pandas_intro

August 23, 2017

0.1 Pandas

Pandas je modul namenjen brzom i efikasnom radu sa podacima.

```
In[8]:  import pandas as pd
```

Za predstavljanje podataka se koriste dve strukture: **Series** i **DataFrame**

Series su strukture koje odgovaraju jednodimenzionim nizovima (čiji elementi mogu biti raznorodni) i koje imaju pridruženi niz obeležja tzv. indeks.

DataFrame su strukture koje odgovaraju tabelarnim podacima (eng. spreadsheet-like data structures) i koje imaju indekse i vrsta i kolona.

NaN označava nedostajuću vrednost. Sa **isnull** i **notnull** se može proveriti da li je neka vrednost nedostajuća ili ne.

```
In[2]:  # stampa
        #\?
```

```
In[9]:  # kreiranje serije
        p= pdS[(1, 1) , (2, 2), (3, 3)])
        p[0]))
        p
        import numpy as np
        p = pary[[1, 1] , (2, 2), (3, 3)])
        p[1]))
```

```
<s'p>
0    (1, 1)
1    (2, 2)
2    (3, 3)
 b
<s'pdy>
```

```
In[10]:  # ocitavanje elemenata serije
         #print(points.values)
         p
```

```
In[44]:  # ocitavanje indeksa kolona
         df.columns

Out[44]: Index(['age', 'city', 'name'], dtype='object')

In[45]:  # ocitavanje indeksa vrsta
         df.index

Out[45]: RangeIndex(start=0, stop=4, step=1)

In[46]:  # ispis vrednosti
         df.values

Out[46]: array([[24, 'Berlin', 'Ana'],
                [30, 'New York', 'John'],
                [33, 'Belgrade', 'Petar'],
                [21, 'London', 'Lana']], dtype=object)

In[47]:  # pristup pojedinacnim kolonama
         print(df['age'])
         print(df.age)

0    24
1    30
2    33
3    21
Name: age, dtype: int64
0    24
1    30
2    33
3    21
Name: age, dtype: int64


In[50]:  # pristup pojedinacnim vrstama
         df.ix[3]

Out[50]: age            21
         city       London
         name         Lana
         Name: 3, dtype: object

In[51]:  # dozvoljeno je i zadavanje opsega ako su u pitanju numericki indeksi
         df.ix[1:5]

Out[51]:    age    city  name
         1   30  New York  John
         2   33  Belgrade Petar
         3   21    London  Lana
```

```
In[52]:  # pristup pojedinacnim elementima
         ex1['na
```

```
Ot[52]: 'Jo
```

```
In[53]: pr
```

```
Ot[53]: pa
```

```
In[54]: pr
```

```
Ot[54]: pa
```

```
In[55]: pr
```

```
Ot[55]: pr
```

Mogu se koristiti različiti metodi filtriranja podataka:

```
In[63]:  # izdvajanje korisnika koji imaju vise od 25 godina
         pr
         pr
         pr
```

```
   ag
1   30
2   33


0
1
2
3
Na
```

Postoji mogucnost promene indeksa bilo za serije bilo za tabelarne podatke. Ukoliko se dodaju nove vrste / kolone, njihove vrednosti su NaN (konstanta np.nan)

```
In[68]: ex
         pr

         ex
         pr
```

```
   ag
0   24
1   30
2   33
3   21
```

```
    ag ab an al
0   24.0     Bh Aa      NaN
1   30.0   NwYk Jb    NaN
x   NaN         NaN    NaN     NaN
y   NaN         NaN    NaN     NaN
z   NaN         NaN    NaN     NaN
w   NaN         NaN    NaN     NaN
```

Provera nedostajucih vrednosti je moguca kroz **isnull** ili **notnull**

```
In[80]: pdl(e['al'])

Out[80]: 0     Te
         1     Te
         2     Te
         3     Te
         Nam al, p bb
```

