Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Miha Kranjc, Marko Adžaga, Matic Hrastelj

Domača naloga 1 - Poročilo

Iskanje in ekstrakcija podatkov s spleta,
Domača naloga 1

MENTOR: prof. dr. Marko Bajec, Timotej Knez

Posebnosti, odločitve in opis implementacije

Implementacija pajka je ločena po komponentah, ki opravljajo različne naloge. Glavni del je razred *Crawler,* ki nadzoruje celoten proces ekstrakcije. Na začetku izvajanja ustvari določeno število niti oziroma delavcev. Vsaka nit izvaja neskončno zanko, ki iz podatkovne baze prevzame spletno stran, iz nje izlušči podatke ter rezultate zapiše nazaj v podatkovno bazo. Dostopi do baze so zavarovani s pomočjo ključavnice. Samo luščenje podatkov se izvaja v ločenem razredu *Extractor,* s pomočjo knjižnic *Selenium* in *BeautifulSoup*. Med obiskovanjem spletnih strani vedno upoštevamo potencialen časovni zamik pri nalaganju ter vse možne izjeme, ki se lahko zgodijo pri poizvedovanju. Rezultati se naknadno vpišejo nazaj v podatkovno bazo. Tokrat nam ni potrebno skrbeti za hkraten dostop, saj ne more priti do konflikta. V podatkovno bazo v tabelo *page* smo dodali tudi stolpec, ki vsebuje hash vrednost HTML strani, za lažje primerjanje med stranmi.

Parametri pajka in težave

Implementiran pajek ima zgolj tri parametre, ki jih lahko določi uporabnik. V datoteki *main.py* lahko nastavimo semenska spletna mesta oziroma izhodišča (trenutno so podana štiri spletna mesta iz navodil naloge), število niti z argumentom *worker_count*, s katerimi pajek prečesava splet ter minimalni pretečeni čas med klici na spletno mesto *default_access_period,* ki je privzeto nastavljen na 5 sekund. Opcijsko se lahko nastavijo še parametri za povezavo na podatkovno bazo v samem razredu pajka *Crawler*. Parametrov, ki pa jih uporabnik ne more nastaviti in se nastavijo ob vzpostavitvi pajka, pa je kar precej. Od ključavnice, do slovarja dostopnih časov vseh spletnih mest in samih ekstraktorjev.

Med samo implementacijo in testiranjem smo imeli tudi kar nekaj težav. Nekatere omembe vredne težave so bile problem s certifikatom SSL, kjer smo morali dovoliti, da ni potreben, nekaj težav smo imeli tudi s ključavnico, ter tudi s počasnim dodajanjem zapisov v bazo, kjer smo morali uporabiti poseben ukaz za hitrejše vnašanje.

Statistika

Spletna mesta iz seznama semen:

