Python - Pandas

Jan Popko
Python Grundkurs

Business Trends Academy (BTA) GmbH
BUSINESS TRENDS ACADEMY
Nestorstraße 36
D-10709 Berlin
Tel.: +49 (0) 30 894 087 57
Fax: +49 (0) 30 895 429 94
Geschäftsführer:
Gabriele Fleischmann-Hahn
Maxi-Marlen Fleischmann
Hauptsitz des Unternehmens:
Nestorstraße 36, D-10709 Berlin
HRB 115251 B / Amtsgericht Berlin Charlottenburg
Steuer-Nr. 27/248/31179

Pandas

- Der Name leitet sich von dem englischen Begriff „Paneldaten“ ab
- Tool mit dem man sehr komfortabel tabellarische Daten einlesen, filtern und verarbeiten kann
- z. B. CSV oder Excel
- auch Pandas muss nachinstalliert werden
- Befehl für die Konsole: **pip3 install pandas**
- für Excel-Unterstützung: **pip install xlrd**
- eingebunden wird pandas mit **import pandas as pd**

Pandas

- **Series** entspricht in etwa einer eindimensionalen Liste, beispielsweise einer Zeitreihe, einer Liste, einem Dict, oder einem Numpy-Array.

- **Dataframe** besteht aus einer zweidimensionalen Tabelle. Die einzelnen Reihen beziehungsweise Spalten dieser Tabelle können wie Series-Objekte bearbeitet werden.

- **Panel** besteht aus einer dreidimensionalen Tabelle. Die einzelnen Ebenen dieser Tabelle bestehen wiederum aus Dataframe-Objekten.

Pandas

```
import pandas as pd
# ein Dataframe-Objekt wird erzeugt und die Quelle angeben aus
# der die Daten kommen, der delimiter spezifiziert das
# Trennzeichen in der CSV Datei, default ist das Komma
df = pd.read_csv("data/astronauts.csv", delimiter=",")

# hier wird der Inhalt des Dataframes ausgeben, head() sorgt
# dafür, dass nur die ersten 5 Einträge ausgeben werden
print(df.head())

# tail() sorgt dafür, dass nur die letzten 5 Einträge ausgeben
# werden
print(df.tail())

# die Anzahl der Einträge des Dataframes wird mit len() bestimmt
print(len(df))

# Zugriff auf eine einzelne Spalte
print(df["Name"])
```

Pandas

```
# Zugriff auf einen Datensatz, es wird auf den Index des
# Datensatzes zugegriffen
print(df.iloc[0])

# Zugriff auf den Namen dieses Datensatzes
entry = df.iloc[0]
print (entry["Name"])

# Bereichsabfrage(Slicing), Ausgabe der Datensätze 4 bis
# einschließlich 7
print (df.iloc[4:8])

# über alles Zeilen(Datensätze) des Dataframes iterieren,
# Rückgabewert ist ein Tupel
for row in df.iterrows():
    print(row)
```

Pandas

```
# in pos steht der erste Wert des Tupel(der Index), in data
# stehen die eigentlichen Daten des Tupels, auf die einzelnen
# Einträge der Daten kann man wieder mit dem Namen der Spalte
# zugreifen

for row in df.iterrows():
    pos = row[0]
    data = row[1]
    print(data["Gender"])
# sorgt dafür, dass nur der erste Datensatz angezeigt wird
    break
```

Pandas

Filtern:

```
import pandas as pd

df = pd.read_csv("data/astronauts.csv", delimiter=",")

# einzelne Spalte ausgeben
print(df["Year"])
# nur die Jahre ab 1990 zwischenspeichern
year = df["Year"] > 1990
# ausgeben, Rückgabewert False oder True
print(year)

#Filter auf Dataframe anwenden und ausgeben
print(df[year])

# Filter nur Military Rank ist Colonel
df2 = df[df["Military Rank"] == "Colonel"]
print(df2)
```

Pandas

Sortieren:

```
import pandas as pd

df = pd.read_csv("data/astronauts.csv", delimiter=",")

#sortieren nach Name aufsteigend
print(df.sort_values("Name"))

#sortieren nach Name absteigend
df2 = df.sort_values("Name",  ascending=False)

# Nur die Namen absteigend sortiert ausgeben lassen
for name in df2["Name"]:
    print(name)
```

Pandas

Einlesen einer Excel-Datei:

```
import pandas as pd

# Excel-Datei wird eingelesen
df = pd.read_excel("data/weather.xlsx")

# Spalte Sunshine Duration wird in sun gespeichert
sun = df["Sunshine Duration"]

# Datensatz mit dem Index 120 wird ausgeben
#print(type(sun[120]))
```