Primer korišćenja raspoloživih podataka

```
In[15]:  # ucitavanje podataka iz CSV fajla
         # podaci su preuzeti sa https://www.kaggle.com/the-guardian/extinct-langua
         # i predstavljaju informacije o jezicima koji su ugrozeni ili u izumiranju
         # Kategorije koje se razmatraju su:
         # vulnerable - jezik se uci u nekom uzem kontekstu
         # definitely endangered - jezik se ne uci kao maternji jezik
         # severely endangered - jezik govore pripadnici starijih generacija, dok g
         # critically endangered - jezik govore samo pripadnici starijih generacija
         # extinct - jezici ciji govornici vise ne postoje

         lag= pedc('da/lagt')

In[70]:  # dodatne informacije o read_csv funkciji
         # stampa \?

         # na raspolaganju su i funkcije za analiziranje Excel formata, JSON format

In[71]:  # rezultat citanje je DataFrame struktura
         p(lag)

Out[71]: pd.sefaDataFam

In[72]:  # ispis broja vrsti ucitanog skupa podataka
         d(lag)

Out[72]: 2722

In[1]:  # ispis prvih 5 redova
        # stampa languages.head(5)
```

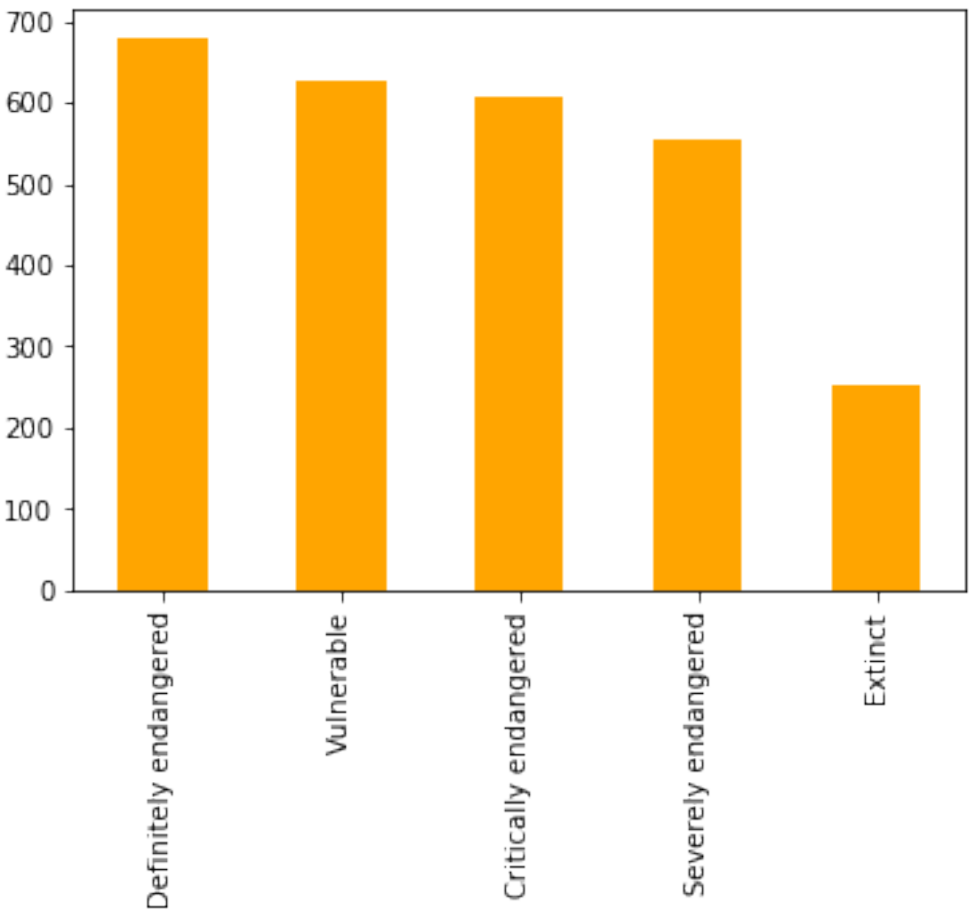
```
In[80]:  # value_counts prebrojava razlicite vrednosti u kolonama
         # stampa \?

In[81]:  # izdvajamo informaciju o broju jezika iz svake kategorije
         # kategorije koje su nam na raspolaganju su sadrzane u koloni Degree of en
         lang['Degree of ...].value_...()

Out[81]: Definitely endangered    680
         Vulnerable               628
         Critically endangered    607
         Severely endangered      554
         Extinct                  253
         Name: Degree of ..., dtype: int64

In[82]:  # iscrtavamo histogram na osnovu izracunatih podataka
         lang['Degree of ...].value_...().plot('bar', color='...')

Out[82]: <matplotlib.axes._subplots.AxesSubplot at 0x...>
```



** Dodatni materijali: **

- https://pandas.pydata.org/pandas-docs
- **knjiga**: Python for Data Analysis, Wes McKinney

```
In[ ]:
```