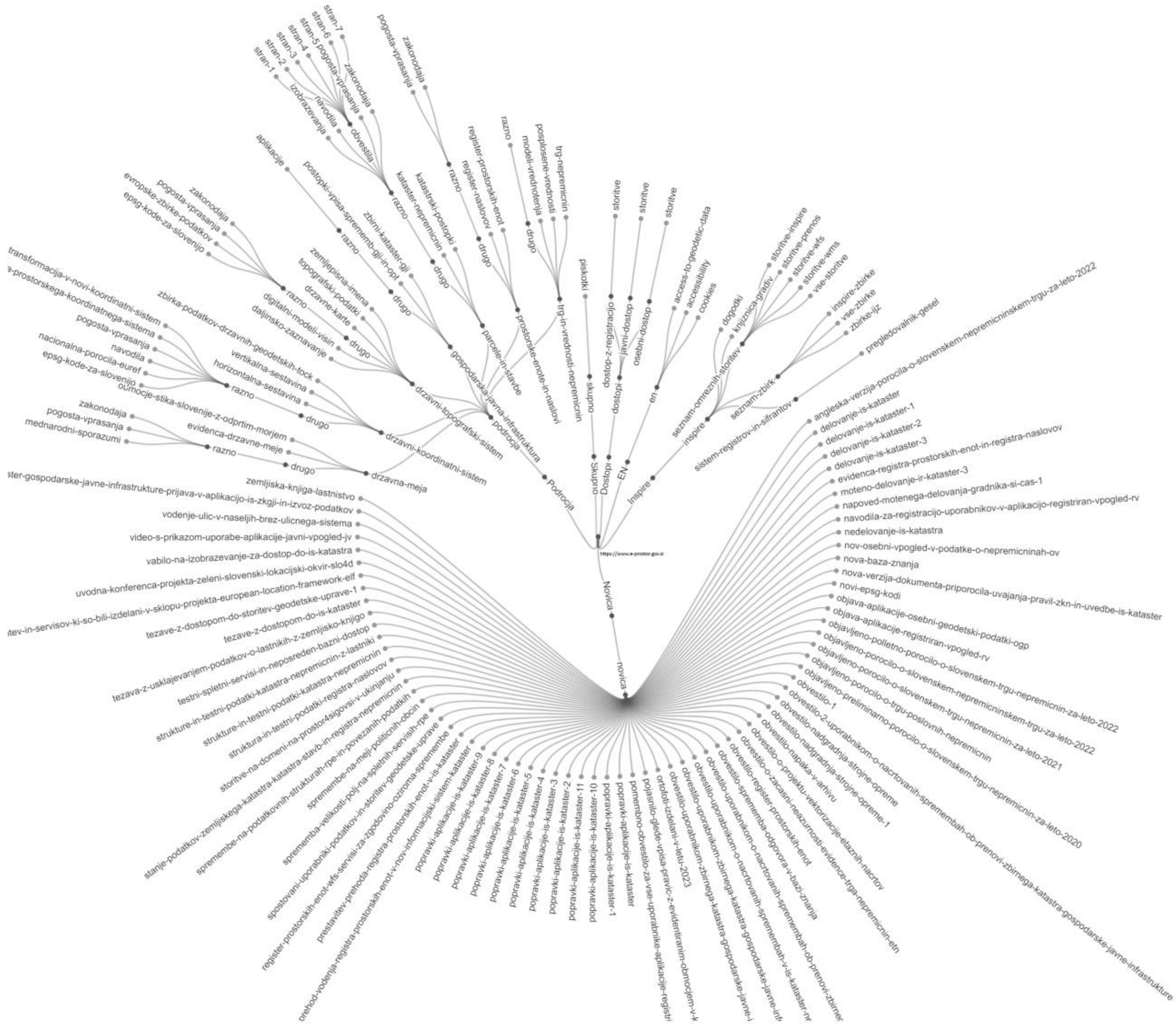
- Število spletnih mest: **4**
- Število spletnih strani (HTML, BINARY, FRONTIER, DUPLICATE): **29.888**
- Število spletnih strani (HTML): **23.303**
- Število duplikatov: **768**
- Število binarnih elementov glede na tip:
    - PDF: **25.475**
    - DOC: **3.157**
    - DOCX: **8.803**
    - PPT: **11**
    - PPTX: **198**
- Število slik: **52.831**
- Povprečno število slik na spletno stran: **2,4**

