29 ) | 김호영 | -빅데이터 전처리 | -데이터 전처리 방법  -R의 데이터 타입 |
| 4일차  ( 5 / 30 ) | 김호영 | -빅데이터 처리 방법(Hadoop) | -HDFS  -MapReduce  -Hadoop Eco System |
| 5일차  ( 5 / 31 ) | 김호영 | -빅데이터 분석  (통계 분석1) | -통계 기초 개념, 대푯값  -표본과 모집단, 정규분포 |
| 6일차  ( 6 / 3 ) | 김호영 | -빅데이터 분석  (통계 분석2) | -t검증, 분산분석  -상관분석, 회귀분석 |
| 7일차  ( 6 / 4 ) | 김호영 | -빅데이터 분석  (데이터마이닝1) | -데이터마이닝의 개념과 분류  -텍스트마이닝 개념 및 절차 |
| 8일차  ( 6 / 5 ) | 안상현 | -빅데이터 분석  (데이터마이닝2) | -군집분석, SVM, 주성분분석  -의사결정나무, 인공신경망 |
| 9일차  ( 6 / 7 ) | 안상현 | -빅데이터 분석  (신기술 동향) | -딥러닝 개념 및 절차  -딥러닝 소프트웨어(tensorflow, keras 등) |
| 10일차  ( 6 / 10 ) | 안상현 | -빅데이터 시각화 | -시각화 이론 및 프로세스  -R을 활용한 시각화 방법(ggplot2) |

