

UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Miha Zidar

Dostop do podatkov Svetovne banke v orodju Orange

DIPLOMSKO DELO
UNIVERZITETNI ŠTUDIJSKI PROGRAM RAČUNALNIŠTVO
IN INFORMATIKA

MENTOR: prof. dr. Blaž Zupan

Ljubljana, 2016

Rezultati diplomskega dela so intelektualna lastnina avtorja in Fakultete za računalništvo in informatiko Univerze v Ljubljani. Za objavlanje ali izkoriščanje rezultatov diplomskega dela je potrebno pisno soglasje avtorja, Fakultete za računalništvo in informatiko ter mentorja.

Besedilo je oblikovano z urejevalnikom besedil \LaTeX .

Namesto te strani **vstavite** original izdane teme diplomskega dela s podpisom mentorja in dekana ter žigom fakultete, ki ga diplomant dvigne v študentskem referatu, preden odda izdelek v vezavo!

IZJAVA O AVTORSTVU DIPLOMSKEGA DELA

Spodaj podpisani Miha Zidar, z vpisno številko **63060317**, sem avtor diplomskega dela z naslovom:

Dostop do podatkov Svetovne banke v orodju Orange

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal samostojno pod mentorstvom prof. dr. Blaža Zupana,
- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela
- soglašam z javno objavo elektronske oblike diplomskega dela v zbirki "Dela FRI".

V Ljubljani, dne 24. avgust 2016

Podpis avtorja:

Zahvalil bi se mentorju, prof. dr. Blažu Zupanu in članom laboratorija za bioinformatiko za pomoč in usmerjanje med izdelavo diplomskega dela. Prav tako bi se zahvalil svojim staršem, prijateljem in svojemu partnerju za spodbudo.

Kazalo

Povzetek

Abstract

1	Uvod	1
1.1	Motivacija	2
1.2	Cilji in struktura diplomske naloge	2
2	Podatkovne zbirke Svetovne Banke	3
2.1	Podatki indikatorjev razvoja držav	4
2.2	Podatki podnebnih meritev	11
2.3	Težave pri dostopu	12
3	Knjižnica in gradniki za Orange	15
3.1	Knjižnica simple_wbd	16
3.2	api_wrapper	18
3.3	Grafični vmesnik	19
4	Primeri uporabe	23
4.1	Napoved temperature s pomočjo CO2 emisij	23
5	Sklepne ugotovitve	25

Seznam uporabljenih kratic

kratica	angleško	slovensko
API	Application Programming Interface	programski vmesnik
REST	Representational State Transfer	predstavitvena arhitektura za prenos podatkov
XML	Extensible Markup Language	razširljivi označevalni jezik
JSON	JavaScript Object Notation	Javascript objektna nota- cija

Povzetek

Naslov: Dostop do podatkov Svetovne banke v orodju Orange

Avtor: Miha Zidar

Program Orange je orodje za podatkovno rudarjenje, v katerem lahko za namene analiz uporabimo različne podatkovne vire. Sam program Orange vsebuje predpripravljene zbirke podatkov, dodatne zbirke podatkov si lahko pripravi in uvozi tudi uporabnik sam, ali pa uporabi katerega od že obstoječih dodatkov za uvoz podatkov. Za namen diplomske naloge smo izdelali dodatek Orange data sets, s katerim je mogoče dostopati do podatkov s programskega vmesnika Svetovne banke. Trenutno Svetovna banka omogoča uporabo štirih različnih programskih vmesnikov: gospodarski indikatorji, projekti Svetovne banke, finančni podatke in klimatski podatki. Dodatek Orange data sets vsebuje dva gradnika, ki sta namenjena lažjemu branju in uporabi podatkov indikatorjev in klimatskih podatkov. S tem bo uporabnikom programa Orange omogočena enostavnejša uporaba velikega števila podatkov iz omenjenih dveh programskih vmesnikov.

Ključne besede: Podatkovno rudarjenje, programski vmesnik, Svetovna banka, gospodarski indikatorji, podnebni podatki, Orange.

Abstract

Title: Access to World bank data with Orange

Author: Miha Zidar

TODO: Orange is an open source data-mining software, capable of using multiple sources for data analysis. There are a few test data sample already present in Orange, and the user can import their own data sets with the use of one of Orange input widgets. For this thesis we created a new widget "Orange data sets" for accessing free data from World bank application program interface (API). The World bank exposes four different data APIs; indicator, project, finance and climate. Our Orange data sets widget will be able to read data from the indicators and climate APIs.

Key words: Data mining, API, World bank, indicators, climate, Orange.

Poglavje 1

Uvod

Na svetovnem spletu je dosegljivih vedno več prosto dostopnih programskih vmesnikov (ang. application programming interface). Ti vmesniki omogočajo dostop do zelo raznolikih zbirk podatkov. Nekaj primerov prosto dostopnih podatkovnih zbirk je seznam stopnje ogroženosti živali po državah ¹, podatki meritev in slike vesolja agencije NASA ², seznam knjig z ocenami in povezavami med uporabniki ³, zgodovina meteoroloških meritev ⁴, razni indikatorji stopenj razvoja držav ⁵.

