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

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 25. 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

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 notacija

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 razlicnih 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 podatkovnh 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 dosezemo 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.

Ker je mogoče z eno poizvedbo dostopati do velike količine podatkov, ima programski vmesnik za dostop do podatkov indikatorjev implementirano ostranjevanje, s katerim je omejeno število podatkov ki jih lahko dobimo z eno poizvedbo. Tako so podatki razdeljeni na skupine ki jih imenujemo strani.

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, s sitom ki ne vrača nobenih 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 (Pri-

[4] `http://api.worldbank.org/indicators?format=json`
[5] `http://api.worldbank.org/lendingTypes?format=json`
[6] `http://api.worldbank.org/topics`

```
[
    {
        "page": 1,
        "pages": 137,
        "per_page": "50",
        "total": 6831
    },
    [
        <podatki>,
        ...
    ]
]
```

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

```
[
    {
        "message": [
            {
                "id": "120",
                "key": "Parameter 'country' has an invalid value",
                "value": "The provided parameter value is not valid"
            }
        ]
    }
]
```

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

mer 4). Programski vmesnik nam tudi omogoča dostop do podatkov posameznega indikatorja dolocenega s kodo in presejanje seznama indikatorjev glede na vir podatkov 3. V nasem programu smo uporabili le poizvedbo za celotn seznam indikatorjev, da smo omogocili iskanje in presejanje po vseh poljih indikatorjev.

```
http://api.worldbank.org/indicators?format=json
http://api.worldbank.org/indicators?format=json&source=5
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"

```
{
    "id": "1.0.HCount.2.5usd",
    "name": "Poverty Headcount (\$2.50 a day)",
    "source": {
        "id": "37",
        "value": "LAC Equity Lab"
    },
    "sourceNote": "The poverty headcount index measures the
                   proportion of the population with daily per
                   capita income (in  2005 PPP) below the poverty
                   line.",
    "sourceOrganization": "LAC Equity Lab tabulations of SEDLAC
                          (CEDLAS and the World Bank).",
    "topics": [
        {
            "id": "11",
            "value": "Poverty "
        }
    ]
}
```

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 (Premer 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 naštete 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”,

```
{
    "id": "ABW",
    "iso2Code": "AW",
    "name": "Aruba",
    "region": {
        "id": "LCN",
        "value": "Latin America & Caribbean "
    },
    "adminregion": {
        "id": "",
        "value": ""
    },
    "incomeLevel": {
        "id": "HIC",
        "value": "High income"
    },
    "lendingType": {
        "id": "LNX",
        "value": "Not classified"
    },
    "capitalCity": "Oranjestad",
    "longitude": "-70.0167",
    "latitude": "12.5167"
},
```

Primer 6: Izsek podatkov veljavne poizvedbe držav.

indicator_id polje "id" indikatorja ki je prikazano v Primeru 4.

parametri Dodatni parametri GET

Za poizvedbe do podatkov indikatorjev so poleg osnovnih parametrov GET `per_page`, `page` in `format`, opisanih v poglavju 2.1, na voljo tudi dodatni parametri za presejanje rezultatov poizvedbe:

MRV Stevilska 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.

gapfill Zastavica `'y'` ali `'n'` za manjkajoče vrednosti meritev. Vrednost `'y'` kombinaciji z poljem `mrv` poskrbi da programski vmesnik ne izpusti nobenega časovnega intervala.

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

```
1  http://api.worldbank.org/en/countries/<country>/indicators/<↩
       indicator_id>?<parametri>
```

Primer 7: Osnovna oblika poizvedbe za podatke enega indikatorja.

Privzeta vrednost za količino podatkov na stran `per_page` je 50. Zgornja meja pa ni strogo določena, vendar je odvisna od velikosti odgovora. Ugotovili smo, da se zanesljivost programskega vmesnika manjša z večjo količino podatkov na stran. V našem programu smo se omejili na 1000 podatkov na stran, kar se je izkazalo za uporabno razmerje med hitrostjo in zanesljivostjo programskega vmesnika. Privzeto bo programski vmesnik vrnil podatke za vse časovne vrednosti. V odgovoru API-ja dobimo seznam objektov (Primer 8) z datumom, indikatorjem, državo in vrednostjo.

Slabosti programskega vmesnika indikatorjev Svetovne banke za uporabo v namene podatkovnega rudarjenja so v tem, da vmesnik ni namenjen prenosu večje kolicine podatkov z eno samo poizvedbo. Zaradi ostranjevanja

