1. 시작하기 전에

1.1 이 책에서 다루는 내용

주로 파이썬 내용을 다룰 것이다.

1.1.1 어떤 데이터를 사용하다

여기서 '데이터'는 정확히 무슨 뜻일까? 주된 의미는 구조화된 데이터이다. 일부러 구조화된 데이터라는 모호한 표현을 썼는데, 다음과 같은 여러 가지 형태의 데이터를 포함한다.

- 각 컬럼의 형식이 문자열, 숫자, 날짜 등으로 서로 다른 표 혹은 스프레드시프와 비슷한 데이터, 이는 관계형 DB혹은 탭이나 쉼표로 구분되는 텍스트 파일 형식으로 저장되는 대부분의 데이터를 포함한다.
- 다차원 배열(행렬)
- SQL에서 기본키나 외래키 같은 키 컬럼에 의해 서로 연관되는 여러 가지 표
- 일정하거나 일정하지 않은 간격의 시계열

이 목록에 있는 형식이 전부는 아니다. 항상 명백하지는 않겠지만 대부분의 데이터는 모델링이나 분석을 위해 좀 더 쉬운 구조로 형태를 바꿀 수 있다. 또는 데이터 안에서 어떤 특성을 추출해서 구조회된 형태로 만들 수 있다. 예를 들어 뉴스 기사 모음은 사용 단어 빈도표를 만들어 감성 분석에 사용할 수도 있다.

아마도 전 세계적으로 가장 널리 사용되고 있는 데이터 분석 툴인 엑셀 같은 스프레드시트 프로그램 사용자는 이런 종류의 데이터가 낯설지 않을 것이다.

1.2.3 파이썬을 사용하면 안 되는 이유

파이썬이 분석 어플이나 범용 시스템을 개발하는 데 훌륭한 환경이긴 하지만 특수한 경우에는 파이썬이 아닌 다른 언어가 해답인 경우도 있다.

파이썬은 인터프리터 언어이므로 자바나 C++ 같은 컴파일 언어보다 많이 느리다. 하지만 개발자의 시간 비용은 CPU의 시간비용보다 비싸므로 대개는 이런 등가 교환에 만족해한다. 그러나 실시간 거래 시스템처럼 매우 짧은 응답 시간을 필요로 하는 어플에서는 가능한 최고의 성능을 내고자 생산성은 떨어지지만 C++ 같은 저수준 언어로 개발을 한다.

파이썬은 동시다발적인 멀티스레드를 처리하거나 CPU에 집중된 많은 스레드를 처리하는 어플에 적합한 언어가 아니다. 바로 GIL(전역 인터프리터 잠금)때문인데, 이 메커니즘은 인터프리터가 한 번에 하나의 파이썬 명령만 실행하도록 한다. 왜 GIL이 존재하는지에 대한 기술적인 이유는 이 책에서 다루는 내용을 벗어난다.

1.3 필수 파이썬 라이브러리

이 책에서 사용하는 파이썬 데이터 환경과 라이브러리에 익숙하지 않은 독자를 위해 그중 일부를 간단히 소개한다.

1.3.1 NumPy

넘파이는 Numerical Python의 줄임말로, 파이썬 산술 계산의 주춧돌 같은 라이브러리다. 자료구조, 알고리즘 산술 데이터를 다루는 대부분의 과학 계산 어플에서 필요한 라이브러리를 제공한다.

- 빠르고 효율적인 다차원 배열 객체 ndarray
- 배열원소를 다루거나 배열 간의 수학 계산을 수행하는 함수

- 디스크로부터 배열 기반의 데이터를 읽거나 쓸 수 있는 도구
- 선형대수 계산, 푸리에 변환, 난수 생성기
- 파이썬 확장과 C,C++코스에서 NumPy의 자료구조에 접근하고 계산 기능을 사용할 수 있도록 해주는 CAPI

1.3.2 pandas

팬더스는 구조화된 데이터나 표 형식의 데이터를 빠르고 쉽고 표현적으로 다루도록 설계된 고수준의 자료구조와 함수를 제공한다. pandas의 주된 자료구조는 표 형태의 로우와 컬럼 이름을 가지는 DataFrame 과 1차원 배열 객체인 Series다.

pandas는 'NumPy의 고성능, 배열 연산 아이디어'에 스프레드시트와 관계형 DB(SQL같은)의 유연한 데이터 처리 기능을 결합한 것이다. 세련된 색인 기능을 제공하여 데이터 변형, 자르기, 취합 그리고 데이터의 부분집합을 선택할 수 있도록 해준다. 데이터를 처리하고 준비하고 다듬는 과정은 데이터 분석에서 중요한 부분이므로, pandas는 이 책에서 우선적으로 집중하는 라이브러리다.

pandas 라이브러리는 개발환경을 간단히 알아보자.

R은 파이썬과는 다르게 data.frame은 R프로그래밍 언어의 표준 라이브러리에 포함되어 있다. 결과적으로 pandas의 많은 기능은 R핵심 구현의 일부 또는 애드온 패키지에서 따왔다. pandas라는 이름은 다차원으로 구조화된 데이터를 뜻하는 경제학 용어인 패널 데이터와 파이썬 데이터 분석에서 따온 이름이다