Programski vmesniki so oblikovani tako, da je omogočena raznolika uporaba podatkov iz podatkovnih zbirk. To pa ima tudi slabost, ki je v tem, da je podatke potrebno predhodno obdelati za vsak namen posebej. Tako bi na primer moral vsak uporabnik programa Orange podatke predhodno pretvoriti v obliko, primerno za njegovo konkretno analizo.

¹<http://apiv3.iucnredlist.org/api/v3/docs>

²<https://api.nasa.gov/>

³<https://www.goodreads.com/api>

⁴<http://climatedataapi.worldbank.org/>

⁵<http://api.worldbank.org/>

1.1 Motivacija

Povezava programskega vmesnika za dostop do podatkov in orodja za analizo podatkov je pogosto prezapletena za navadnega uporabnika. Z dodatkom Orange data sets želimo podatke programskega vmesnika Svetovne banke spraviti v obliko, primerno za nadaljno uporabo v orodju Orange. Ta dodatek bi pomagal združiti programe za obdelavo podatkov in prosto dostopne zbirke podatkov. S tem dobimo enostavnejši dostop do podatkov iz prek 16.000 indikatorjev in številnih klimatskih meritev, s čimer bomo lažje analizirali in iskali morebitne zakonitosti v podatkih. Če bi imeli en sam ustrezen dodatek za dostop do podatkov programskega vmesnika Svetovne banke, bi poenostavili posodabljanje in vzdrževanje kode v primeru sprememb programskega vmesnika za vse uporabnike istega orodja hkrati. S tem odpravimo potrebo, da bi moral vsak uporabnik sam skrbeti za uskladitvene posodobitve.

1.2 Cilji in struktura diplomske naloge

Cilj diplomske naloge je izdelati knjižnico za uporabo programskega vmesnika Svetovne banke ter izdelati dodatek za program Orange, ki s pomočjo omenjene knjižnice omogoča uporabniku dostop do podatkov Svetovne banke preko grafičnega vmesnika.

V diplomski nalogi najprej predstavimo spletna vira indikatorjev držav sveta in meritev podnebnih podatkov Svetovne banke, ter opišemo delovanje njunih programskih vmesnikov. Nato podrobneje opišemo našo implementacijo knjižnice za dostop do programskega vmesnika Svetovne banke in gradnikov za program Orange, ki to knjižnico uporabljajo. V nadaljevanju prikažemo še nekaj praktičnih primerov uporabe dodatka Orange data sets. Na koncu še popisemo opravljeno delo, navedemo vire kode in omenimo možne načine za izboljšavo ali nadgradnjo našega dodatka.

Poglavje 2

Podatkovne zbirke Svetovne Banke

Pri diplomski nalogi smo se osredotočili na dva programska vmesnika za dostop podatkov Svetovne banke, to sta “ClimateAPI” s katerim dostopamo do podatkovne zbirke meteoroloških meritev in “IndicatorAPI” s katerim dostopamo do zbirke podatkov raznih indikatorjev stopenj razvoja držav. Za uporabo podatkovne zbirke Svetovne banke smo se odločili, ker združuje in na enovit način predstavi podatke iz večih različnih virov. Podatkovni viri za indikatorje stopnje razvoja držav so:

- Svetovni indikatorji razvoja [1]
- Globalni finančni razvoj [2]
- Afriški indikatorji razvoja [3]
- Poslovanje [4],
- Podjetniške raziskave [5],
- Razvojni cilji [6],
- Statistike izobraževanja [7],
- Statistike spolov [8],

- Statistike zdravja in prehranjevanja [9],
- Rezultati meritev IDA [10].

Podatkovni vir zbirke podnebnih meritev pa je osnovan na podatkih oddelka za podnebne raziskave (ang. Climatic Research Unit) [11].

Svetovna banka omogoča dostop do podatkov preko programskega vmesnika REST, ki ponuja veliko možnosti za iskanje in presejanje rezultatov. Pri vsaki poizvedbi REST lahko določimo željeno obliko odgovora. Za poizvedbe o informacijah indikatorjev sta na voljo obliki XML in JSON. Programski vmesnik meteoroloških meritev pa ponuja samo obliko JSON. Za konsistentnost in lažjo berljivost smo na obeh programskih vmesnikih uporabili obliko JSON. To na programskem vmesniku indikatorjev dosežemo take da nastavimo parameter `GET format` na vrednost `json`.

