AiDA Lab.
Artificial Intelligence & Data Analysis Laboratory

2021 인공지능 소수전공

9~12차시: Pandas

2021.07.21 18:30~19:15

Seokhwan Yang

Pandas

- Pandas : Panel Data System
  - 데이터 분석을 위해 널리 사용되는 파이썬 라이브러리 패키지
  - R의 데이터프레임(Dataframe)과 유사한 형태의 DataFrame 객체가 핵심
  - 엑셀과 유사한 2차원 데이터 구조로 되어있어 데이터 전처리 및 가공 용이
  - 구글, 페이스북 및 데이터를 분석하는 주요 회사의 데이터 과학자는 거의 대부분이 Pandas를 이용 중

Pandas

- **Pandas의 개발 원인**
  - **개발자: 월스트릿의 투자운용회사(헤지펀드) AQR에 근무중이던 웨스 맥키니**
  - **개발 이유: 회사에서 사용하는 데이터 핸들링 툴이 마음에 들지 않음**
  - **계기**
    - 2008년 동료가 파이썬으로 간단한 알고리즘을 작성하는 것을 보고 파이썬에 입문
    - 파이썬의 SciPy를 접한 후 상용 통계도구를 대체하는 오픈소스 도구가 많음을 발견
    - 스탠포드 대학 조나단 테일러 통계학 교수의 오픈소스 패키지에서 관련 모듈 발견
    - 오픈소스를 참고하여 R의 DataFrame 객체를 파이썬으로 이식하는 작업 도전
    - 도전 1개월 만에 Pandas 초기버전 출시(2008년 초)

Pandas

- **회사에서 사용하던 데이터 분석 도구에서 원하던 기능**
  - 자동적, 명시적으로 축의 이름에 따라 데이터를 정렬할 수 있는 자료구조
  - 잘못 정렬된 데이터에 의한 일반적인 오류 예방
  - 다양한 소스에서 가져온 다양한 방식으로 색인된 데이터를 다루는 기능
  - 통합된 시계열 데이터 처리 기능
  - 시계열 데이터와 비 시계열 데이터를 함께 다룰 수 있는 통합 자료구조
  - 산술 연산과 한 축의 모든 값을 더하는 등의 데이터 축약연산은 축의 이름 같은 메타데이터로 전달 가능해야 함
  - 누락된 데이터를 유연하게 처리할 수 있는 기능
  - SQL 같은 일반 데이터베이스처럼 데이터를 합치고 관계연산을 수행하는 기능

Pandas의 자료구조

- **Pandas의 대표적인 자료구조**

DataFrame

- 표 같은 스프레드시트 형식의 자료구조
- 여러 개의 칼럼을 가지며 각 칼럼은 서로 다른 종류의 값을 담을 수 있음
- Row나 Column에 대하여 색인(Index)을 보유

Series

- 일련의 객체를 담을 수 있는 1차원 배열 같은 자료구조
- 어떤 NumPy 자료형이라도 담을 수 있음
- 배열의 데이터에 연관된 이름을 가진 색인(Index)을 보유

Index

- 표 형식의 데이터에서 각 Row와 Column에 대한 이름과 다른 메타데이터 (축의 이름)를 저장하는 객체
- DataFrame이나 Series 객체에서 사용됨

Pandas의 자료구조

- Pandas에서는
  - DataFrame과 Series만 알면(특히 DataFrame) 대부분의 애플리케이션에서 사용하기 쉽고 탄탄한 기반을 제공할 수 있음
  - 다른 자료구조도 있긴 있지만 위의 두 가지가 가장 중요함
  - DataFrame은 색인의 모양이 같은 Series 객체를 담고 있는 (파이썬 기본 자료형인)딕셔너리라고 생각하면 편함

Pandas의 자료구조

실습

DataFrame ~ Series ~ Index

Pandas의 자료구조

- **Pandas의 주요 Index 객체**

클래스	설명
Index	가장 일반적인 Index 객체 NumPy 배열 형식으로 축의 이름을 표현
Int64Index	정수 값을 위한 특수한 Index
MultiIndex	단일 축에 여러 단계의 색인을 표현하는 계층적 Index 객체 튜플의 배열과 유사
DatetimeIndex	나노초 타임스탬프를 저장 NumPy의 datetime64 자료형으로 표현됨
PeriodIndex	기간 데이터를 위한 특수한 Index

Pandas의 자료구조

- **Index 메소드와 속성**

메소드	설명
append	추가적인 Index 객체를 덧붙여 새로운 색인을 반환
diff	색인의 차집합 반환
intersection	색인의 교집합 반환
union	색인의 합집합 반환
isin	넘겨받은 값이 해당 색인 위치에 존재하는지 알려주는 불리언 배열 반환
delete	i 위치의 색인이 삭제된 새로운 색인 반환
drop	넘겨받은 값이 삭제된 새로운 색인 반환
insert	i 위치에 값이 추가된 새로운 색인 반환
is_monotonic	색인이 단조성을 가지는 경우 True 반환
is_unique	중복되는 색인이 없다면 True 반환
unique	색인에서 중복되는 요소를 제거하고 유일한 값만을 반환

Pandas의 자료구조

실습