Duomenų tyryba
Darbo su atviraisias duomenimis pavyzdžiai

Dr. Bronius Skūpas

Vilniaus licėjus

30.3. Duomenų tyrybos ir informacijos mokymo(si) turinys.

- 30.3.1. Duomenų rinkimas. Mokomasi rinkti duomenis iš įvairių šaltinių: atvirų duomenų (pavyzdžiui, https://data.gov.lt, https://data.europa.eu/euodp/lt ir kt.), interneto svetainių (pavyzdžiui, naudojant https://www.knime.com/knime-analytics-platform arba Python), apklausų (pavyzdžiui, Google Forms, Microsoft Forms), mikrovaldiklių jutiklių (pavyzdžiui, Arduino, Micro:bit, RaspberryPi), vaizdų (pavyzdžiui, nuotraukų, piešinių) analizės, programų ir programėlių žurnalų (ang. log file), geografinių informacinių sistemų ir pan.).
- 30.3.2. Didelių duomenų tyrinėjimas. Supažindinama su įvairių šaltinių duomenų sujungimu, įkėlimu į duomenų bazę, apdorojimu ir rezultatų išvedimu. Sprendžiamos skirtingo žymėjimo, duomenų matavimo vienetų suderinamumo problemos. Paaiškinami pagrindiniai užklausų, duomenų atrinkimo principai, rakto sąvoka. Naudojamasi skaičiuokle (įskaitant debesų kompiuterija paremtą), duomenų bazių valdymo sistema (pavyzdžiui, LibreOffice Base, MySQL, SQLite), programavimo kalbomis (pavyzdžiui, Python, JavaScript, PHP).

Duomenų šaltiniai

- atviri duomenys (pavyzdžiui, https://data.gov.lt, https://data.europa.eu/euodp/lt, https://open.vilnius.lt/
- interneto svetainių duomenys (pavyzdžiui, naudojant https://www.knime.com/knime-analytics-platform, Python, Word),
- apklausų rezultatai (pavyzdžiui, Google Forms, Microsoft Forms),
- mikrovaldiklių jutiklių duomenys (pavyzdžiui, Arduino, Micro:bit, RaspberryPi),
- vaizdų (pavyzdžiui, nuotraukų, piešinių) analizės,
- programų ir programėlių žurnalų (ang. log file),
- geografinių informacinių sistemų ir pan.

Informatikos bendrojo ugdymo programa

- Duomenų tyryba.
  - Ar tai nauja?
  - Ar visus paminėtus įrankius būtina išbandyti su mokiniais?
  - Ar galima mokant naudoti Excel?
  - Koks santykis tarp teorijos ir praktikos?
- Didieji ir atvirieji duomenys
  - Ar tikrai reiks dirbti su didžiaisiais duomenimis?

Pastebėjimai

- Duomenų šaltiniai gali būti naudojami ir ne automatiniu būdu. Duomenys gali būti atrenkami filtruojant, analizuojant diagramas, demonstruojant supratimą apie įvairius matavimo duomenis.
- Visai reali užduotis gali būti susijusi su skirtingų duomenų sulyginimu, pavyzdžiui, kritulių kiekio per vasarą palyginimas su užaugintu javų kiekiu. Duomenys gali būti pasiekiami iš skirtingų šaltinių – vieni iš atvirų duomenų lentelės, kiti iš diagramų ar ...
- Gana svarbu darbo su duomenimis efektyvumas. Tam padeda gebėjimas formuluoti užklausas

Atviri duomenys ir jų formatai

- CSV
  - Subtilumas – į Excel geriau įterpti į lakštą, o ne tiesiogiai atidaryti, nes skirtukai gali būti ne kabliataškiai, koduotė ne windows-1257.
  - Jei dirbti su Python - yra csv paketas, tačiau žymiai smagiau dirba pandas.
  - Patogiai importuojama į bet kokias duomenų bazes, duomenų analizės sistemas

- JSON

Atviri duomenys ir jų formatai

- JSON
  - Dažnai įvardijama kaip API. Rodoma svetainėse dinamiškai užkraunant duomenis.
  - Excel neturi galimybės atverti json, tačiau galima pasinaudoti keliais sprendimais:
    - VBA Excel makrokomandos https://medium.com/swlh/excel-vba-parse-json-easily-c2213f4d8e7a (Basic - dar viena programavimo kalba...)
    - Google Sheets makrokomandos (Javascript, galima automatizuoti duomenų rinkimą, papildant automatiškai kas kiek laiko paleidžiamais skriptais, dauguma mokinių turi Google paskyras)
    - PHP kalbos standartinės priemonės (dar viena kalba. Patogu kuriant svetaines)
    - Python – ko gero patogiausia (galima naudoti https://colab.research.google.com/).
      Jei dirbama su NŠA Python - rekomenduotina įvykdyti
      pip install pandas
      pip install openpyxl