2.1 Podatki indikatorjev razvoja držav

Programski vmesnik indikatorjev razvoja držav Svetovne banke omogoča dostop do podatkov preko 16.000 raznih indikatorjev. Podatki indikatorjev so merjeni mesečnem, četrtletnem ali letnem intervalu. Začetek meritev podatkov posameznega indikatorja je odvisna od vira podatkov. Najstareši podatki segajo do leta 1960. Poleg podatkov indikatorjev nam ta programski vmesnik omogoča tudi dostop do večine metapodatkov s katerimi lahko presejamo in natančneje določimo našo poizvedbo. Seznami metapodatkov so:

- viri podatkov in njihovi opisi (ang. Catalog Source Queries ¹),
- seznam držav, skupin držav in regij z identifikatorji (ang. Country Queries ²),
- razdelitev višin dohodkov z identifikatorji (ang. Income Level Queries ³),

¹<http://api.worldbank.org/sources?format=json>

²<http://api.worldbank.org/countries?format=json>

³<http://api.worldbank.org/incomeLevels?format=json>

- seznam indikatorjev (ang. Indicator Queries ⁴),
- seznam tipov posojil (ang. Lending Type Queries ⁵),
- seznam tem (ang. Topics ⁶).

Za pridobitev podatkov indikatorjev potrebujemo metapodatke o indikatorjih in državah. Primere teh metapodatkov si bomo podrobneje pogledali v nadaljevanju.

Vsi odgovori na veljavne poizvedbe po podatkih in metapodatkih, ki so na voljo s programskim vmesnikom indikatorjev razvoja, imajo enako osnovno obliko. Poizvedbe vračajo seznam z dvema elementoma, kjer je ima prvi element informacije o količini podatkov in trenutnem izboru podatkov, drugi element pa vsebuje seznam izbranih podatkov (Primer 1). Privzeta vrednost števila elementov na stran je 50, kar lahko spremenimo tako da poizvedbi nastavimo parameter GET `per_page` na poljubno vrednost. Če želimo pridobiti podatke z večih strani, moramo za vsako stran poslati novo poizvedbo v kateri podamo željeno stran s parametrom GET `page`, . Veljavne poizvedbe, ki ne vračajo podatkov, imajo vrednost drugega elementa osnovnega seznama `null`. Za neveljavne poizvedbe, pa programski vmesnik vraca seznam z enim elementom, ki vsebuje podatke o napaki poizvedbe (Primer 2).

2.1.1 Opis seznama indikatorjev

Programski vmesnik Svetovne banke za indikatorje razvoja nam ponuja seznam vseh indikatorjev z imeni, opisi, kodami in drugimi metapodatki (Primer 4). Programski vmesnik nam tudi omogoča dostop do podatkov posameznega indikatorja določenega s kodo in presejanje seznama indikatorjev glede na vir podatkov 3. V našem programu smo uporabili le poizvedbo za celotn seznam indikatorjev, da smo omogočili iskanje in presejanje po vseh poljih indikatorjev.

⁴<http://api.worldbank.org/indicators?format=json>

⁵<http://api.worldbank.org/lendingTypes?format=json>

⁶<http://api.worldbank.org/topics>

```
1  [  
2      {  
3          "page": 1,  
4          "pages": 137,  
5          "per_page": "50",  
6          "total": 6831  
7      },  
8      [  
9          <podatki>,  
10         ...  
11     ]  
12 ]
```

Primer 1: Osnovna oblika odgovora programskega vmesnika Svetovne banke, za veljavno poizvedbo indikatorjev.

```
1  [  
2      {  
3          "message": [  
4              {  
5                  "id": "120",  
6                  "key": "Parameter 'country' has an invalid value",  
7                  "value": "The provided parameter value is not valid"  
8              }  
9          ]  
10     }  
11 ]
```

Primer 2: Osnovna oblika odgovora programskega vmesnika Svetovne banke, za neveljavne poizvedbe.

```
1  http://api.worldbank.org/indicators?format=json  
2  http://api.worldbank.org/indicators?format=json&source=5  
3  http://api.worldbank.org/indicators/A10i?format=json
```

Primer 3: Primeri poizvedb po seznamu indikatorjev. 1) seznam vseh indikatorjev, 2) seznam indikatorjev glede na vir podatkov, 3) podatki indikatorja "A10i"

```
1  {
2    "id": "1.0.HCount.2.5usd",
3    "name": "Poverty Headcount (\$2.50 a day)",
4    "source": {
5      "id": "37",
6      "value": "LAC Equity Lab"
7    },
8    "sourceNote": "The poverty headcount index measures the
9                  proportion of the population with daily per
10                 capita income (in 2005 PPP) below the poverty
11                 line.",
12    "sourceOrganization": "LAC Equity Lab tabulations of SEDLAC
13                          (CEDLAS and the World Bank).",
14    "topics": [
15      {
16        "id": "11",
17        "value": "Poverty "
18      }
19    ]
20 }
```

Primer 4: Podatki indikatorja stopnja revščine pri dohodku 2,5 dolarja na dan.

2.1.2 Opis seznama držav

Seznam držav na programskem vmesniku Svetovne banke vsebuje podatke o imenih, opisih, ISO-3166-1 alpha kodah, regijah in druge metapodatke (Primer 6). Programski vmesnik nam tudi omogoča presejanje seznama držav po kodi države, regiji, visini dohodka, in tipu posojil (Primer 5)

```
1 http://api.worldbank.org/countries?format=json
2 http://api.worldbank.org/countries/svn?format=json
3 http://api.worldbank.org/countries?format=json&incomeLevel=HIC&region←
  =ECS
```

Primer 5: Primeri poizvedb po seznamu držav. 1) seznam vseh držav, 2) podatki ene države, 3) seznam držav v Evropi in Osrednji Aziji, z visoko višino dohodka.

