[ 데이터 분석 ]

A. 절차

1. 연구목표 설정

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

- WBS, 기획서, 요구사항 분석 등

- 목표를 명확하게 설정

- 기존 시스템의 예측 정확도를 3% 증가시킨다

- 상관관계를 분석해서 예측을 수행한다

- 의사결정을 위한 레포트를 구상

- 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 인터넷 상에서 구하는 방법이다 -----------------------------------

- level 2

- open API 활용

- 네이버(developers.naver.com), 카카오(https://developers.kakao.com/) 등에서 api를 신청해서 쿼리를 통해 획득

- level 3

- Web Scraping(웹스크래핑)

- 웹을 긁어서(읽어서) html을 확보 후 → DOM 생성(파싱) → 추출

- 사이트 로딩이 끝나면 사람의 조작의 개입 없이 바로 수집할 수 있는 수준의 사이트

- Beautifulsoup(bs4), request + 웹의 이해 + css selector 이해

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

- level 4

- Crawling(크롤링) ↔ 매크로

- 사람이 조작해서 데이터가 나오는 형태인 경우 해당

- 웹기술로는 ajax, 로그인 등이 적용되어 있음

- selenium(파이썬)

- https://www.selenium.dev/

- 자동화(매크로)

- 데이터 수집을 스케줄 걸어서 특정 시간마다 특정 횟수만큼 수행하여 데이터를 획득

- 운영체제 영향 받음, 윈도우 혹은 리눅스에서 수행

- 매크로는 약간 어둠의 느낌, 매크로는 불법이 아니지만 매크로로 불법적인 일 많이 함

→ 여기까지 jupyter로 진행

→ 여기서부터는 구글의 colab을 이용해 진행

3. 데이터 준비(전처리)

- 데이터의 품질을 향상시키는 데 목적

- 분석, 학습, 예측 등에 좋은 성과를 낼 수 있게 데이터를 전처리

- 데이터 가공, 적재, 전처리 등을 수행하는 단계

- 4단계(분석)와 거의 병행해서 움직임 (because 검증을 해봐야 향상된건지 아닌지 알 수 있음)

- 핵심 라이브러리

- python → 정규식

- 연속데이터 타입 : 리스트 [], 딕셔너리 {}, 튜플 (), 집합/set {'','',''}(딕셔너리에 키가 없는 형태), 문자열(str)

- numpy

- 선형대수학, 푸리에 등 기본 수학/과학용 라이브러리

- pandas, 기타 딥러닝 엔진들의 근간(베이스 라이브러리)

- 기본자료구조 : 배열(ndarray) → 행렬

- nd : n-dimension(n차 배열)

- pandas

- 연속자료형 : Series(1-D, 1차원), DataFrame(2-D, 2차원)

4. 데이터 분석

- 데이터를 분석함으로써, 그 안에 숨겨진 패턴을 찾아내(과거 분석) 근미래를 예측 → 빅데이터 분석으로 확장 可(스파크, 하둡 등)

- 핵심 라이브러리

- numpy, scipy

- pandas

- 시각화

- 차트(바, 파이, 산포도, 산포행렬, ... ← 무엇을 설명할 때 가장 적합한 차트인가)

- matplotlib, seaborn, folium 등

- 분석의 최종 산출물

- 보고서(레포트)로 끝날 수 있고, 6단계까지 가서 서비스로도 갈 수 있음

- 근 미래에 대한 예측 수행

5. 데이터 모델링 구축(비중이 약하다/생략 가능)

6. 시스템 통합(산출물 or 서비스)

[ 머신러닝 ]

A. 절차

1. 연구목표 설정

2. 데이터 획득(수집)

3. 데이터 준비(전처리)

4. 데이터 분석(생락가능)

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(생락가능)

5. 예측 모델 구축

- 연속자료형 : Tensor(행렬이라고 생각하면 됨)

6. 시스템 통합(산출물 or 서비스)

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

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

[], {}, (), set, str, ndarray, Series, DataFrame, Tensor

[], {}, (), set : 구성원들의 타입이 같을 필요 없음

str

ndarray : 모든 값의 타입이 동일

Series : 동일 타입

DataFrame : Column끼리만 타입 동일, 물론 다 같아도 OK

Tensor : 모든 값이 수치(정수 또는 부동소수)