Apps Script Temperaturu pavyzdys

Deploy

B

Files

macros.gs

makrokomandos.gs

Libraries

Services

Run Debug dabarLietuvoje Execution log

```
/** @OnlyCurrentDoc */

function dabarLietuvoje() {
  var response = UrlFetchApp.fetch("https://eismoinfo.lt/weather-conditions-service");
  var dataAll = JSON.parse(response.getContentText());
  // Logger.log(dataAll[0]);

  var spreadsheet = SpreadsheetApp.getActive();
  var sheet = spreadsheet.getActiveSheet();

  var row = 1;
  for (i in dataAll){
    Logger.log(dataAll[i]["id"] + " "+ dataAll[i]["irenginys"] + " " + dataAll[i]["oro_temperatura"]);
    sheet.getRange(row, 1).setValue(dataAll[i]["id"]);
    sheet.getRange(row, 2).setValue(dataAll[i]["irenginys"]);
    sheet.getRange(row, 3).setValue(dataAll[i]["oro_temperatura"]);
    row++;
  }
};
```

temperaturos.ipynb

Failas Redaguoti Rodinys Įterpti Vykdymo aplinka Įrankiai Pagalba Visi pakeitimai išsaugoti

Komentuoti Bendrinti

Failai

..
sample_data
data_file.csv
irenginiai.csv
irenginiai.xlsx

Diskas 84.49 GB laisvos vietos

+ Kodas + Tekstas

RAM
Diskas

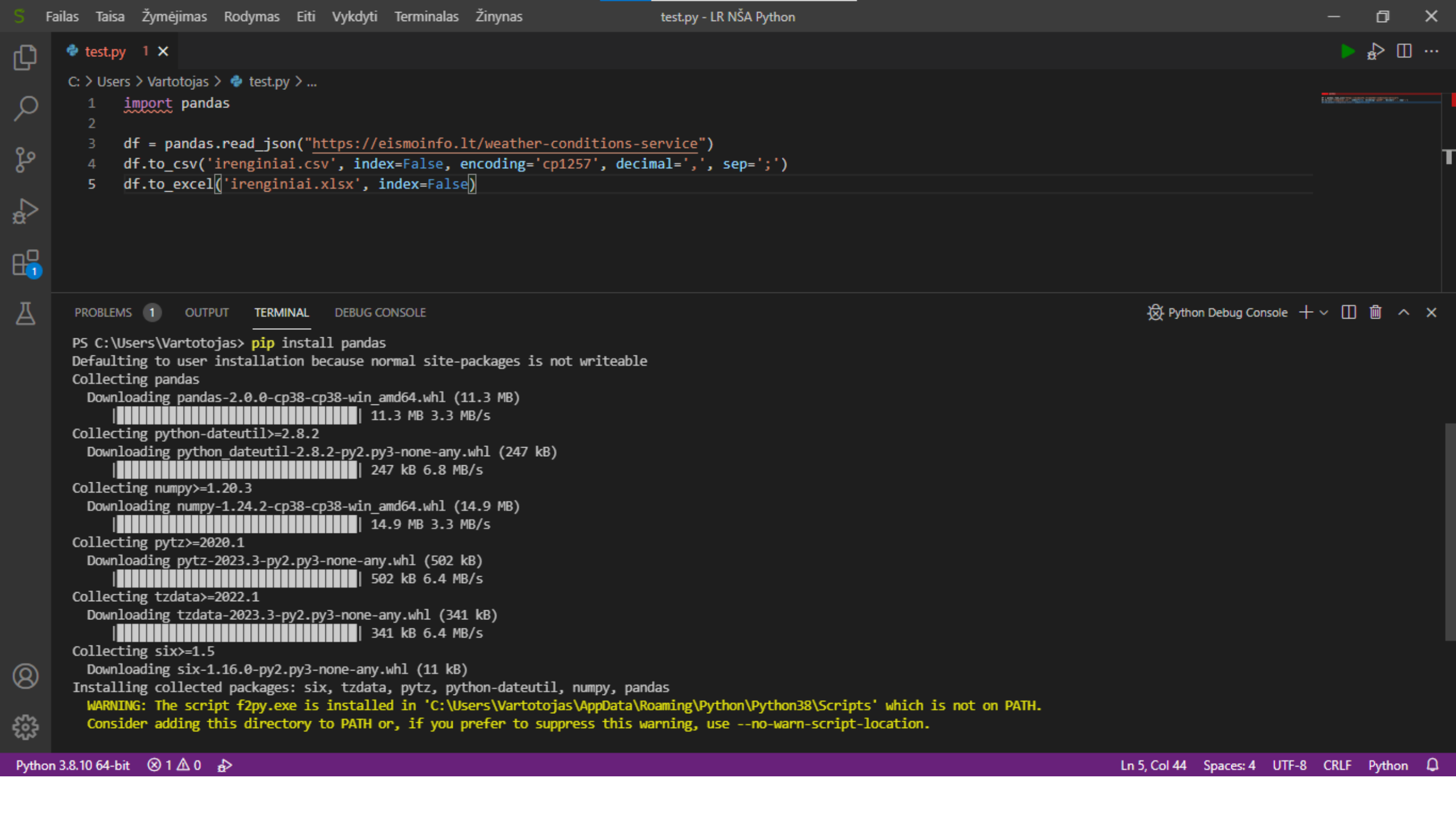
Excel

1 sek.

```
import pandas

df = pandas.read_json("https://eismoinfo.lt/weather-conditions-service")
df.to_csv('irenginiai.csv', index=False, encoding='cp1257', decimal=',', sep=';')
df.to_excel('irenginiai.xlsx', index=False)
```