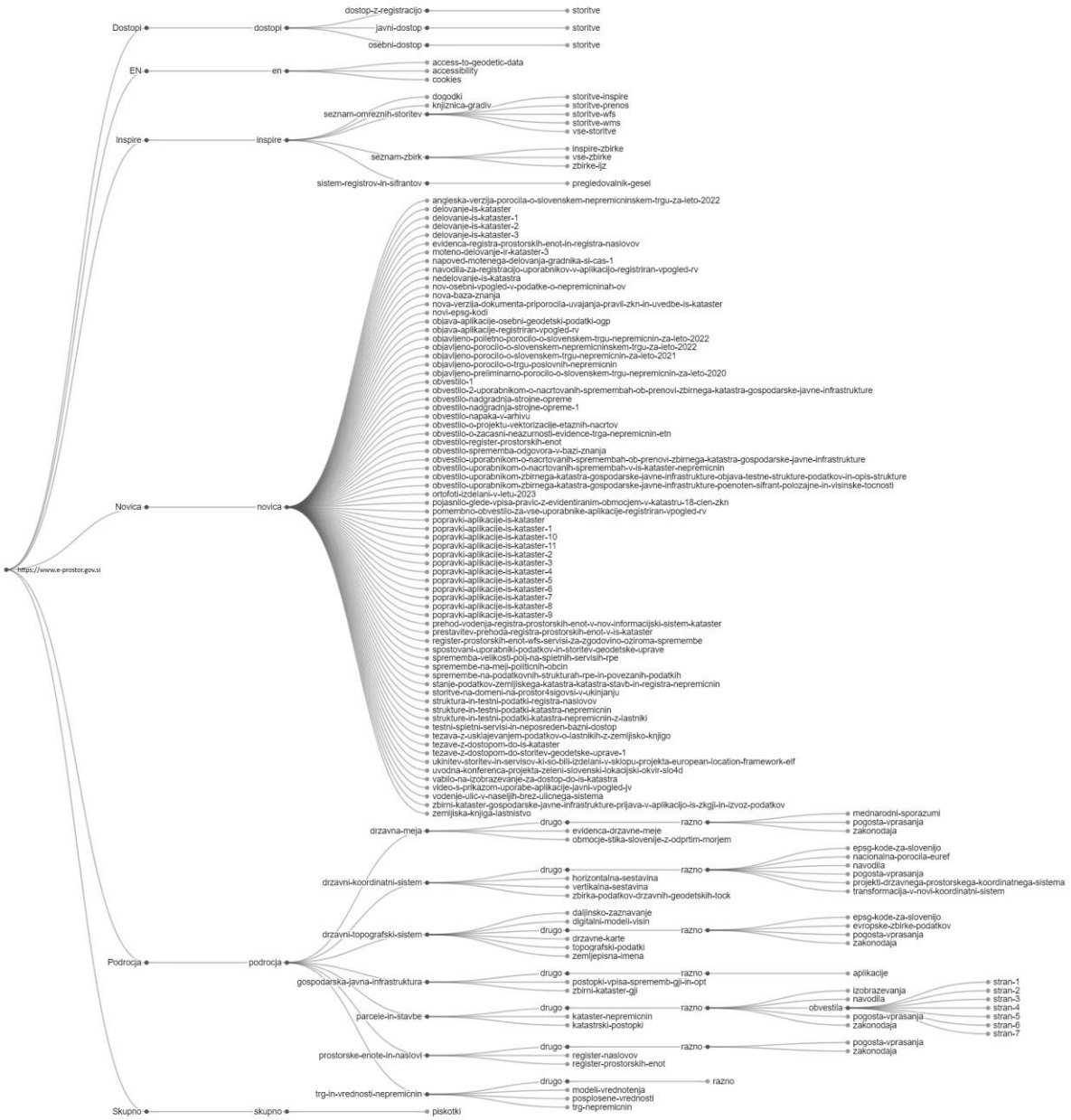
Celotna podatkovna baza crawldb:

- Število spletnih mest: **318**
- Število spletnih strani (HTML, BINARY, FRONTIER, DUPLICATE): **133.301**
- Število spletnih strani (HTML): **52.735**
- Število duplikatov: **6.674**
- Število binarnih elementov glede na tip:
    - PDF: **55.343**
    - DOC: **7.565**
    - DOCX: **14.499**
    - PPT: **78**
    - PPTX: **259**
- Število slik: **488.941**
- Povprečno število slik na spletno stran: **11,2**

Vizualizacija povezav

Zaradi prevelikega števila skupnih url-jev smo se odločili vizualizirati samo url-je z spletne domene https://www.e-prostor.gov.si. Vizualizacijo smo dodatno omejili z poddomenami: dostopi, inspire, podrocja, novica, en, skupno.