**학습 정리**

|  |  |  |  |
| --- | --- | --- | --- |
| 팀 | 준비데이터 | 구성원 | 김호영, 안상현 |

|  |  |  |  |
| --- | --- | --- | --- |
| 일정 | 발제자 | 주제 | |
| 5/27 | 안상현 | 1. 빅데이터 정의 및 관련 개념  2. 인터넷 강의 요약 | |
| 주요 내용 요약 | | | |
| **Ⅰ오늘의 ADsP 요약**   1. 빅데이터의 정의   기존 DBMS(데이터베이스 관리시스템)에 보관하기 힘든 양의 데이터 집합  이 집합에는 정형 데이터, 비정형 데이터 등 다양한 타입이 포함될 수 있다.  최근 SNS 사용 활성화, 데이터 처리 비용 감소, 클라우드 발달 등으로 빅데이터가 유망 산업에 포함되고 있다.   1. 빅데이터의 특징(3V + α)   ---- 전통 3V  Volume(양) : 데이터의 양이 매우 커졌다.(수십 TB ~ 수 PB)  Variety(다양성) : 생성/수집되는 데이터의 종류가 매우 다양하다(텍스트, 이미지, 동영상, 숫자 등)  Velocity(속도) : 데이터의 생성 속도가 매우 빠르다  ---- 새로운 V  Value(가치) : 방대한 양의 데이터에서 가치있는 정보를 이끌어낼 수 있다  Veracity(정확성) : 가치를 이끌어내기 전에 수집한 데이터가 정확한지 살펴야 한다  Variability(가변성) : 데이터는 맥락에 따라 의미가 달라질 수 있다  Visualization(시각화) : 빅데이터를 분석한 내용을 이해하기 쉽게 시각화해야 한다   1. 빅데이터 분석 방법   - 통계 분석 : 기술통계(빈도분석 등), 추리통계(카이제곱 검정, T검정, 분산분석, 상관분석, 회귀분석, 주성분 분석)  - 데이터 마이닝 : 로지스틱 회귀분석, 인공신경망, 의사결정나무, 앙상블 모형, 군집분석, 연관분석, 텍스트마이닝   1. 빅데이터 관련 개념   - RDBMS(관계형 데이터베이스)  엑셀과 비슷한 방식으로 데이터를 저장하는 소프트웨어이다. 엑셀에서 행과 열은 RDBMS에서 각각 row와 column으로 부른다.  컬럼은 데이터를 식별하는 역할을 맡는데, 구체적으로 기본키, 후보키 등으로 나눌 수 있다.  규칙적으로 데이터를 저장하므로 SQL을 통해 데이터를 삽입, 조회, 삭제, 수정이 가능하다.  자주 사용되는 소프트웨어로는 Oracle, MySQL, SQLite이 있다.  - NoSQL(Not Only SQL)  기존 RDBMS와 다른 방식으로 데이터를 저장하는 방식이다. 소프트웨어마다 각기 다른 방식으로 데이터를 저장하므로 통일성이 없다.  일반적으로 자주 사용하는 저장 방식은 JSON과 같이 key와 value로 저장한다.  따라서 SQL 대신 Key에 대한 Put(저장)과 Get(읽기)만 지원한다.  자주 사용되는 소프트웨어로는 몽고DB, 카산드라, Hbase가 있다.  - 데이터, 정보, 지식, 지혜  데이터는 가공 전의 순수한 수치나 기호를 나타낸다. raw data라고 불리며, 그 자체로는 어떠한 의미를 가지지 않는다.  정보는 데이터를 가공하고 연결시킴으로써 패턴을 찾아 의미를 부여한 경우에 사용한다.  지식은 정보를 연결함으로써 만든 결과물이다. 지식을 통해 어떻게 행동할지 예측할 수 있다.  지혜는 지식을 내면화시켜 개인별로 만든 창의적 아이디어이다.  **Ⅱ 오늘의 강의 요약**   1. 러닝패킷 : 통계의 기초 1강 - 공통   - 통계가 필요한 이유  통계는 사회 및 다양한 현상을 이해해 규칙성을 발견할 수 있는 기술이므로 통계를 바탕으로 예측함으로써 위험을 대비할 수 있다.  - 자료의 2가지 형태  ① 비연속변수/질적변수/범주형변수 : 몇 개의 유형으로 분석대상을 구별하기 위한 척도, 사칙연산을 적용하기 어렵다.(명목변수 + 순서변수)  ② 연속변수/양적변수 : 분석대상의 성질 크기가 얼마인지 파악하는 척도, 사칙연산이 가능하다.(등간변수 + 수치변수)  ③ 명목변수 : 각 항목에 임의로 숫자를 붙여 구별하는 척도, 항목 간에 연관성이 없고 순서가 우열을 나타내지 않는다. ex) 성별, 종교, 혈액형  ④ 순서변수 : 각 항목을 순서적으로 나열하지만 항목 간 차이가 일정하지 않은 경우  ex) 계층(상/중/하), 만족(1/2/3/4/5)  ⑤ 등간변수 : 수치 간 차이는 일정하지만 절대 0점이 없는 경우 ex) 온도, IQ점수  ⑥ 수치변수 : 절대 0점이 있어 상대적인 크기비교가 가능한 척도 ex) 키, 몸무게  - 범주형 자료의 요약  범주형 자료는 사칙연산이 불가능하므로 우열을 나타낼 수 없다. 따라서, 단순 빈도로 비교하거나 빈도를 잘 보여주는 그래프를 이용해야 한다.  ex) 도수분포표, 막대그래프, 원도표  - 양적 자료의 요약  양적 자료는 사칙연산이 가능하므로 비교가 가능하다. 다양한 대표값과 그래프를 통해서 표현할 수 있다. 다만, 범위가 넓으므로 이용자가 한눈에 이해할 수 있게 그래프를 표시해야 한다.  ex) 산점도, 히스토그램, 막대상자, 꺾은선 그래프   1. 빅데이터 분석과 R프로그래밍 - 김호영  * R이란?   1995년 Ross Ihaka and Robert Gentleman 이 S언어를 개선하여 개발한 오픈소스 통계분석 프로그램으로 빅데이터 분석 도구이다.  다양한 통계기법과 시각화 도구 제공하며, 통계 기법에는 통계기법, 데이터마이닝, 기계학습, 텍스트마이닝, 인공지능 분석툴이 있고,  시각화 도구에는 Ggplot을 이용한 시각적분석, 공간분석 맵이 있다.  C++, JAVA, Python 등 다른 프로그램과 쉽게 연동 가능하다는 특징이 있다.   * 전세계 학계나 연구기관에서 범용적 사용하며 기업체에서도 R을 탑재엔진으로 기업형분석 플랫폼개발 중   + 확장자는 \*.r * R 패키지 종류   데이터마이닝 툴   * + svm, cluster, tree, …   텍스트마이닝 툴   * + NLP, wordcloud, tm   그래픽툴   * + ggplot, lattice * 추가패키지 설치(install.packages)   항상 인터넷이 연결된 상태로 사용해야한다.  설치는 한번만 사용  해당 패키지를 사용할때 library는 사용할때마다 필요하다.  패키지 이름을 모를 경우에는 웹검색을 통해 찾을 수 있으며, 설치 후 help를 이용하여 상세설명을 알 수 있다.   * R프로그램을 수행하는 단축키   + Ctrl + Enter   + 수행하고자 하는 프로그램 부분을 선택하고 수행   + 주석은 #으로 처리가능  1. 머신러닝기법과 R프로그래밍 - 안상현  * 오늘 밤이나 내일 정리해서 적을 예정임   **참고 문헌**  - 데이터 관련 개념  <https://ko.wikipedia.org/wiki/%EB%B9%85_%EB%8D%B0%EC%9D%B4%ED%84%B0>  <http://egloos.zum.com/yjhyjh/v/39721>  - 러닝패킷 강의  <http://www.kocw.net/home/search/kemView.do?kemId=694004>  이종익, 박민석 편저, 「사회조사분석사 2급 필기」, 시대고시기획, 2014.  - 빅데이터 분석과 R프로그래밍  <https://pabi.smartlearn.kr/courses/course-v1:POSTECH+DSB101+P902/course/> | | | |
|
|
|
|
|
|
|