Ta seznam ne vsebuje zgolj samo držav, ampak tudi regije in skupine držav, združenih glede na različne kriterije (višine dohodka, velikost, stopnja razvoja). Poleg tega zgornji seznam vsebuje tudi nekatere izjeme kot je trenutno Kosovo. V nadaljevanju bomo za vse našteje tipe lokacijskih podatkov uporabljali besedo “države”.

2.1.3 Dostop do podatkov indikatorjev

Za dostop do podatkov posameznega indikatorja, potrebujemo kodo indikatorja s seznama vseh indikatorjev in kodo ene ali večih držav. Namesto kode ene ali večih držav lahko uporabimo tudi ključno besedo “all”, ki označuje vse kode držav. Pri večjih količinah podatkov, lahko z dodatnimi parametri določimo število podatkov na stran, in željeno stran podatkov. Primer 7 prikazuje osnovno obliko poizvedbe, kjer so:

country s podpičjem ločen seznam kod izbranih držav, ki jih preberemo iz polja “id” ali “iso2Code”, ki sta prikazana v Primeru 6, ali pa ključna beseda “all”,

```
1  {
2    "id": "ABW",
3    "iso2Code": "AW",
4    "name": "Aruba",
5    "region": {
6      "id": "LCN",
7      "value": "Latin America & Caribbean "
8    },
9    "adminregion": {
10     "id": "",
11     "value": ""
12   },
13   "incomeLevel": {
14     "id": "HIC",
15     "value": "High income"
16   },
17   "lendingType": {
18     "id": "LNX",
19     "value": "Not classified"
20   },
21   "capitalCity": "Oranjestad",
22   "longitude": "-70.0167",
23   "latitude": "12.5167"
24 },
```

Primer 6: Izsek podatkov veljavne poizvedbe držav.

parametri Dodatni parametri GET

MRV Stevilška vrednost, ki določi maksimalno število zadnjih meritev, ki jih programski vmesnik vrne. Ko uporabljamo polje `mrv` bo programski vmesnik izpustil ničelne vrednosti za obdobja v katerih ni meritev.

date Polje obilke 'leto' ali 'leto:leto' ki omeji rezultate poizvedbe na določeno leto ali interval med določenimi leti.

Primer 7: Osnovna oblika poizvedbe za podatke enega indikatorja.

Slabost programskega vmesnika indikatorjev Svetovne banke je v tem, da ne moremo z eno poizvedbo dostopati do podatkov večih indikatorjev hkrati.

```
1  {  
2    "indicator": {  
3      "id": "SP.POP.TOTL",  
4      "value": "Population, total"  
5    },  
6    "country": {  
7      "id": "IL",  
8      "value": "Israel"  
9    },  
10   "value": "6289000",  
11   "decimal": "0",  
12   "date": "2000"  
13 }
```

Primer 8: Podatki za indikator SP.POP.TOTL (populacija države) za Izrael leta 2000.

2.2 Podatki podnebnih meritev

Programski vmesnik Svetovne banke za podnebne podatke omogoča dostop do podatkov napovednih modelov in zgodovinskih meritev meteoroloških postaj. V tej diplomski nalogi smo se odločili uporabiti samo podatke zgodovinskih meritev, saj si s temi podatki lahko uporabnik programa Orange sam sestavi svoje napovedne modele.

Za razliko od uporabe programskega vmesnika indikatorjev, lahko pri tem programskem vmesniku uporabljamo veljavne ISO 3166-1 alpha-2 ali ISO 3166-1 alpha-3 kode držav, ali pa številski identifikator (TODO link) vodotočnega območja.

Ta programski vmesnik nam omogoča dostop do podatkov o povprečnih temperaturah in padavinah v časovnih obdobjih enega leta, desetletja ali pa nam omogoča dostop do mesečnih povprečij skozi vsa leta meritev.

2.2.1 Dostop do podatkov podnebnih meritev

Dostop do podatkov podnebnih meritev je mogoč na naslovu <http://climatedataapi.worldbank.org/climateweb/rest/>

v1/**loc_type**/cru/**data_type**/**interval**/**location**

kjer je:

loc_type vrsta identifikatorja območja (“basin” za vodotocno območje, “country” za države),

data_type vrsta meritev (“pr” za padavine, “tas” za temperature),

interval vrsta meritvenega obdobja (“month” za mesečno, “year” za letno in “decade” za desetletno),

location identifikator države ali vodotočnega območja.

Izsek podatkov (primer za mesečno povprečje količine padavin za Slovenijo): <http://climatedataapi.worldbank.org/climateweb/rest/v1/country/cru/pr/month/SVI>

```
1 {  
2     "month": 0,  
3     "data": 68.93643  
4 },  
5 {  
6     "month": 1,  
7     "data": 64.23069  
8 },  
9 {  
10    "month": 2,  
11    "data": 81.098724  
12 },  
13 ...
```

2.3 Težave pri dostopu

Težave pri uporabi SB API-ja lahko razdelimo v dve skupini. Prvo skupino sestavljajo težave pri dokumentaciji in z manjkajočimi podatki, drugo pa napake v samih pridobljenih podatkih.