https://www.e-prostor.gov.si
Podrocja
Skupno
Dostopi
EN
Inspire
Novica
novica
drzavna-meja
drzavni-koordinatni-sistem
drzavni-topografski-sistem
gospodarska-javna-infrastruktura
parcele-in-stavbe
prostorske-enote-in-naslovi
trg-in-vrednotenje-nepremicnin
podrocja
skupno
piskotki
dostop-z-registracijo
javni-dostop
osebni-dostop
storitve
en
access-to-geodetic-data
accessibility
cookies
inspire
seznam-omreznih-storitev
seznam-zbirk
sistem-registrov-in-sifrantov
dogodki
knjiznica-gradiv
storitve-inspire
storitve-prenos
storitve-wfs
storitve-wms
vse-storitve
inspire-zbirke
vse-zbirke
zbirke-ljz
pregledovalnik-gesel
zemljiska-knjiga-lastnistvo
vodenje-ulic-v-naseljih-brez-ulicnega-sistema
video-s-prikazom-uporabe-aplikacije-javni-vpogled-jv
vabilo-na-izobrazevanje-za-dostop-do-is-katastra
uvodna-konferenca-projekta-zeleni-slovenski-lokacijski-okvir-slo4d
tezave-z-dostopom-do-storitev-geodetske-uprave-1
tezave-z-dostopom-do-is-katastra
tezave-z-dostopom-do-zemljisko-knjigo
tezava-z-usklajevanjem-podatkov-o-lastnikih-z-zemljiski-knjigo
testni-spletni-servisi-in-neposreden-bazni-dostop
strukture-in-testni-podatki-katastra-nepremicnin-z-lastniki
strukture-in-testni-podatki-katastra-nepremicnin
strukture-in-testni-podatki-registra-naslovov
struktura-in-testni-podatki-registra-nepremicnin
storitve-na-domeni-na-prostor4sigovsi-v-ukinjanju
spremembe-na-meji-politicnih-obcin
spremembe-na-podatkovnih-strukturah-rpe-in-povezanih-podatkih
sprememba-velikosti-polj-na-spletnih-servisih-rpe
spostovani-uporabniki-podatkov-in-storitev-geodetske-uprave
popravki-aplikacije-is-kataster
popravki-aplikacije-is-kataster-1
popravki-aplikacije-is-kataster-2
popravki-aplikacije-is-kataster-3
popravki-aplikacije-is-kataster-4
popravki-aplikacije-is-kataster-5
popravki-aplikacije-is-kataster-6
popravki-aplikacije-is-kataster-7
popravki-aplikacije-is-kataster-8
popravki-aplikacije-is-kataster-9
popravki-aplikacije-is-kataster-10
popravki-aplikacije-is-kataster-11
pomembno-obvestilo-za-vse-uporabnike-aplikacije-registr
pojasnilo-glede-vpisa-pravic-z-evidentiranim-obmocjem-v-k
ortofoti-izdelani-v-letu-2023
obvestilo-1
obvestilo-2-uporabnikom-o-nacrtovanih-spremembah-ob-prenovi-zbirnega-katastra-gospodarske-javne-infrastrukture
obvestilo-nadgradnja-strojne-opreme
obvestilo-nadgradnja-strojne-opreme-1
obvestilo-napaka-v-arhivu
obvestilo-o-projektu-vektorizacije-etaznih-nacrtov
obvestilo-o-zacasni-neazurnosti-evidence-trga-nepremicnin-etn
obvestilo-register-prostorskih-enot
obvestilo-sprememba-odgovora-v-bazi-znanja
angleska-verzija-porocila-o-slovenskem-nepremicninskem-trgu-za-leto-2022
delovanje-is-kataster
delovanje-is-kataster-1
delovanje-is-kataster-2
delovanje-is-kataster-3
evidenca-registra-prostorskih-enot-in-registra-naslovov
moteno-delovanje-ir-kataster-3
napoved-motenega-delovanja-gradnika-si-cas-1
navodila-za-registracijo-uporabnikov-v-aplikacijo-registriran-vpogled-rv
nedelovanje-is-katastra
nov-osebni-vpogled-v-podatke-o-nepremicninah-ov
nova-baza-znanja
nova-verzija-dokumenta-priporocila-uvajanja-pravil-zkn-in-uvedbe-is-kataster
novi-epsg-kodi
objava-aplikacije-osebni-geodetski-podatki-ogp
objava-aplikacije-registriran-vpogled-rv
objavljeno-polletno-porocilo-o-slovenskem-trgu-nepremicnin-za-leto-2022
objavljeno-porocilo-o-slovenskem-nepremicninskem-trgu-za-leto-2022
objavljeno-porocilo-o-trgu-poslovnih-nepremicnin-za-leto-2021
objavljeno-preliminarno-porocilo-o-slovenskem-trgu-nepremicnin-za-leto-2020
zakonodaja
pogosta-vprasanja
mednarodni-sporazumi
razno
drugo
evidenca-drzavne-meje
oumocje-stika-slovenije-z-odprtim-morjem
epsg-kode-za-slovenijo
nacionalna-porocila-euref
navodila
transformacija-v-novi-koordinatni-sistem
zbirka-podatkov-drzavnih-geodetskih-tock
vertikalna-sestavina
horizontalna-sestavina
evropske-zbirke-podatkov
zemljepisna-imena
topografski-podatki
drzavne-karte
digitalni-modeli-visin
daljinsko-zaznavanje
aplikacije
postopki-vpisa-sprememb-gji-in-opt
zbirni-kataster-gji
katastrski-postopki
kataster-nepremicnin
register-prostorskih-enot
register-naslovov
modeli-vrednotenja
posplosene-vrednosti
trg-nepremicnin
obvestila
izobrazevanja
stran-1
stran-2
stran-3
stran-4
stran-5
stran-6
stran-7


