[ 데이터 분석 ]

A. 절차

1. 연구목표 설정

- 6단계 산출물을 위해서 어떤 프로젝트를 진행할 것인지 계획(기획)

- WBS, 기획서, 요구사항분석등...

- 목표를 명확하게 설정

- 기존 시스템의 예측 정확도를 3% 증가시킨다?

- 상관관계를 분석해서 예측을 수행한다

- 의사결정을 위한 레포트를 구성한다

- 데이터가 존재하는가? 확보할 수 있는가?

2. 데이터 획득(수집)

- 데이터가 없으면 프로젝트 수행 불가

- 데이터의 수준에 따라 4단계에서 수집 가능

- level 1

- 공공데이터, 사내데이터, 연구용 데이터, 데이터 바우처를 이용한 구매 데이터 = > 이미 정제된 형태로 제공된다.

- 데이터 거래소(https://kdx.kr/main)

----------------- 아래 난이도를 가진 방식은 인터넷상에서 구하는 방법이다 ------------------

- level 2

- open API 활용

- 네이버, 카카오 등 api를 신청해서 쿼리를 통해 획득

- https://developers.kakao.com/docs, https://developers.naver.com/main/

- level 3

- Web Scrapping (웹스크레핑)

- 웹을 긁어서(읽어서) html을 확보 후 -> DOM(document object model)생성(파싱) -> 추출

- 사이트 로딩이 끝나면 사람의 조작 개입없이 바로 수집할 수 있는 수준의 사이트

- Beautifulsoup(bs4), request + 웹의 이해 + css selector 이해 (웹요청할 때 쓰는 것)

- https://www.crummy.com/software/BeautifulSoup/bs4/doc/

- level 4

- Crawling (크롤링)

- 사람이 조작해서 데이터가 나오는 형태인 경우 해당

- 웹기술로는 ajax, 로그인 등이 적용되어 있다.

- selenium

- 자동화(매크로)

- (데이터 수집을 스케줄 걸어서, 특정 시간마다, 특정 횟수반큼 수행하여 데이터를 획득)

- 윈도우, 리눅스에서 수행

=> 여기서부터는 구글의 colab을 이용하여 진행

3. 데이터 준비(전처리)

- 데이터의 품질을 향상시키는 목적

- 분석, 학습, 예측 등의 좋은 성과를 낼 수 있게 데이터를 전처리한다.

- 데이터 가공, 적제, 전처리 등 수행하는 단계

- 4단계와 거의 병행해서 움직인다.

- 핵심 라이브러리

- python => 정규식

: 연속데이터 타입 : [리스트], {딕셔너리}, (듀플), {",","}set집합, 문자열(Str)

- numpy : 선형대수학, 푸리에 등 기본 수학/과학용 라이브러리

pandas, 기타 딥러닝 엔진들의 근간(베이스 라이브러리)

기본 자료 구조 => 배열(ndarray) => 행렬

nd : n-demension n차원

- pandas :

: Series(1-Demension), DataFrame(2-D)

4. 데이터 분석

- 데이터를 분석함으로써 그 안에 숨겨진 패턴을 찾아내서(과거를 분석) 가까운 미래를 예측하는 것. => 빅데이터 분석 확장(스파크, 하둡..)

- 핵심 라이브러리

- numpy, scipy : 수학, 과학적인 계산의 기본모형제공 및 알고리즘제공

- pandas : 데이터 전처리, 분석 진행

- 시각화 -> 차트 (바,파이,산포도,산포행렬, ... <= 무엇을 설명할 때, 가장 적합한 차트인가?)

matplotlib, seaborn, folium,...

- 분석의 최종 산출물

=> 보고서(레포트)로 끝날 수 있고, 6단계까지 가서 서비스로도 갈 수 있다.

=> 근 미래에 대한 예측 수행

5. 데이터 모델 구축 (비중이 약하다, 생략)

6. 시스템 통합(산출물 or 서비스)

[ 머신러닝 ]

A. 절차

1. 연구목표 설정

2. 데이터 획득(수집)

3. 데이터 준비(전처리)

4. 데이터 분석(생략가능)

5. 예측 모델 구축

6. 시스템 통합(산출물 or 서비스)

[ 딥러닝 ]

A. 절차

1. 연구목표 설정

2. 데이터 획득(수집)

3. 데이터 준비(전처리)

4. 데이터 분석(생략가능)

5. 예측 모델 구축

- Tensor : 행렬이라 생각하면 됨

6. 시스템 통합(산출물 or 서비스)

-----------------------------------------------------------------------------

앞으로 python 기반으로 딥러닝을 수행한다면, 만날 수 있는 모든 연속형 자료구조

[].{},(),set (구성원들의 타입이 같을 필요 없다)

str

ndarray(모든 값의 타입을 동일,단일 타입)

Series (동일타입)

DataFrame(colomn끼리 타입이 같다,물론 타입이 다 같아도 상관없음.)

Tensor(모든 값이 수치(정수 or 부동소수)