기상청

3. 데이터 다루기

Python을 활용한 분석 기초 가이드 북

기상청

Ⅲ. 데이터 다루기

1. 데이터 로드
2. 데이터 생성
3. 데이터 접근
4. 데이터 변환
5. 데이터 저장

1. 데이터 로드 (1/3)

데이터를 불러올 때 Pandas 패키지를 이용하면 편리합니다. pandas 패키지를 이용한 데이터 로드는 다음과 같습니다.

● pandas 데이터 로드 함수

pandas 함수	설명
read_csv()	쉼표(,)로 구분된 파일을 읽어 옵니다.
read_excel()	Excel 파일을 읽어 옵니다.
read_table()	탭(₩t)으로 구분된 파일을 읽어 옵니다.
read_fwf()	구분자가 없는 파일을 읽어 옵니다.

• .csv 파일 불러오기

[구조]

```
import pandas as pd

데이터명 =
pd.read_csv('파일경로/파일명.csv')
```

[예제]

```
import pandas as pd

mydata =
pd.read_csv('C:/Users/user/Desktop/example/mydata.csv')
```

* 함수의 이름이 길거나 어떤 필요에 의해 함수의 이름에 가명(Alias)를 주고 싶은 경우, 위 예시처럼 “함수명 as Alias” 와 같은 표현을 사용할 수 있습니다.
* 본 교육자료의 예시 코드에 사용되는 주요 패키지에 대한 설명은 ’01_[edu]Python_소개’ 교육컨텐츠의 ‘4.사용 패키지 소개’에 기술되어 있습니다.

1. 데이터 로드 (2/3)

데이터를 불러올 때 pandas 패키지를 많이 이용합니다. pandas 패키지를 이용한 데이터 로드는 다음과 같습니다.

● pandas 를 이용해 데이터 로드 시 옵션 키워드

옵션 종류	설명
sep / delimiter	구분자를 지정합니다.
header	header를 컬럼이름으로 지정합니다.
index_col	index로 사용할 컬럼을 지정합니다.
encoding	파일의 인코딩을 지정합니다(한글일 경우 매우 중요)

- 파일 불러올 때 index 지정하기

[구조]

```
import pandas as pd
데이터명 =
pd.read_csv('파일경로/파일명.csv',
index_col='컬럼명or컬럼번호')
```

[예제] * 첫번째 컬럼을 index로 지정하는 경우

```
import pandas as pd
mydata =
pd.read_csv('C:/Users/user/Desktop/ex
ample/mydata.csv', index_col='0')
```

- .txt 파일 불러오기 (구분자가 | 로 되어 있는 경우)

[구조]

```
import pandas as pd
데이터명 =
pd.read_csv('파일경로/파일명.csv',
sep='구분자')
```

[예제]

```
import pandas as pd
mydata =
pd.read_csv('C:/Users/user/Desktop/ex
ample/mydata.csv', sep='|')
```

1. 데이터 로드 (3/3)

데이터를 불러올 때 Pandas 패키지를 많이 이용합니다. pandas 패키지를 이용한 데이터 로드는 다음과 같습니다.

● 기타 옵션 키워드

옵션 종류	설명
skiprows	읽지 않을 row 번호 list를 지정합니다.
na_values	NA 로 인식할 값 list를 기정합니다.
comment	주석으로 분류할 문자열을 지정합니다.
parse_date	날짜로 구분할 컬럼 list를 지정합니다.
date_parser	날짜 변환 시 사용할 함수를 지정합니다.
converters	컬럼을 읽어올 때 적용할 함수를 지정합니다.
nrows	몇 번째 줄까지 읽을 것인지 지정합니다.
skipfooter	무시할 파일의 마지막 줄 수를 지정합니다.

2. 데이터 생성

파이썬에서 생성한 데이터를 DataFrame 형식으로 변환할 수 있습니다.

● 데이터 생성(dictionary 사용해 입력하여 생성)

• 간단한 데이터를 생성해 DataFrame으로 변환하는 것은 다음과 같습니다.