Težave prve skupine:

- ob posodabljanju spletne strani SB se izgubijo posamezne povezave do primerov, dokumentacije in opisov API-ja,

- nepopolna dokumentacija:

polje za datum je opisano, vendar ni dokumentirano, kakšne so vse možne vrednosti,

delovanje polj za obdobje (date), zadnje vrednosti (mrv) in za manjkajoče vrednosti (gapfill) ni ustrezno opisano.

Težave druge skupine:

- manjkajoči identifikatoriji za polja na posameznih indikatorjih (primer je manjkajoča vrednost v polju id države),
- datum vsebuje naključne vrednosti (“last known value” “2001 - 2015” “2040”
- zgornja meja števila izbranih lokacij na 250 ni navedena,
- nemogoče je ugotoviti pogostost vzorčenja indikatorja (frequency).

Poglavje 3

Knjižnica in gradniki za Orange

V okviru diplomske naloge smo razvili tri ločene komponente za programerje in končne uporabnike programa Orange.

Prva komponenta je programska knjižnica `simple_wbd`, ki omogoča enostaven dostop do programskega vmesnika indikatorjev in klimatskih podatkov Svetovne banke. Ta knjižnica je narejena s čim manj odvisnosti in je namenjena splošni uporabi v Python programih. Poudarka pri zasnovi knjižnice `simple_wbd` sta predvsem enostavnost razširitve in zanesljivost. Ta cilja dosežemo z mehanizmom za vključevanje lastne kode v komponente knjižnice in mehanizmi za popravljanje ali odstranjevanje pokvarjenih podatkov.

Drugi sestavni del je razširitev knjižnice `simple_wbd` s funkcionalnostmi, potrebnimi za lažje delo v programu Orange. To predvsem zavzema pretvorbo pridobljenih podatkov v Orange in numpy tabele. Ta sklop je namenjen skriptnemu delu s programom Orange in je dostopen kot `api_wrapper` Python modul.

Tretji sestavni del je grafični vmesnik za uporabo `api_wrapper` modula. Namen grafičnega vmesnika je omogočiti ne-programerjem dostop do podatkov programskega vmesnika Svetovne banke znotraj programa Orange za namen obdelave, analize in iskanja zakonitosti med podatki.

3.1 Knjižnica `simple_wbd`

Knjižnica `simple_wbd` programerjem olajša dostop do podatkov programskega vmesnika Svetovne banke. Glavna lastnost te knjižnice je združevanje večjega števila zahtev po podatkih in enostavna predstavitev dobljenih podatkov. Druga lastnost je pretvorba podatkov iz več dimenzij v dvo-dimenzionalno polje, primerno za uporabo v programu Orange. Glavna vmesnika te knjižnice sta `IndicatorAPI` in `ClimateAPI`. Prvi omo”goca pridobivanje podatkov iz programskega vmesnika indikatorjev Svetovne banke, drugi pa s programskega vmesnika podnebnih meritev.

3.1.1 Pomocnik `IndicatorAPI`

`IndicatorAPI` je razred namenjen pridobivanju podatkov indikatorjev razvoja. Ker ima programski vmesnik Svetovne banke omejitve koliko podatkov lahko prenesemo z eno poizvedbo, nam ta razred združuje rezultate vseh poizvedb, ki so potrebne za prenos celotne zahteve. To poskrbi tako da se po prvi poizvedbi sprehodi čez število preostalih strani (Primer 1) ki so na voljo. Za razliko od obstoječih knjižnic^{1 2} za delo z programskim vmesnikom Svetovne banke, katerih cilj je čim bolj natančno predstaviti programski vmesnik, je cilj nase knjižnice le poenostaviti poizvedbe. V ta namen smo s to knjižnico razsirili programski vmesnik da lahko z enim klicem funkcije prenesemo podatke tudi več kot le enega indikatorja.

Poleg tega da skrbi za prenos vseh strani podatkov, tudi beleži število narejenih in število potrebnih poizvedb za celoten prenos. Do teh števil lahko dostopamo z drugih niti in jih uporabimo za prikaz napredka in preostalega časa do prenosa celotne zahteve.

Glavne metode ki jih ponuja `IndicatorAPI` razred so:

`get_indicators` vrne seznam vseh možnih indikatorjev z imeni, opisi in identifikatorji,

¹<https://pypi.python.org/pypi/wbdata>

²<https://pypi.python.org/pypi/wbpy/2.0.1>

get_countries vrne seznam držav in regij z kodami in metapodatki,

get_dataset vrne razred (IndicatorDataset) ki vsebuje vse podatke z poizvedbe in metode za oblikovanje predstavitev podatkov: `api_responses`, `as_list`, `as_dict`.