```
{
    "indicator": {
        "id": "SP.POP.TOTL",
        "value": "Population, total"
    },
    "country": {
        "id": "IL",
        "value": "Israel"
    },
    "value": "6289000",
    "decimal": "0",
    "date": "2000"
}
```


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

moramo za en sam indikator narediti več poizvedb, da prenesemo podatke z vseh strani. Prav tako podatkovni vmesnik ne podpira poizvedb po večih indikatorjih hkrati, kar potrebujemo za iskanje zakonitosti med posameznimi indikatorji.

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 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

Za dostop do podnebnih podatkov preko programskega vmesnika Svetovne banke, potrebujemo ISO-3166-1 alpha3 kodo države, ali številski identifikator vodotočnega območja (Slika 2.1). Programski vmesnik nam omogoča dostop do meritev povprecnih količin padavin in temperatur za letno ali desetleno obdobje. Poleg letnega in desetletnega obdobja pa nam programski vmensik ponuja tudi povprecno kolicino padavin in temperatur za posamezne mesece skozi vsa leta meritev. Obliko poizvedbe prikazuje primer 9, kjer je:

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

data_type vrsta meritev ("pr" za padavine, "tas" za temperature),

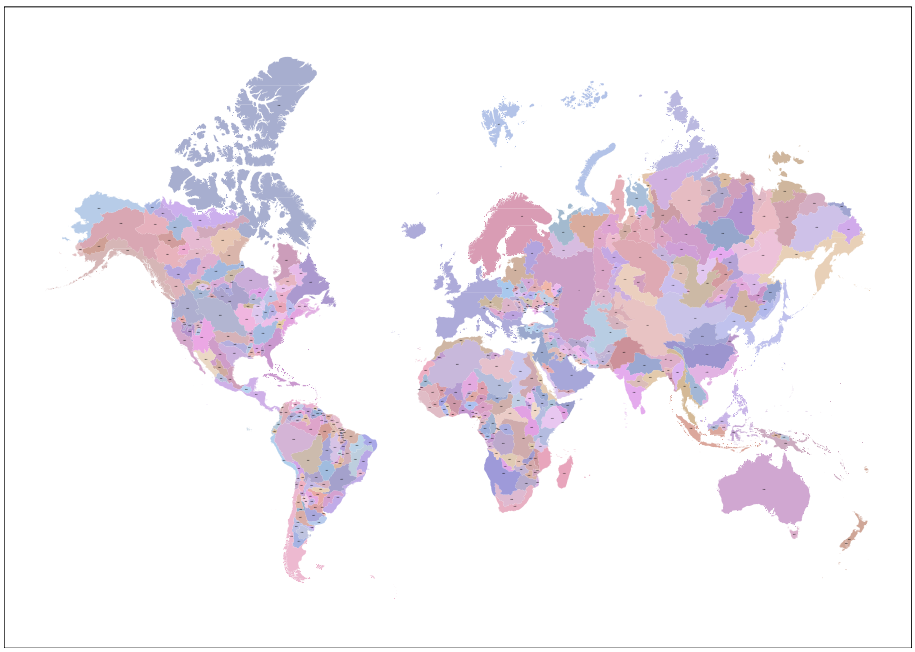
interval vrsta meritvenega obdobja ("month" za mesečno, "year" za letno in "decade" za desetletno),

location koda države ali stevilski identifikator vodotočnega območja.

Za razliko od programskega vmesnika indikatorjev, nam programskem vmesniku podnebnih meritev z eno poizvedbo omogoca dostop do podatkov le za eno državo. To pomeni da je količina podatkov dovolj omejena da nam programski vmesnki, vedno vrne vse podatke brez ostranjevanja, kot prikazuje primer 10,

```
1  http://climatedataapi.worldbank.org/climateweb/rest/v1/<loc_type>/cru↩
       /<data_type>/<interval>/<location>
```

Primer 9: Osnovna oblika poizvedbe za podnebne podatke.



Slika 2.1: Prikaz vodotočnih območij sveta.

