3/14 강의

파이썬 기초는 추가로 따로 배우자

판다스에서 데이터베이스를 다루니 그 구조를 이해하는 것이 좋다

넘파이 : 파이썬으로 행렬연산 (선형대수)를 구현, 수학 라이브러리이며, 어렵게 말하면 다차원 개념 배열을 사용하는 라이브러리 머신러닝의 선행

시각화 : 사람은 머신러닝이 인지하는 패턴을 인지하기 어려움. 그래프 등으로 인간이 패턴을 인지할수 있도록 시각화함

데이터 분석 프로젝트 : 머신러닝의 선행, 이전까지 배운 내용들을 복습하는 과정이기도 함

머신러닝 5주

딥러닝 5주

커리큘럼이 모든 것을 완성시켜주지 않음 스스로 공부할 수 있도록 기초를 잡아주는 거임

6개월차에 3인 이상 팀을 구성해서 프로젝트를 진행할거임 ㅇㅇ

개인컴퓨터에 파이썬이 여러 개 깔리면 충돌 일어나니까 잘 설치하셈 ㅇㅇ

R과 파이썬

프로그래밍 언어 : 컴퓨터의 의사수단 소통

R : 1. 연구용 언어 2. 배우기 어려움 (통계학 지식이 필요함, 지식이 있다면 오히려 쉬울수도 있음) 3. 그래프 그리기는 쉬움 4. 통계학 기반 데이터 분석

파이썬 : 1. 개발용 언어 2. 배우기 쉬움 3. 빅데이터 분석에 용이함 4. 최신 인공지능 연구 트렌드 반영

데이터 분석하는데는 엑셀도 쓸수 있으나… 대용량 처리는 어렵기에 R (이것도 한계가 있음) -> 파이썬을 사용함

파이썬만 사용하는 회사는 많은데 R만 사용하는 회사는 거의 없음

빅데이터/AI 분야에서 사용하는 라이브러리가 있긴한데 아나콘다가 그러한 라이브러리들을 다같이 설치해줌

파이썬 기본 툴은 직관적이지 않고 코드 재활용에 불리함 아나콘다 같은 통합 개발 환경(IDE)가 필요함 (메모장 – 워드와 비슷한 관계) 정확히는 아나콘다가 IDE를 포함 = 빅데이터에서 사용하는 파이썬의 배포판

빅데이터가 뭔데?

* 말 그대로 존나게 큰 데이터, 쉽게 말해서 더블클릭해서 열면 컴퓨터 터지는 용량의 데이터 (일반적으로 수용할수 없는 정도의 크기의 데이터)

빅데이터의 특징

양, 다양성, 속도 = 3V

양 : 수용 한계를 넘어섬

다양성 : 숫자 뿐 아니라 다양한 형식의 데이터들을 분석이 가능한 형태로 정리가 되어야 함 (비정형 포함)

속도 : 데이터의 생성 속도 + 데이터의 분석 속도 = 실시간으로 빠르게 생성되고 처리되어서 바로 쓸수 있어야 함

+ 2V

Veracity : 정확성, 데이터가 유효하여야 함

Value : 가치, 데이터를 통해서 뭔가를 얻을 수 있어야 함 (다양한 시각도 좋다)

프로젝트 (항상 이렇게는 아님)

분석 기획 – 데이터 수집 & 전처리 – 데이터 분석 & 시각화 – 웹 기능 구현

역할

데이터 엔지니어 : 빅 데이터를 저장 관리하고 데이터를 사용하는 쪽에서 원활하게 사용할 수 있도록 데이터를 관리되고 보관되고 처리하고 쓸수 있게 하는 것 (데이터 파이프라인 구현) 을 함 데이터를 잘 알거나 클라우드/네트워크 잘 하고 데이터 수집, 전처리, 관리하는 내용을 잘 다루면 좋다 (스펙 : 하둑, 쿠버네티스, AWS, DCP)

머신러닝 엔지니어 (이번 강의의 초점) : 머신러닝 모델링 담당. 데이터 사이언티스트의 요청을 받아서 작업 (체계적이지 않은 회사는 업무가 지랄남 ㅋㅋㅋ)

데이터 분석가 : 데이터 시각화를 기반으로 대쉬보드 개발, 경영판단에 도움이 되는 분석자료를 통하여 인사이트(통찰력)를 뽑아냄. 머신러닝은 존나게 필요함

데이터 사이언티스트 : 어느 곳은 데이터 분석가이기도 함, 혹은 나머지 포지션의 어드밴스드 형태, 전반적인 프로젝트의 설계 및 운영을 담당함

빅데이터 활용을 위한 3가지 요소

= 자원(데이터, 그와 관련된 업무와 장비… 좋은 컴퓨터라던가)

+ 기술(자연어 처리 = 챗GPT같은 텍스트 기반, 컴퓨터 기반 = 이미지 기반, 음성인식 = STT, TTS등 음성기반…. 얘가 가장 어려움)

+ 요소(필요 역량 : 수학, 통계학, 컴퓨터공학, 인문, 사회지식 등등등)