Implementirali lastni caching: - obstoječe knjižnice (`vcrapy`, `requests-cache`) za hranjenje narejenih poizvedb so se izkazale za zelo počasne v primeru ko delamo z večjim številom podatkov. - nasa resitev se zanaša na edinstvenost URL naslovov - za vsak url si shranimo odgovor in če je mlajši od enega tedna, isti odgovor uporabim pri vsaki poizvedbi za določeni URL.

mehanizmem za odpravo napak: - pridobiti podatke o državi za posamezni indikator: - ob manjkajočih id-jih poskusamo določiti id iz imena - ob manjkajočih imenih poskusamo dobiti ime iz id-ja - če ni nobene informacije ta podatek ignoriramo.

ko zelimo dobiti polje v obliki časovne vrste: - pridobivanje datuma iz polja 'date' - ob neveljavnih stringih probamo upoštevati le začetni del. - za text za obdobje '2002 - 2006' bomo uporabili le datum 2002 - neveljavne stringe kot so "most rescent value" ignoriramo.

Razred IndicatorDataset

Razred IndikatorDataset je osnovni razred v katerem dobimo zahtevane podatke indikatorjev. Ta razred vsebuje vse potrebne metode in podatke za predstavitev rezultatov programskega vmesnika, na dva glavna načina; kot slovar slovarjev in dvo dimenzionalen seznam. Poleg omenjenih načinov predstavitev podatkov lahko dostopamo tudi do neobdelanih podatkov prejetih z programskega vmesnika za vsako poizvedbo posebej.

Posamezne vrednosti teh podatkov so določene z državo, časovno komponento, in kodo indikatorja. Te podatke lahko predstavimo na dva glavna načina:

- kot gnezdeni slovar, kjer je na prvem nivoju ime indikatorja, na drugem država, in na tretjem nivoju časovna komponenta.

- Kot dvo-dimenzionalno polje, kjer imamo v vrsticah eno oznako, v stolpcih pa kartezicni produkt ostalih dveh. ponujene možnosti so: - vrstice = država, stolpci = cas x indikator - vrstice = cas, stolpci = država x indikator
Indicator

3.1.2 Pomcnik ClimateAPI

IndicatorAPI je

api dovoli le podatke za en tip za eno vrsto obdobja in eno državo hkrati. mi naredim kartezicni produkt med vsemi temi zgradimo vse url-je in naredimo vse potrebne poizvedbe za pridobitev podatkov.

Razred IndicatorDataset

Isto kot pri indicator apiju ko zelimo dobiti polje v obliki casovne vrste: - pridobivanje datuma iz polja 'date' - ob neveljavnih stringih probamo upoštevati le zacetni del. - za text za obdobje '2002 - 2006' bomo uporabili le datum 2002 - neveljavne stringe kot so "most rescent value" ignoriramo.

as_dict

glede na poizvedbo dobimo tu gnezden slovar s polji

država / tip podatkov / tip casovnega obdobja / vrednost casovnega obdobja / vrednost

as_list

3.2 api_wrapper

razširitev simple wbd vmesnikov z dedovanjem pravega dataset razreda.

```
class ClimateDataset(simple_wbd.ClimateDataset):
```

```
    def as_numpy(self):
        raise NotImplemented()
```

```
def as_orange_table(self):
    raise NotImplemented()

class ClimateAPI(simple_wbd.ClimateAPI):

    def __init__(self):
        super().__init__(ClimateDataset)
```

- razsiri `as_list` v `as_numpy_array` ki tudi odstrani vse stolpce ki nimajo veljavne vrednosti.

- doda `as_orange_table` ki `numpy array` spremeni v `orange tabelo`. - za indikator `api` doda se metapodatke držav ko ne prikazujemo v obliki časovne vrste.

`api wrapper` je tudi zelo uporaben za skriptno uporaba programa `Orange` (referenca <http://www.jmlr.org/papers/volume14/demsar13a/demsar13a.pdf>)

in tukaj si lahko vsak programer sam oblikuje podatke v katerokoli željeno obliko.

3.3 Graficni vmesnik

nova skupina `data sets` - v katero je mogoče dodati nove gradnike za druge programske vmesnike.

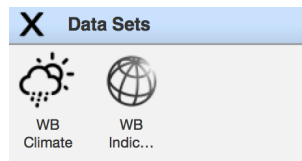
2 gradnika - `wb indicators` in `wb climate`