1 sek. baigta 04:39


Failas Taisa Žymėjimas Rodymas Eiti Vykdyti Terminalas Žinynas
test.py - LR NŠA Python
test.py 1
C: > Users > Vartotojas > test.py > ...
1 import pandas
2
3 df = pandas.read_json("https://eismoinfo.lt/weather-conditions-service")
4 df.to_csv('irenginiai.csv', index=False, encoding='cp1257', decimal=',', sep=';')
5 df.to_excel('irenginiai.xlsx', index=False)
PROBLEMS 1 OUTPUT TERMINAL DEBUG CONSOLE
Python Debug Console
PS C:\Users\Vartotojas> pip install pandas
Defaulting to user installation because normal site-packages is not writeable
Collecting pandas
Downloading pandas-2.0.0-cp38-cp38-win_amd64.whl (11.3 MB)
11.3 MB 3.3 MB/s
Collecting python-dateutil>=2.8.2
Downloading python_dateutil-2.8.2-py2.py3-none-any.whl (247 kB)
247 kB 6.8 MB/s
Collecting numpy>=1.20.3
Downloading numpy-1.24.2-cp38-cp38-win_amd64.whl (14.9 MB)
14.9 MB 3.3 MB/s
Collecting pytz>=2020.1
Downloading pytz-2023.3-py2.py3-none-any.whl (502 kB)
502 kB 6.4 MB/s
Collecting tzdata>=2022.1
Downloading tzdata-2023.3-py2.py3-none-any.whl (341 kB)
341 kB 6.4 MB/s
Collecting six>=1.5
Downloading six-1.16.0-py2.py3-none-any.whl (11 kB)
Installing collected packages: six, tzdata, pytz, python-dateutil, numpy, pandas
WARNING: The script f2py.exe is installed in 'C:\Users\Vartotojas\AppData\Roaming\Python\Python38\Scripts' which is not on PATH.
Consider adding this directory to PATH or, if you prefer to suppress this warning, use --no-warn-script-location.
Python 3.8.10 64-bit
Ln 5, Col 44 Spaces: 4 UTF-8 CRLF Python

Atviri duomenys ir jų formatai

- HTML ir XML
  - Python BeautifulSoup4 paketas (Pavyzdys: https://github.com/justinas2314/tamo-homework-scraper)
    Prieš dirbant reiktų įsidiegti
    pip install beautifulsoup4
  - PHP verta pažiūrėti į http://sourceforge.net/projects/simplehtmldom/
  - Javascript labai lankstus ir modernus įrankis darbui su html ir xml.
    Yra nemažai bibliotekų: cheerio, jsdom. Verta paskaityti
    https://www.scrapingbee.com/blog/web-scraping-javascript/
  - Kodėl kartais nepanaudoti ir Word? Pavyzdžiui, lentelių tvarkymui.

Įrankiai duomenims tyrinėti

- Microsoft Office Excel, Google Sheets, LibreOffice Calc
- Duomenų bazė (pavyzdžiui, LibreOffice Base, MySQL, SQLite) Labai rekomenduoju https://sqlitebrowser.org/ Maža, greita, paprasta, labai populiari Android aplikacijose.
- Specializuoti įrankiai: Orange, Knime, Octave, Python su bibliotekomis...

- Labai gražus naujas https://lookerstudio.google.com/

LookerStudio privalumai

- Nemokama
- Daugybė duomenų šaltinių
- Duomenų vizualizavimas realiu laiku
- Google paskyra

- Trūkumas – tinklinė.
- Kartais sugriūna, „pakimba“ puslapis
- Ne visur intuityvu

temperaturos

Reset | Share | Edit

Untitled Page

Untitled Page

	surinkimo_data	oro_te...
1.	2023-04-11 01:50	7.8
2.	2023-04-11 01:35	7.7
3.	2023-04-11 01:20	7.9
4.	2023-04-11 01:05	8.5
5.	2023-04-11 00:50	9.2
6.	2023-04-11 00:35	9.5
7.	2023-04-11 00:20	9.3
8.	2023-04-11 00:05	9.8
9.	2023-04-10 23:50	10
10.	2023-04-10 23:35	9.1
11.	2023-04-10 23:20	8.9
12.	2023-04-10 23:05	9
13.	2023-04-10 22:50	9
14.	2023-04-10 22:35	8.9
15.	2023-04-10 22:20	8.9
16.	2023-04-10 22:05	9.8
17.	2023-04-10 21:50	9.9
18.	2023-04-10 21:35	10.5
19.	2023-04-10 21:20	11.1
20.	2023-04-10 21:05	11.1
21.	2023-04-10 20:50	11.3
22.	2023-04-10 20:35	11.5
23.	2023-04-10 20:20	11.9
24.	2023-04-10 20:05	12.3
25.	2023-04-10 19:50	13