```
[
    {
        "month": 0,
        "data": 68.93643
    },
    {
        "month": 1,
        "data": 64.23069
    },
    {
        "month": 2,
        "data": 81.098724
    },
    ...
]
```

Primer 10: Primer odgovora za poizvedbo količine padavin v posameznih mesecih v Sloveniji.

2.3 Težave pri uporabi programskih vmesnikov Svetovne banke

Programski vmesniki Svetovne banke zajema podatke različnih virov, zato je teško zagotoviti pravilnost in konsistentnost podatkov. Poleg tega pa se programski vmesnik in spletna stran z dokumentacijo občasno spremenita, kar povzroča še dodatne težave pri uporabi. Nekatere težave, ki smo jih opazili so:

- nekaterim delom dokumentacije se je med izdelavo te diplomske naloge spremenil spletni naslov, tako da do tistit delov sedaj nimamo več dostopa,
- polje za datum v odgovoru je opisano, vendar ni dokumentirano, kakšne so vse možne vrednosti (nekaj primerov nedokumentiranih vrednosti: “last known value” “2001 - 2015” “2040”),
- delovanje sita z različnimi kombinaciji polj `mrv`, `gapfill` in `date` ni ustrezno opisano,
- v odgovoru poizvedbe po podatkih indikatorjev, ponekod manjkajo vrednosti kot so koda države, ime države ali ime indikatorja.
- zgornja meja števila izbranih lokacij na 250 ni navedena in napaka ki jo programski vmesnik vrne ni dokumentirana,
- 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 razsiritve in zanesljivost. Ta cilja dosežemo z mehanizmom za vkljucevanje 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 podatkovno tabelo Orange in tabelo numpy. Ta sklop je namenjen skriptnemu delu s programom Orange [12] 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 v podatkih.

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 prejetih podatkov. Druga lastnost je pretvorba podatkov iz več dimenzij v dvo-dimenzionalno polje, primerno za uporabo v programu Orange. Glavna razreda te knjižnice sta `IndicatorAPI` in `ClimateAPI`. Prvi omogoča pridobivanje podatkov iz programskega vmesnika indikatorjev, drugi pa s programskega vmesnika podnebnih meritev.

