



Python Technológiák

Pandas, RDBMS

Lukács Gergely

Pandas

- Táblázatos adatok, heterogén típusú oszlopokkal
 - DB reláció, Excel táblázat
 - Idősorok (rendezett, rendezetlen, változó lépésű)
 - Mátrixok sor és oszlopcímkekkel
- NumPy alapú, jól integrálható mindenféle környezetbe
- Tudományos és ipari alkalmazások
 - Pénzügy, neurobiológia, közgazdaságtan, statisztika, hirdetés, Web elemzés

Pandas

- Series (1D)
 - data
 - Index
 - ...
- DataFrame (2D)

Jellemzők

- Hiányzó adatok (NaN)
- Oszlopok beszűrhetőak, törölhetőek
- Adat elrendezés címkék szerint
- Group by, aggregáció, transzformálás
- Kiválasztás
- Join
- Pivotálás
- Hierarchikus címkék
- CSV, Excel, RDBMS, HDF5

Gyakorlatok

- <http://pandas.pydata.org/pandas-docs/stable/tutorials.html#lessons-for-new-pandas-users>
- [01 - Lesson:](#) - Importing libraries - Creating data sets - Creating data frames - Reading from CSV - Exporting to CSV - Finding maximums - Plotting data
- [02 - Lesson:](#) - Reading from TXT - Exporting to TXT - Selecting top/bottom records - Descriptive statistics - Grouping/sorting data
- [04 - Lesson:](#) - Adding/deleting columns - Index operations
- [05 - Lesson:](#) - Stack/Unstack/Transpose functions
- [06 - Lesson:](#) - GroupBy function
- [08 - Lesson:](#) - Read from Microsoft SQL databases

Pandas vs SQL

- Pandas
 - egyszerű
 - hatékony
 - sokoldalú
- SQL
 - join tetszőleges kritériummal
 - nagyon nagy adatbázis (Pandas: memóriahatár)
 - deklaratív: végrehajtás nagyon jól optimalizált (költségalapú lekérdezés tervezés; ha jó az DB adminisztrátor)