pandas & matplotlib

Nikola Bebić

Malo o obradi podataka

1. prikupljanje podataka
2. obrada "sirovih podataka"
3. eksplorativna analiza
4. upotreba svega toga

pandas biblioteka

- korisna za svakakvu obradu podataka

```
import pandas as pd
```

- osnovni koncept: DataFrame (tabela)

```
>>> df = pd.DataFrame({'a': [1, 2, 3],
...                    'b': [4, 5, 6]})

>>> df
   a  b
0  1  4
1  2  5
2  3  6
```

- osnovni koncept br 2: Series (jedna kolona)

```
>>> df.a # ili df['a']
0    1
1    2
2    3
Name: a, dtype: int64
```

selekcija

- vraća redove koji zadovoljavaju uslov

```
>>> df[df.a < 3]
   a  b
0  1  4
1  2  5

>>> df[(df.a < 3) & (df.b > 4)]
   a  b
1  2  5
```

učitavanje podataka

```
>>> pd.read_csv('imefajla.csv')
  col1 col2  col3
0    a    b     1
1    a    b     2
2    c    d     3
```

statistika

```
>>> df.mean()
a    2.0
b    5.0
dtype: float64

>>> df.a.mean()
2.0
```

- `max`, `min`, `sum`, `prod`, `std`, ...

spajanje

```
>>> first = pd.DataFrame({'key': ['K0', 'K1', 'K2'],
...                       'A': ['A0', 'A1', 'A2']})
>>> other = pd.DataFrame({'key': ['K0', 'K1', 'K2'],
...                       'B': ['B0', 'B1', 'B2']})
>>> pd.merge(first, other, on='key')
  key   A   B
0  K0  A0  B0
1  K1  A1  B1
2  K2  A2  B2
```

grupisanje

```
>>> df
   a  b  c
0  1  1  4
1  1  2  5
2  2  3  6
3  2  4  7
>>> df.groupby('a').sum()
   b  c
a
1  3  9
2  7  13
```

- sve one statističke metode

grupisanje (generalno)

```
>>> df.groupby('a').agg({'b': np.sum, 'c': np.max})
   b  c
a
1  3  5
2  7  7
```

matplotlib

- crtkaranje

```
>>> import matplotlib.pyplot as plt
```

vizualizacija podataka

- vremenska serija
- rangiranje
- odnos sa celinom
- frekventna distrubucija
- korelacija
- ...

https://en.wikipedia.org/wiki/Data_visualization

vremenske serije

- kretanje podataka kroz vreme
  - ili sličnu "kontinualnu" vrednost
- **linijski plot**

```
plt.plot(data, '-')
plt.plot(xs, ys, '-')
df_or_series.plot.line()
```

- primer: broj narudžbi u toku dana

rangiranje

- poređenje nekih vrednosti
- **bar chart**

```
plt.bar(labels, data)
series.sort_values().plot.bar()
```

- primer: najčešće kupljeni proizvodi

odnos sa celinom

- uticaj par klasa na celinu
- **pie chart**

```
plt.pie(data)
plt.pie(labels, data)
series.plot.pie()
```

- primer: "popularnost" različitih departmana

frekventna distribucija

- koliko se često neke vrednosti pojavljuju
- **histogram**

```
plt.hist(data) # bins, range, ...
series.plot.hist() # bins, ...
```

- primer: raspodela broja narudžbi po kupcu

korelacija

- da li su i koliko neke vrednosti u vezi
- **scatter plot**

```
plt.plot(xs, ys, 'o') # '.' za manje tačkice
df.plot.scatter('xlabel', 'ylabel')
```

- primer: da li češći kupci kupuju više/manje?
- korisno: `df.corr()`, `np.polyfit()`