Čeprav za dostop do programskega vmesnika Svetovne banke že obstajajo rešitve kot sta knjižnici `wbdata`[1] in `wbpy`[2], smo se odločili za lastno implementacijo podobne knjižnice. Glavni razlog za to je, da obstoječe rešitve poskušajo čim bolj natančno predstaviti programski vmesnik Svetovne banke, ne pa olajšati dostop do čim večje količine podatkov.

Za potrebe te knjižnice smo razvili lastno rešitev za predpomnenje poizvedb, saj so se bolj splošne rešitve, kot naprimer `vcrpy`[3] in `requests-cache`[4], izkazale za prepočasne ko delamo z večjimi količinami podatkov. Naša rešitev za predpomnenje izkorišča dejstvo da je vsaka poizvedba določena le z naslovom URL, in da so vsi odgovori oblike JSON. Za vsak URL naredimo novo datoteko v sistemskem začasnem imeniku, v kateri hranimo serializirane JSON podatke. Ker se podatki na programskem vmesniku Svetovne banke redko posodabljajo, smo za čas veljavosti začasnih datotek izbrali en teden.

3.1.1 Razred IndicatorAPI

`IndicatorAPI` je razred namenjen pridobivanju podatkov indikatorjev razvoja držav. Ker ima programski vmesnik Svetovne banke omejitev koliko

[1] `https://pypi.python.org/pypi/wbdata`
[2] `https://pypi.python.org/pypi/wbpy/2.0.1`
[3] `https://pypi.python.org/pypi/vcrpy/1.10.0`
[4] `https://pypi.python.org/pypi/requests-cache`

podatkov lahko prenesemo z eno poizvedbo in nam dovoli tvoriti poizvedbe le za en indikator na enkrat, smo napisali razred, ki v ozadju tvori in izvede poizvedbe za vse strani vseh zahtevanih indikatorjev. To poskrbi tako da se po prvi poizvedbi za en indikator sprehodi čez število preostalih strani (Primer 1), ki so na voljo, in pridobljene podatke večih strani združi in predstavi kot rezultat ene same poizvedbe. Ta postopek ponovi za vse zahtevane indikatorje, in njihove rezultate vrne v obliki slovarja, ki ima za ključ kodo indikatorja posamezne zahteve.

Poleg tega da skrbi za prenos vseh strani podatkov, tudi beleži število izvedenih in število potrebnih poizvedb za celoten prenos. Ta števila se lahko uporablja za prikaz napredka prenosa podatkov.

