Danologia - Laboratorium nr 9

Jakość danych

1. **Wektoryzacja danych.**

   Pakiet NumPy wraz z pakietami Pandas i Matplotlib jest częścią składową biblioteki SciPy.

   NumPy jest podstawowym pakietem wykorzystywanym do obliczeń naukowych w języku Python. Pozwala między innymi na wykonywanie wydajnych operacji na macierzach, obliczenia numeryczne, obliczenia z zakresu algebry liniowej, FFT etc.

   Umiejętność przeywarzania wektorów, macierzy i innych rodzajów tablic tworzonych za pośrednictwem pakietu NumPy jest niezmiernie ważna.

   Załadowanie pakietu NumPy:

   >>> import numpy as np.

   Przykładowy wektor. Zacznijmy od stworzenia wektora o sześciu elementach typu całkowitego.

   >>> x = np.array([-2, -1, 0, 1, 2, 3])

   >>> type(x)

   <class ‘numpy.ndarray’>

   >>> x

   array([-2, -1, 0, 1, 2, 3])

   Sprawdź liczbę dla uwtorzonej tablicy:

   >>> x.dim

   Pole shape określa kształt tablicy.

   >>> x. shape

   Zadanie.

   Stwórz tablice 2 wymiarową 4 elementową. Sprawdź wymiar i kształt macierzy.

   Zmień kształt tablicy na tablicę 2 i 6 elementową.

   Zadanie.

   Stwórz tablicę zawiera ciąg arytmetyczny o zadanym przyroście.

   Stwórz tablicę zawiera ciąg arytmetyczny złożony z wartości z przedziału [a,b].

Zadanie.

Stwórz macierz: jednostkową, wypełnioną zerami i jedynkami.

Zadanie.

Stwórz macierz 4x2, a następnie dodaj do niej dowolny element.

Zadanie.

Stwórz 2 różne macierze, a następnie wykonaj na nich operacje: +, -, *, /, **, //, %.

2. Agregacja danych.

Funkcje i metody służące do agregacji d-elementowych wektorów:

Operacja	Funkcja
suma	numpy.sum(x)
iloczyn	numpy.prod(x)
Średnia arytmetyczna	numpy.mean(x)
Odchylenie standardowe	numpy.std(x,ddof=0)
Wariacja próbkowa	numpy.var(x,ddof=0)
Element największy	numpy.amax(x)
Element najmniejszy	numpy.amin(x)
mediana	numpy.median(x)

Zadanie.

Stwórz macierz o 150 wierszach i 4 kolumach:

>>> import seaborn as sns

>>> iris = np.array(sns.load_dataset(”iris”).iloc[:,0:4])

Dokonaj standaryzacji wszystkich zmiennych, tj. w każdej kolumnie od każdej obserwacji odejmij wartość średniej arytmetycznej i podziel ja przez odchylenie standardowe.

Zadanie.

Zaimplementuj algorytm sortowania szybkiego, wykorzystując funkcję/metodę partition().

Zadanie.

Napisz funkcję, która wywołując unique() (z odpowiednimi argumentami), wyznacza modę (dominantę) wartości w podanym wektorze.

3. **Ramki danych.**

   Załadowanie pakietów:

   >>> import numpy as np

   >>> import pandas as pd

   >>> import seaborn as sns

   >>> flights = sns.load_dataset(“flights”)

   >>> tips = sns.load_dataset(“tips”)

   >>> iris = sns.load_dataset(“iris”)

   Zadanie.

   Sprawdź rozmiar ramki danych flights, tips, iris. Wywołaj metody head(), tail() oraz info() na obiektach flights, tips, iris.

   Zadanie.

   Ramka danych flights zawiera kolumny year i month. Wygneruj na ich podstawie nową zmienną typu data i czas.

   Zadanie.

   Zastosuj wektor etykiet hierarchicznych do nazwania wierszy ramki danych flights. Wykorzystaj do tego celu dane zawarte w kolumnie year i month.

   Zadanie.

   Podziel ramkę danych irys na dwie rozłączne części: pierwsza 80% obserwacji, druga 20%.

   Zadanie.

   Dokonaj standaryzacji wszystkich zmiennych liczbowych w ramce danych iris.

   Zadanie.

   Wyznacz podstawowe statystyki próbkowe (średnia, odchylenie standardowe, kwartyle) dla liczby pasażerów (flights) w każdym roku z sobna. Przedstaw wyniki w taki sposób, by zagregowane wartości przechowywane były w kolumnach, a kolejne lata w wierszach.

Literatura:

1. Marek Gągolewski, Maciej Bartoszuk, Anna Cena, Przetwarzanie i analiza danych w języku Python, Wydawnictwo Naukowe PWN, Warszawa, 2017
2. Alberto Boschetti, Luca Massaron, Python. Podstawy nauki o danych. Wydanie II, Helion, 2017