https://www.e-prostor.gov.si
Dostopi
dostopi
dostop-z-registracijo
storitve
javni-dostop
storitve
osebni-dostop
storitve
EN
en
access-to-geodetic-data
accessibility
cookies
Inspire
inspire
seznam-omrežnih-storitev
dogodki
knjiznica-gradiv
storitve-inspire
storitve-prenos
storitve-wfs
storitve-wms
vse-storitve
seznam-zbirk
inspire-zbirke
vse-zbirke
zbirke-jlz
sistem-registrov-in-sifrantov
pregledovalnik-gesel
Novica
novica
angleska-verzija-porocila-o-slovenskem-nepremicninskem-trgu-za-leto-2022
delovanje-is-kataster
delovanje-is-kataster-1
delovanje-is-kataster-2
delovanje-is-kataster-3
evidenca-registra-prostorskih-enot-in-registra-naslovov
moteno-delovanje-if-kataster-3
napoved-motenega-delovanja-gradnika-si-cas-1
navodila-za-registracijo-uporabnikov-v-aplikacijo-registriran-vpogled-rv
nedelovanje-is-katastra
nov-osebni-vpogled-v-podatke-o-nepremicninah-ov
nova-baza-znanja
nova-verzija-dokumenta-priporocila-uvajanja-pravil-zkn-in-uvedbe-is-kataster
novi-epsg-kodi
objava-aplikacije-osebni-geodetski-podatki-ogp
objava-aplikacije-registriran-vpogled-rv
objavljeno-poletno-porocilo-o-slovenskem-trgu-nepremicnin-za-leto-2022
objavljeno-porocilo-o-slovenskem-nepremicninskem-trgu-za-leto-2022
objavljeno-porocilo-o-slovenskem-trgu-nepremicnin-za-leto-2021
objavljeno-porocilo-o-trgu-poslovnih-nepremicnin
objavljeno-preliminarno-porocilo-o-slovenskem-trgu-nepremicnin-za-leto-2020
obvestilo-1
obvestilo-2-uporabnikom-o-nacrtovanih-spremembah-ob-prenovi-zbirnega-katastra-gospodarske-javne-infrastrukture
obvestilo-nadgradnja-strojne-opreme
obvestilo-nadgradnja-strojne-opreme-1
obvestilo-napaka-v-arhivu
obvestilo-o-projektu-vektorizacije-etaznih-nacrtov
obvestilo-o-zacasni-neazurnosti-evidence-trga-nepremicnin-etn
obvestilo-register-prostorskih-enot
obvestilo-sprememba-odgovora-v-bazi-znanja
obvestilo-uporabnikom-o-nacrtovanih-spremembah-ob-prenovi-zbirnega-katastra-gospodarske-javne-infrastrukture
obvestilo-uporabnikom-o-nacrtovanih-spremembah-v-is-kataster-nepremicnin
obvestilo-uporabnikom-zbirnega-katastra-gospodarske-javne-infrastrukture-objava-testne-strukture-podatkov-in-opis-strukture
obvestilo-uporabnikom-zbirnega-katastra-gospodarske-javne-infrastrukture-poenoten-sifrant-polozajne-in-visinske-tocnosti
ortofoti-izdelani-v-letu-2023
pojasnilo-glede-vpisa-pravic-z-evidentiranim-obmocjem-v-katastru-18-clen-zkn
pomembno-obvestilo-za-vse-uporabnike-aplikacije-registriran-vpogled-rv
popravki-aplikacije-is-kataster
popravki-aplikacije-is-kataster-1
popravki-aplikacije-is-kataster-10
popravki-aplikacije-is-kataster-11
popravki-aplikacije-is-kataster-2
popravki-aplikacije-is-kataster-3
popravki-aplikacije-is-kataster-4
popravki-aplikacije-is-kataster-5