- lažja uporaba `apija` - večja preglednost

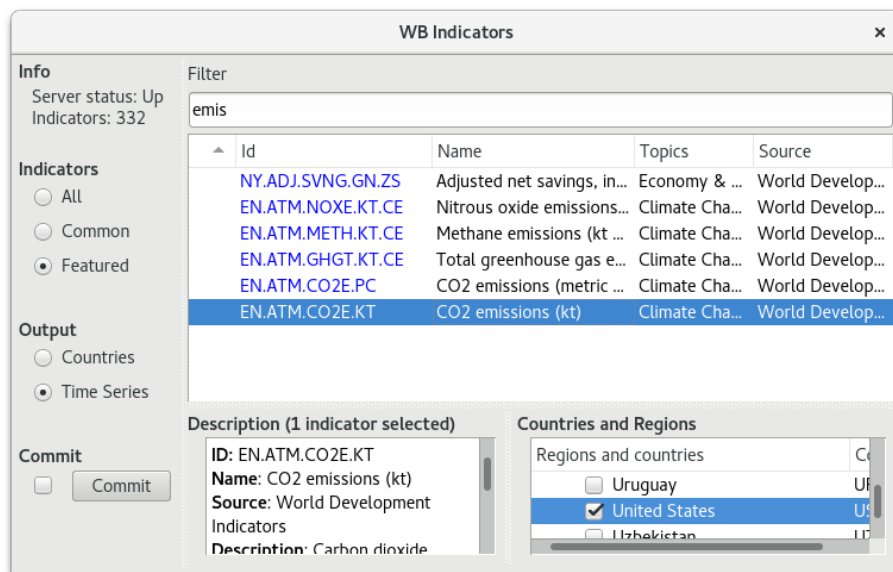
za oba gradnika smo razvili in uporabili `base class` - skupni podatki

- razvili smo tudi gradnik za gnezden prikaz urejenih slovarjev. ta se uporablja za prikaz držav po kontinentih v `climate` gradniku, in za prikaz držav in skupin držav in drugih agregatov v gradniku `indicators`.

za te gradnike smo tudi napisali enotske teste.



Slika 3.1: Skupina gradnikov data sets



Slika 3.2: Odločitveno drevo za izbor primerne metode.

3.3.1 wb indicators gradnik

za sestavo smo si pomagali z gradniki Orange.gui

elementi gradnika

2 osnovna filtra:

- izbor indikatorjev ki se pokazujejo v seznamu all/common/featured ki ustreza seznamu indikatorjev na strani: all - vse (tudi nekateri ki jih na strani ni nastetih) common - <http://data.worldbank.org/indicator?tab=all> featured - <http://data.worldbank.org/indicator?tab=featured> - text filter

gradnik ima sistem za prikaz (progress bar?)

moznost izbire tipa izhoda (countries in time series - opis)

The image shows a software window titled "WB Climate" with a close button (X) in the top right corner. The window is divided into two main sections: "Info" on the left and "Countries" on the right.

Info Section:

- Server status: Up
- Selected countries: 1
- Average intervals:**
 - ☐ Month
 - ☒ Year
 - ☐ Decade
- Data Types**
 - ☒ Temperature
 - ☐ Precipitation
- Output**
 - ☐ Countries
 - ☒ Time Series
 - ☐ Use Country names
- Commit**
 - ☐

Countries Section:

A table titled "Regions and countries" with a "Code" column. It lists various countries and regions, each with a checkbox and a corresponding code. The "United States of America" is selected.

Regions and countries	Code
<input type="checkbox"/> Martinique	MTQ
<input type="checkbox"/> Mexico	MEX
<input type="checkbox"/> Montserrat	MSR
<input type="checkbox"/> Nicaragua	NIC
<input type="checkbox"/> Panama	PAN
<input type="checkbox"/> Puerto Rico	PRI
<input type="checkbox"/> Saint Barthélemy	BLM
<input type="checkbox"/> Saint Kitts and Nevis	KNA
<input type="checkbox"/> Saint Lucia	LCA
<input type="checkbox"/> Saint Martin	MAF
<input type="checkbox"/> Saint Pierre and Mique...	SPM
<input type="checkbox"/> Saint Vincent and the ...	VCT
<input type="checkbox"/> Sint Maarten	SXM
<input type="checkbox"/> Trinidad and Tobago	TTO
<input type="checkbox"/> Turks and Caicos Islan...	TCA
<input type="checkbox"/> United States Virgin Is...	VIR
<input checked="" type="checkbox"/> United States of Ameri...	USA
▶ <input type="checkbox"/> Oceania	
▶ <input type="checkbox"/> SouthAmerica	

Slika 3.3: Odločitveno drevo za izbor primerne metode.

3.3.2 wb climate gradnik

dovoli izbiro posameznih drzav

moznost izbire tipa izhoda (countries in time series - opis) za razliko od indikator apija tukaj nismo dodali metapodatkov drzav

