3/16 강의

빅데이터의 기술 분류



크게 2가지로 나누면 “빅데이터 저장과 관리” + “빅데이터의 활용”으로 나눌 수 있음

빅데이터의 활용을 주로 수업하게 됨, 빅데이터 생성은 거의 다루지 않음 (생성 + 수집)

저장 : 빅데이터를 쉽고 빠르게 다루기 위해서 정리함 (분산파일시스템은 배우지 않음)

처리 : 분산병렬처리 = 다수의 컴퓨터를 활용 “데이터를 쪼개서 컴퓨터에 분배” + “각 컴퓨터의 연산능력을 합쳐서 하나의 데이터를 다루기” (다루지 않음)

분석 (강의에서 주로 다룸) : 데이터마이닝 (데이터풀에서 쓸만한 데이터를 뽑아낸다), 텍스트마이닝(텍스트에서 데이터를 뽑아냄)

시각화도 강의에서 주로 다룸

분산파일시스템 (소개만)

* 컴퓨터 네트워크로 공유하는 여러 호스트 컴퓨터 파일에 접근할 수 있는 파일 시스템
* 대표적으로 하둡이라는 것이 있음 (HDFS)

병렬 처리 : 여러 연산 시스템이 데이터를 나누어 처리하는 것

자격증….

텍스트이(가) 표시된 사진

자동 생성된 설명

ADsP가 국가 공인이며 쓸만함 (필기만 있고 난이도는 쉬움)

난이도를 좀 높인게 빅데이터분석기사… 기사니까 권위 있음

ADP… 개같이 어려움 실무 3년이라도 6개월~1년공부 각오해야됨 (거의 박사급) 있으면 취업이 문제가 아님

빅데이터의 수집

크롤링 : 인터넷에 공개되어 있는 데이터를 컴퓨터를 이용하여 원하는 것만 수집하는 기술

데이터과학

텍스트이(가) 표시된 사진

자동 생성된 설명

컴퓨터 공학 + 분석학 + 통계학이 합쳐짐

도메인에 대한 지식이 필요함

데이터 수집 시작

노이즈를 포함한 원시 데이터 : 여러가지 소스에서 데이터를 수집… 빅 데이터는 다다익선 = 일단 최대한 많이 수집하는 것이 좋다 (나중에 사용할 데이터만 분류하면 됨)

이후 목적에 따라 2가지로 나뉨

인공지능 프로젝트 : 기계학습용 데이터준비 – 기계학습, 모델최적화 – 지능시스템 품질평가

분석 프로젝트 : 데이터분석, 모델 선정 – 모델최적화, 패턴추출 – 분석 결과 시각화/평가

인공지능, 머신러닝, 딥러닝

도표이(가) 표시된 사진

자동 생성된 설명

알파고, 챗GPT, 게임 봇 등… = 전부다 인공지능

사람의 행동을 유사하게 구현한것들이 전부다 인공지능임

머신러닝 : 기계가 **“코드로 명시하지 않은”** 동작까지 실행하는 것. (인공지능과의 가장 큰 차이)

딥러닝 : **“깊은”** 인공신경망을 사용하는 머신러닝 (정의는 좀 애매한데 가장 큰 차이는 다각화된 판단이 가능함)

데이터의 양이 많지 않으면 머신러닝이 성능이 좀 더 좋음 (어느정도 규격화가 되어 있음) 그대신 성능의 한계점이 있음, 사람의 개입이 많음

딥러닝은 깊은 인공신경망 작성을 해야되기 때문에 데이터가 적으면 성능이 떨어짐 (신생아와 비슷함) 데이터가 많으면 딥러닝 ㄱ, 사람의 개입이 적음 (일반인이 생각하는 인공지능에 더 가까운 개념)

머신러닝과 딥러닝의 성능 교차점은 아무도 모름

데이터가 많을수록 좋긴한데… 처리할 데이터가 적을수록 데이터의 질이 좋음

머신러닝의 종류

도표이(가) 표시된 사진

자동 생성된 설명

지도학습 : 정답을 알려주고 학습을 시킴, y=f(x)… x를 넣으면 y가 나온다 라고 알려줌

지도학습은 분류와 회귀로 나눠짐

분류 : 어떤 범주에 속하는지 예측 (ex. 스팸메일 분류, 필기체 인식) 데이터를 2차원 상의 평면에서 나눔

회귀 : 어떤 값이 나오는지 예측 (ex. 주가 분석) 데이터를 2차원 상에서 선으로 그을수 있음

비지도학습 : 정답을 안알려주고 학습. 데이터 자체에 집중해야됨. 데이터의 구조와 특징을 직접 발견함

클러스터링 : 데이터를 그룹화함 (마케팅쪽에서 가장 많이 활용) 지도학습-분류와 다른 것은 주어진 정답이 없다는 것

강화학습 : 보상 시스템으로 학습 (알파고가 대표적) 규칙만 던져주고 최선의 결과만 도출하는 과제를 줌

파이썬 다루기!

1.자료형 : 파이썬에서 다루는 데이터가 어떤 형태가 있는지



네모 하나가 셀(Cell)이라고 부름

초록색 : 셀 안에 뭔가를 입력할수 있다

파란색 : 셀의 상태를 제어할 수 있는 상태 (입력이 아닌 설정을 변경할수 있는 단계) 여기서 m을 입력하면 In[]: 이 사라짐 -> 마크다운을 입력할 수 있음

마크다운 : 메모를 작성할 수 있는 형식 (C++에서 ## 이거랑 비슷한듯?)

M을 누르면 마크다운, Y를 누르면 마크다운 해제 가능

Shift + Enter : 셀을 실행함

마크다운에서 #은 제목을 의미, 한칸을 띄워야 작동하고 # 4개까지 붙여서 소제목 단계로 내려갈수 있음

파란색 상태일 때 A(Above)를 누르면 위에 셀이 생기고 B(Bottom)을 누르면 아래에 셀이 생김

지우고 싶으면 X 지운거 실행취소는 Z

텍스트이(가) 표시된 사진

자동 생성된 설명

a, b, c라는 변수에 숫자를 정의

print(a/b/c) : 각 a,b,c를 출력

print(type(a/b/c)) : 각 a,b,c가 어떤 종류의 데이터인지 출력

class ‘int’ (integer) 이것은 정수 타입의 숫자 데이터다 라는 의미