popravki-aplikacije-is-kataster-6
popravki-aplikacije-is-kataster-7
popravki-aplikacije-is-kataster-8
popravki-aplikacije-is-kataster-9
prehod-vodenja-registra-prostorskih-enot-v-nov-informacijski-sistem-kataster
preslavitev-prehoda-registra-prostorskih-enot-v-is-kataster
register-prostorskih-enot-wfs-servisi-za-zgodovino-oziroma-spremembe
spostovani-uporabniki-podatkov-in-storitev-geodetske-uprave
sprememba-velikosti-poli-na-spletnih-servisih-rpe
spremembe-na-meji-politicnih-obcin
spremembe-na-podatkovnih-strukturah-rpe-in-povezanih-podatkih
stanje-podatkov-zemljiskega-katastra-katastra-stavb-in-registra-nepremicnin
storitve-na-domeni-na-prostor4sigovsi-v-ukinjanju
struktura-in-testni-podatki-registra-naslovov
strukture-in-testni-podatki-katastra-nepremicnin
strukture-in-testni-podatki-katastra-nepremicnin-z-lastniki
testni-spletni-servisi-in-neposreden-baznl-dostop
tezava-z-uskladjevanjem-podatkov-o-lastnikih-z-zemljisko-knjigo
tezave-z-dostopom-do-is-kataster
tezave-z-dostopom-do-storitev-geodetske-uprave-1
ukinitev-storitev-in-servisov-ki-so-bili-izdelani-v-okviru-projekta-european-location-framework-elf
uvodna-konferenca-projekta-zeleni-slovenski-lokacijski-okvir-slo4d
vabilo-na-izobrazevanje-za-dostop-do-is-katastra
video-s-prikazom-uporabe-aplikacije-javni-vpogled-jv
vodenje-ulic-v-naseljih-brez-ulicnega-sistema
zbirni-kataster-gospodarske-javne-infrastrukture-prijava-v-aplikacijo-is-zkgji-in-izvoz-podatkov
zemljiska-knjiga-lastnistvo
Podrocja
podrocja
drzavna-meja
drugo
evidenca-drzavne-meje
obmocje-stika-slovenije-z-odprtim-morjem
razno
mednarodni-sporazumi
pogosta-vprasanja
zakonodaja
drzavni-koordinatni-sistem
drugo
horizontalna-sestavina
vertikalna-sestavina
zbirka-podatkov-drzavnih-geodetskih-tock
razno
epsg-kode-za-slovenijo
nacionalna-porocila-euref
navodila
pogosta-vprasanja
projekti-drzavnega-prostorskega-koordinatnega-sistema
transformacija-v-novi-koordinatni-sistem
drzavni-topografski-sistem
daljinsko-zaznavanje
digitalni-modeli-visin
drugo
drzavne-karte
topografski-podatki
zemljepisna-imena
razno
epsg-kode-za-slovenijo
evropske-zbirke-podatkov
pogosta-vprasanja
zakonodaja
gospodarska-javna-infrastruktura
drugo
postopki-vpisa-sprememb-gji-in-opt
zbirni-kataster-gji
razno
aplikacije
parcele-in-stavbe
drugo
kataster-nepremicnin
katastrski-postopki
razno
obvestila
izobrazevanja
navodila
pogosta-vprasanja
zakonodaja
stran-1
stran-2
stran-3
stran-4
stran-5
stran-6
stran-7
prostorske-eriote-in-naslovi
drugo
register-naslovov
register-prostorskih-enot
razno
pogosta-vprasanja
zakonodaja
trg-in-vrednosti-nepremicnin
drugo
modeli-vrednotenja
posplosene-vrednosti
trg-nepremicnin
razno
Skupno
skupno
piskotki