판다스(Pandas)란?

Pandas Overview

Pandas는 쉽고 직관적인 관계형 또는 분류된 데이터로 작업 할 수 있도록 설계된

빠르고 유연하며 표현이 풍부한 데이터 구조를 제공하는 Python 패키지이다.

Python에서 실용적인 실제 데이터 분석을 수행하기 위한 고수준의 객체 형태를 목표로한다.

또한, 어떤 언어로도 사용할 수 있는 가장 강력하고 유연한 오픈 소스 데이터 분석 / 조직 도구가되는 더 넓은 목표를 가지고 있다.

Pandas는 다음의 종류의 데이터에 적합한 분석 패키지이다.

- SQL 테이블 또는 Excel 스프레드 시트에서와 같이 이질적으로 유형이 지정된 열이있는 데이블 형식 데이터
- 정렬되고 정렬되지 않은 시계열 데이터
- 행 및 열 레이블이 포함 된 임의의 행렬 데이터
- 다른 형태의 관찰 / 통계 데이터 세트

Pandas의 두 가지 주요 데이터 구조인 Series (1차원) 및 DataFrame (2차원)은 재무, 통계, 사회 과학 및

다양한 엔지니어링 분야의 일반적인 사용 사례의 대부분을 처리한다.

R 사용자의 경우 DataFrame은 R의 data.frame이 제공하는 모든 것을 제공한다.

Pandas는 NumPy를 기반으로하며 다른 많은 타사 라이브러리와 잘 통합되도록 설계되었다.

Pandas는 다음의 수행과정에 적합하다.

- 부동 소수점 데이터뿐만 아니라 누락 된 데이터(NaN)를 손쉽게 처리
- DataFrame 및 상위 차원 개체에서 열을 삽입하고 삭제할 수 있다.
- 입력하고자 하는 내용을 레이블 세트에 이름으로 정렬하거나 사용자가 레이블을 무시하고 Series, DataFrame 등으로 자동으로 데이터를 계산에 사용할 수 있다
- 데이터를 집계 및 변환하기 위해 데이터 세트에 분할 적용 및 유연한 그룹 별 기능 가능
- 다른 Python 및 NumPy 데이터 구조의 비정형 색인 생성 데이터를 DataFrame 객체로 쉽게 변환할 수 있다.
- 지능형 레이블 기반 슬라이싱, 인덱싱 및 대용량 데이터 세트의 하위 집합 가능
- 직관적인 데이터 병합 및 결합 가능
- 데이터 세트의 유연한 재 형성 및 피벗 가능
- 축의 계층적 레이블링 가능
- 플랫 파일(CSV), Excel 파일, 데이터베이스 및 초고속 HDF5 형식의 저장 / 로드 데이터에서 데이터 로드를 위한 견고한 IO 도구 포함
- 날짜 범위 생성 및 빈도 반환, 이동 창 통계, 이동 윈도우 선형 회귀, 날짜 이동 및 지연 사용 가능

이러한 원칙 중 많은 부분이 다른 언어 / 과학 연구 환경에서 자주 경험 한 단점을 해결하기위한 것이다.

데이터 과학자의 경우 데이터 작업은 일반적으로 데이터 정리, 분석 / 모델링 그리고 분석 결과를 플로팅 또는 표 형식으로 표시하기에 적합한 형식으로 구성한다.

Pandas 데이터 구조

차원	이름	설명
1차원	Series	균일한 유형의 배열로 표시된 1차원 데이터
2차원	DataFrame	잠재적으로 이질적으로 유형이 지정된 열이있는 크기가 가변적인 테이블 형식의 2차원 데이터

Pandas 데이터 구조를 생각하는 가장 좋은 방법은 더 낮은 차원의 데이터를 위한 유연한 컨테이너이다.

예를 들어, DataFrame은 Series의 컨테이너이고 Series는 스칼라의 컨테이너이다.

사전적으로 이러한 컨테이너에서 개체를 삽입하고 제거하는 편집 기술을 원한다.

또한 시계열 및 교차 단면 데이터 세트의 일반적인 방향을 고려한 공통 API 함수에 대해 적절한 기본 동작을 원한다.

ndarray를 사용하여 2차원 및 3차원 데이터를 저장할 때 함수를 작성할 때 데이터 집합의 방향을 고려해야하는 부담이 있다.

Pandas에서 축은 데이터에 더 많은 의미론적 의미를 부여한다.

그러므로 Pandas의 목표는 다운 스트림 기능에서 데이터 변환을 코딩하는 데 필요한 정신적 노력의 양을 줄이는 것이다.

예를 들어 표 형식의 데이터 (DataFrame)를 사용하면 축 0과 축 1 대신 행과 열을 생각하는 것이 의미적으로 도움이된다. 따라서 DataFrame의 열을 반복하면 더 읽기 쉬운 코드가 된다.

데이터의 변경 및 복사

모든 Pandas 데이터 구조는 값을 변경할 수 있다.

항상 크기를 변경할 수 있는 것은 아니다. 계열의 길이는 변경 할 수 없지만, 열을 DataFrame에 삽입 할 수 있다.

그러나 대다수의 메소드는 새로운 객체를 생성하고 입력 데이터는 변경하지 않는다.

일반적으로 Pandas는 불변성을 선호한다.