[구조]

```
import pandas as pd

# dictionary 생성
dictionary 파일명
= {'컬럼명1': ['chr1','chr2','chr3'], '컬럼명2':
[num1, num2, num3]}
# 변수타입이 character인 경우 값을 작은 따옴표
사이에 입력, 숫자인 경우 숫자만 입력

# dictionary을 DataFame으로 변환
데이터 프레임 파일명 =
pd.DataFrame(dictionary 파일명,
index=['a','b','c','d','e'])
# index 키워드 뒤에 row 수만큼 index 명칭을 작은
따옴표 사이마다 입력
```

[예제]

```
import pandas as pd


data={'ID': ['A1','A2','A3','A4','A5'],
'X1': [1, 2, 3, 4, 5],
'X2': [3.0, 4.5, 3.2, 4.0, 3.5]}



mydata = pd.DataFrame(data,
index=['a','b','c','d','e'])
```

3. 데이터 접근

데이터에 접근하는 방법은 여러가지 입니다. 그 중 원하는 행 또는 열에 자유롭게 접근하는 것은 데이터 분석 시 매우 유용합니다. 접근하는 방법은 []를 활용합니다.

● 데이터/변수 접근

코드	설명
mydata.columns	데이터가 보유한 변수명을 확인합니다.
mydata.x1 또는 mydata['x1']	변수열(column)을 선택합니다.
mydata.ix[1] 또는 mydata.ix['row2']	특정 행(row)을 선택합니다.
del mydata['x1']	변수를 삭제합니다.
mydata.rename(columns ={'x1':'new_name'}, inplace=True)	변수명을 변경합니다.

[예제]

```
import pandas as pd

# x1변수를 선택해 상위 행 보기
mydata['x1'].head()

# x1변수의 두번째 행부터 4번째 행까지 보기
mydata[1:5]['x1']

# x1변수명을 item1변수명으로 변경
mydata.rename(columns={'x1':'item1'}, inplace=True)
```

4. 데이터 변환

파이썬의 내장 함수 중 데이터를 변환하는 함수는 다음과 같습니다.

● 데이터 변환 함수

내장 함수	설명
int()	base 진법의 수를 10진 정수형으로 변환
long()	base 진법의 수를 10진 long 형으로 변환합니다.
float()	수를 실수형으로 변환합니다.
complex()	복소수를 만듭니다.
str()	객체를 출력할 수 있는 문자열 그 자체로만 변환합니다.
repr()	객체를 출력 가능하고 eval 함수의 입력으로 쓰일 수 있는 문자열로 변환합니다.
eval()	문자열로 된 파이썬 식을 실행합니다.
tuple()	튜플로 변환합니다.
list()	리스트로 변환합니다.
set()	리스트나 튜플 등을 세트로 변환합니다.
dict()	사전 객체를 생성합니다.
frozenset()	변경이 불가능 한 세트로 변환합니다.
chr()	코드 값을 문자로 변환합니다.
unichr()	유니코드 값을 유니코드 문자로 변환합니다.
ord()	문자의 코드 값을 구합니다.
hex()	10진수에서 16진수로 변환합니다.
oct()	10진수에서 8진수로 변환합니다.

5. 데이터 저장

파이썬에서 생성한 데이터를 DataFrame 형식으로 변환하여 pandas 패키지의 DataFrame.to_csv() 함수를 사용해 csv파일로 저장할 수 있습니다.

● 데이터 저장

• 2장에서 생성한 데이터(DataFrame)를 구조화된 파일로 저장하는 것은 다음과 같습니다.

[구조]

```
import pandas as pd

데이터명.to_csv('파일을 저장할
경로/파일명.csv', 옵션)
```

[예제]

```
import pandas as pd

mydata.to_csv('C:/Users/user/Desktop/
example/mydata.csv',
...: sep=',',   # 값 구분자
...: na_rep='NaN')   # 결측값 표기 방법
```

함수	설명
to_csv	csv 파일로 내보냅니다.
to_excel	Excel 파일로 내보냅니다.
to_json	Json 파일로 내보냅니다.

본 문서의 내용은 기상청의 날씨마루(http://big.kma.go.kr) 내
Python 기초 교육 자료 입니다.