

Python średnio zaawansowany

Dzień 11



Blok nr 4:

Analiza danych

AGENDA

- Przypomnienie: instalacja modułów (pip -r)
- Wprowadzenie do środowiska interaktywnego Jupyter
- Wprowadzenie do modułu NumPy
- Wprowadzenie do modułu pandas

Instalacja modułów: pip -r

Przepis na środowisko uruchomieniowe

Każda aplikacja oczekuje środowiska wyposażonego w wymagane moduły.

Istnieje możliwość przygotowania pliku „przepisu”, który informuje o wymaganych modułach i ich wersjach.

Przepis umieszcza się w pliku tekstowym (zwyczajowo: requirements.txt):

```
# Moduł warunek wersja  
numpy == 1.15.2  
pandas == 0.23.4
```

Przepis na środowisko uruchomieniowe

Ściągnięcie i instalacja:

```
pip3 install -r requirements.txt
```

Więcej informacji:

https://pip.pypa.io/en/stable/user_guide/#requirements-files

Środowisko interaktywne Jupyter

Jupyter: Notebook, kernels



- Jupyter Notebook to interaktywne środowisko do przetwarzania danych i dokumentowania tego przetwarzania
- Dostęp realizowany jest przez przeglądarkę, praca oparta o komórki (cells)
- Działa na zasadzie REPL (Read, Eval, Print, Loop) – podobnie do interpretera konsolowego

Jupyter: Notebook, kernels

- Pracę dokumentuje artefakt .ipynb – gotowy do edycji, eksportu i udostępnienia (!)
- Współpracuje z kernelami (jądrami) realizującymi przetwarzanie w wybranym języku programowania (obecnie ponad 40 – m.in., IPython, IRuby, IJulia, Ihaskell, IGo).
- Popularniejszy niż podobne rozwiązania: Maple, Mathematica

Uruchomienie, zatrzymanie

Środowisko Jupyter uruchamiamy w następujący sposób:

```
jupyter notebook
```

Od tej pory korzystamy ze środowiska przez przeglądarkę internetową.

Zatrzymanie: CTRL + C

Uruchomienie, zatrzymanie

Praktyka: notebooks/01_intro.ipynb

Więcej informacji: <https://jupyter.readthedocs.io/en/latest/index.html>

<http://johnlaudun.org/20131228-ipython-notebook-keyboard-shortcuts/>

Kolekcja notebooków: <https://github.com/jupyter/jupyter/wiki/A-gallery-of-interesting-Jupyter-Notebooks>

NumPy

NumPy (Numerical Python)



NumPy - popularna biblioteka wykorzystywana do obliczeń naukowych przy pomocy języka Python.

Pozwala między innymi na wykonywanie wydajnych operacji na tablicach wielowymiarowych, obliczenia numeryczne, obliczenia z zakresu algebry liniowej, FFT (szybka transformata Fouriera).

W zakresie możliwości porównywana do pakietu MATLAB.

NumPy

Podstawowy typ danych: ndarray (N-dimensional array)

Możliwości:

- bardzo wydajna podczas operacji na jednorodnych (**homogenicznych**) tablicach zawierających dane liczbowe:
 - wyodrębnianie podzbiorów, filtrowanie, sortowanie, przekształcanie
 - agregacje, podsumowania, wyszukiwanie
- mniejsze zużycie pamięci

NumPy

Ograniczenia:

- brak możliwości obsługi szeregów czasowych
- wymaga aby przetwarzane dane były jednorodne

Praktyka: notebooks/02 i 03

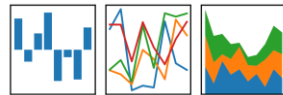
Więcej informacji: <https://docs.scipy.org/doc/numpy/user/quickstart.html>

| **pandas**

pandas (panel data)

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Biblioteka udostępniająca struktury i narzędzia przeznaczone do przetwarzania danych umieszczonych w formie tabel lub danych o charakterze heterogenicznym.

Stosowana w systemach przetwarzających ogromne ilości danych (big data).

pandas

Podstawowe typy danych: series, dataframe

Praktyka: notebooks/04_pandas_basics.ipynb

Dzięki!