Poglavje 4

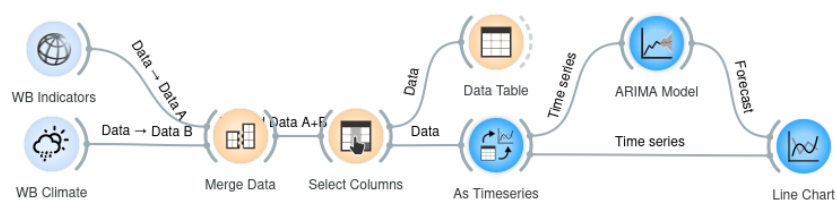
Primeri uporabe

uporaba: slike orange uporabe

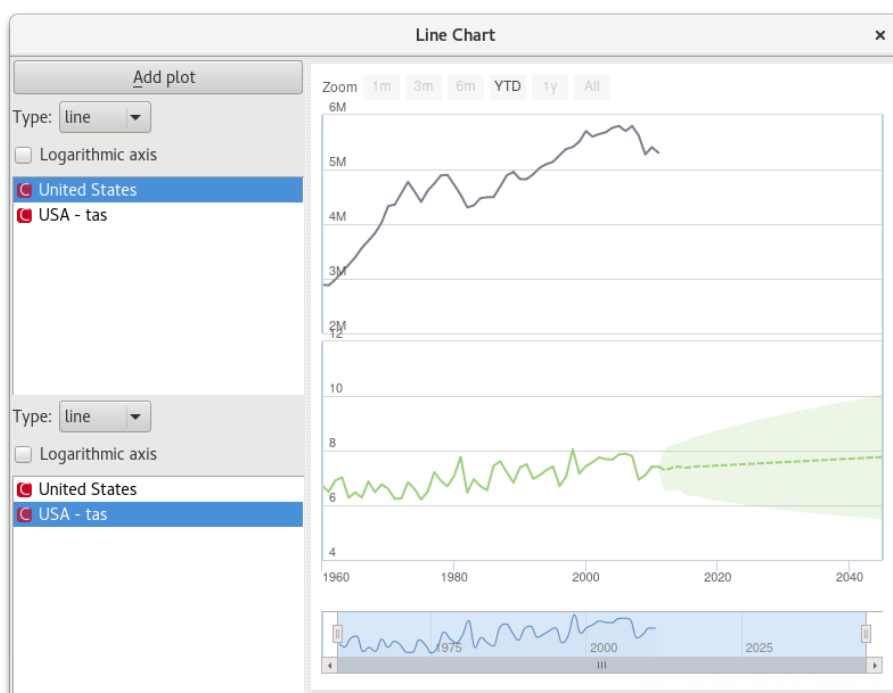
4.1 Napoved temperature s pomocjo CO2 emisij

postavitev kaze slika 3.3

CCCCC



Slika 4.1: Odločitveno drevo za izbor primerne metode.



Slika 4.2: Odločitveno drevo za izbor primerne metode.

Poglavje 5

Sklepne ugotovitve

Z izdelavo dodatka za program Orange smo zaključili delo na diplomski nalogi. Koda izdelanega dodatka se nahaja na git

Nas grafični dodatek za dostop do podatkov indikatorjev lahko nadgradimo tako, da uporabnikom grafičnega vmesnika omogočimo večjo izbiro oblik izhodnih podatkov in natančnejše presejanje rezultatov. Dodamo lahko tudi več metapodatkov na posamezne stolpce Orange tabele, ki nam omogočijo boljšo predstavnost v ostalih Orange gradnikih. V grafični vmesnik za dostop do podnebnih podatkov lahko dodamo še možnost izbire vodotočnih območij meritev. Za boljšo predstavbo bi lahko postopek izbire držav, regij in vodotočnih območij omogočili prek interaktivnega zemljevida sveta.

- dodamo metapodatke tudi climate gradniku - boljša pokritost testov
- V data sets skupino bi lahko dodali še gradnik za katerega od drugih v uvodu nastetih spletnih programskih vmesnikov

Literatura

- [1] World Development Indicators, The World Bank, (August 2016)
URL: <http://data.worldbank.org/data-catalog/world-development-indicators>

- [2] Data source: Global Financial Development Database (GFDD), The World Bank. Methodology citation: Martin Čihák, Aslı Demirgüç-Kunt, Erik Feyen, and Ross Levine, 2012. "Benchmarking Financial Systems Around the World." World Bank Policy Research Working Paper 6175, World Bank, Washington, D.C. (Junij 2016)
<http://data.worldbank.org/data-catalog/global-financial-development>

- [3] Africa Development Indicators, The World Bank (Februar 2013)
<http://data.worldbank.org/data-catalog/africa-development-indicators>

- [4] Doing Business, The World Bank (<http://www.doingbusiness.org>) (Julij 2016)
<http://data.worldbank.org/data-catalog/doing-business-database>

- [5] Enterprise Surveys, The World Bank (Julij 2016)
<http://data.worldbank.org/data-catalog/enterprise-surveys>

- [6] Millennium Development Goals, The World Bank (Julij 2016)
[http://data.worldbank.org/data-catalog/
millennium-development-indicators](http://data.worldbank.org/data-catalog/millennium-development-indicators)
- [7] World Bank EdStats (Junij 2016)
<http://data.worldbank.org/data-catalog/ed-stats>
- [8] Gender Statistics, The World Bank (Julij 2016)
<http://data.worldbank.org/data-catalog/gender-statistics>
- [9] HealthStats, World Bank Group (Julij 2016)
[http://data.worldbank.org/data-catalog/
health-nutrition-and-population-statistics](http://data.worldbank.org/data-catalog/health-nutrition-and-population-statistics)
- [10] IDA Results Measurement System, the World Bank (Julij 2016)
[http://data.worldbank.org/data-catalog/
IDA-results-measurement](http://data.worldbank.org/data-catalog/IDA-results-measurement)
- [11] Climatic Research Unit, University of East Anglia
<http://www.cru.uea.ac.uk/data>