mehanizmem za odpravo napak: - pridobiti podatke o drzavi za posamezni indikator: - ob manjkajocih id-jih poskusamo dolociti id iz imena - ob manjkajocih imenih poskusamo dobiti ime iz id-ja - ce ni nobene informacije ta podatek ignoriramo.

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

Glavne metode ki jih ponuja razred IndicatorAPI so:

get_indicators za pridobivanje seznama indikatorjev s kodami, imeni in opisi,

get_countries za pridobivanje seznama drzav z metapodatki,

get_dataset za pridobivanje podatkov za indikatorjev.

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 nacina; kot

slovar slovarjev in dvo dimenzionalen seznam. Poleg omenjenih nacinov predstavitve podatkov lahko dostopamo tudi do neobdelanih podatkov prejetih z programskega vmesnika za vsako poizvedbo posebej.

Posamezne vrednosti teh podatkov so določene z drzavo, casovno komponento, in kodo indikatorja. Te podatke lahko predstavimo na dva glavna nacina:

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

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

Indicator

3.1.2 Pomcnik ClimateAPI

IndicatorAPI je

api dovoli le podatke za en tip za eno vrsto obdobja in eno drzavo 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 upostevati 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

drzava / 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 drzav ko ne prikazujemo v obliki casovne vrste.

api vrapper 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 zeljeno obliko.

3.3 Graficni vmesnik

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

2 gradnika - wb indicators in wb climate

Slika 3.1: Skupina gradnikov data sets

- lazja uporaba apija - vecja preglednost

za oba gradniko smo razvili in uporabili base class - skupni podatki

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

za te gradnike smo tudi napisali enotske teste.

3.3.1 wb indicators gradnik

za sestavo smo si pomagali z gradniki Orange.gui

elementi gradnika

2 osnovna filtra:

- izbor indikatorjev ki se pokazejo 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

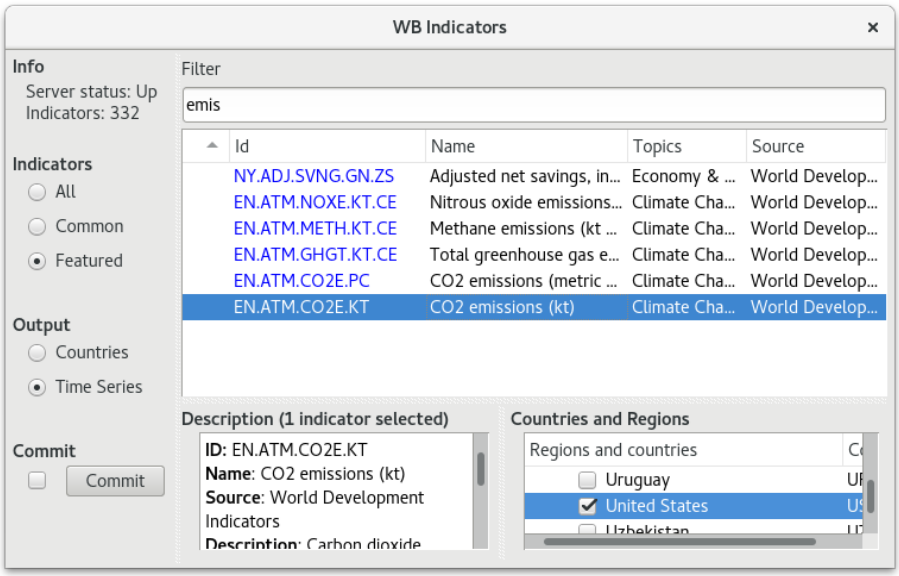
gradnik ima sistem za prikaz (progress bar?)

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

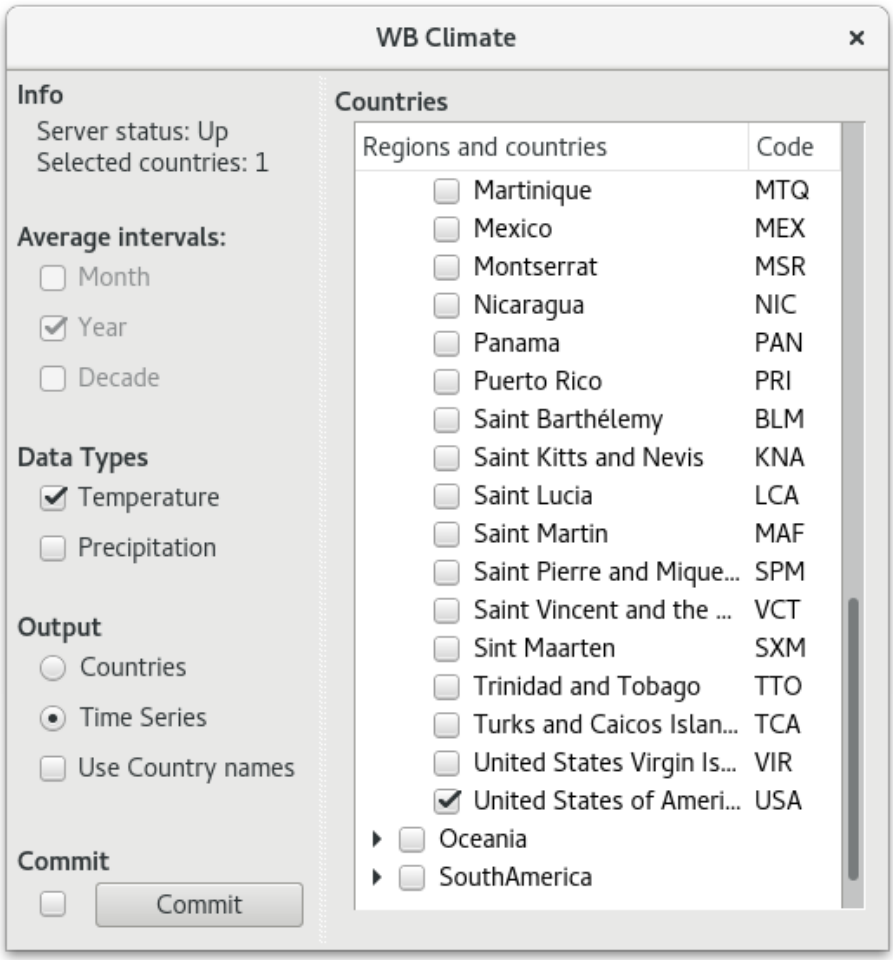
3.3.2 wb climate gradnik

dovoli izbiro posameznih drzav

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



Slika 3.2: Odločitveno drevo za izbor primerne metode.



Slika 3.3: Odločitveno drevo za izbor primerne metode.

Poglavje 4

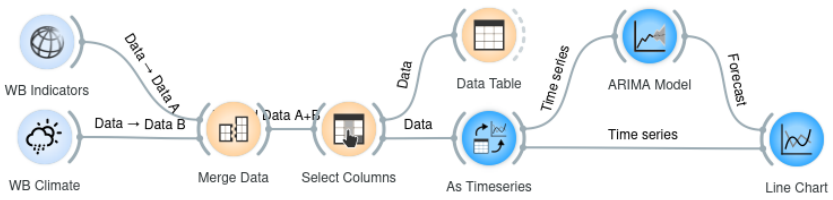
Primeri uporabe

uporaba: slike orange uporabe

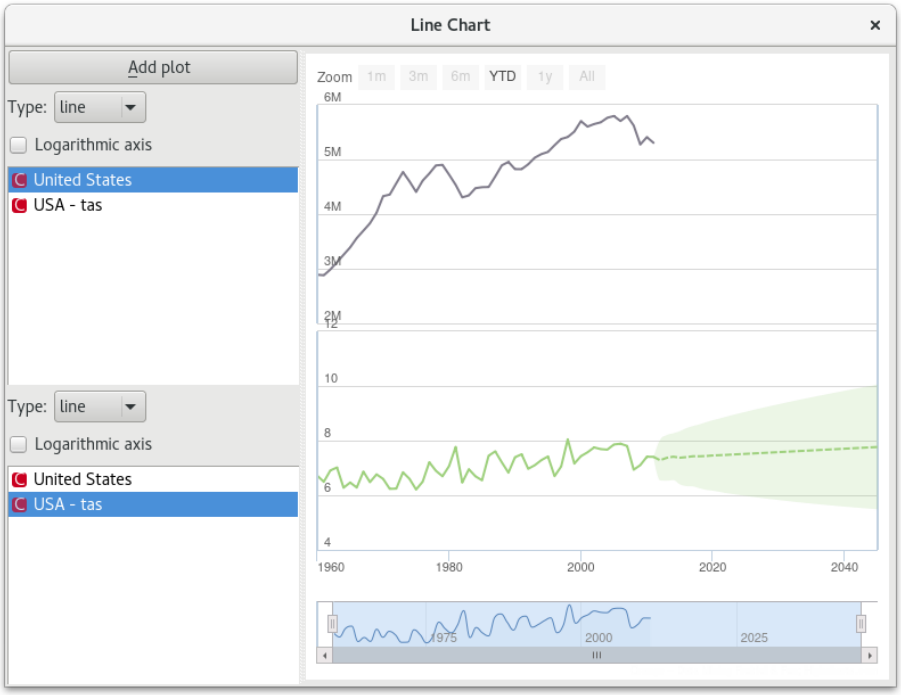
4.1 Napoved temperature s pomocjo CO2 emisij

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 zakljucili 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 natančnejse 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 območji meritev. Za boljšo predstavo bi lahko postopek izbire drzav, regij in vodotočnih območij (Slika 2.1) omogočili prek interaktivnega zemljevida sveta.

- dodamo metapodatke tudi climate gradniku - boljsa pokritost testov

- V data sets skupino bi lahko dodali se 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

[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

[10] IDA Results Measurement System, the World Bank (Julij 2016)
http://data.worldbank.org/data-catalog/
IDA-results-measurement

[11] Climatic Research Unit, University of East Anglia
http://www.cru.uea.ac.uk/data

[12] Janez Demšar and Tomaž Curk and Aleš Erjavec and Črt Gorup and Tomaž Hočevar and Mitar Milutinovič and Martin Možina and Matija Polajnar and Marko Toplak and Anže Starič and Miha Štajdohar and Lan Umek and Lan Žagar and Jure Žbontar and Marinka Žitnik and Blaž Zupan, "Orange: Data Mining Toolbox in Python," Journal of Machine Learning Research, vol. 14, pp. 2349-2353, 2013.

[13] Jernej Kernc, "Orodje za interaktivno analizo časovnih vrst," 2016

[14] Jure Dimec (2002), Medjezično iskanje dokumentov
http://clir.craynaud.com/clir/MEDJEZICNOISKANJEDOKUMENTOV.
pdf