Uvod. V delu je predstavljena prva seminarska naloga in sicer izdelava spletnega pajka. Spletni pajek pregleda določeno količino spletnih strani z izbranih domen, iz njihovih vsebin izloči slike, linke in druge binarne datoteke. Spletni pajek v svoji podatkovni dazi shranjuje seznam zapisov z obiskanih strani, njihove duplikate, strani, ki jih še mora obikati in tiste, ki jih ne sme.

Za izbrane domene smo vzeli domene gov.si in nekaj poljubnih. Na njih smo pognali spletnega pajka in po večurnem delovanju programa zbrali podatke o preiskanih straneh, kolilini slik in drugih binarnih datotek.

Podatki. Za začetne podatke smo dobili seznam državnih domen npr. evem.gov.si, e-uprava.gov.si, podatki.gov.si... Poleg teh smo dodali še nekaj drugih državnih in poljubnih domen.

Domene: evem.gov.si, e-uprava.gov.si, podatki.gov.si, e-prostor.gov.si, mizs.gov.si, mddsz.gov.si, mf.gov.si, www.mgrt.gov.si, ucilnica.fmf.uni-lj.si, 21run.com, dibikibi.com.

Implementacija

Podatkovna baza kot Frontier

Za podatkovno bazo smo, kot je bilo priporočeno, vzeli Progesql in opisano podatkovno shemo. Vsako html stran, ki je bila primerna za obdelavo smo dodali v podatkovno bazo z oznako FRONTIER. Da smo vedeli katera v vrsti je novo dodana spletna stran, smo tabeli s stranmi dodali nov stolpec z imenom $bfslevel$. Ta vsebuje število za katero je stran oddaljena od prve preiskane strani z isto domeno. Vsaka nova stran ima vrednost $bfslevel$ za eno večjo od starša.

Ko izbiramo novo stran za v obdelavo, išemo od najmanjše $bfslevel$, kar je 0, navzgor. Najprej preverimo ali obstaja stran, ki je že 1 oddaljena od osnovne spletne strani, če obstaja, jo vzamemo v obdelavo. V nasprotnem primeru iščemo spletno stran z za ena višjim levelom. Na tak način smo implementirali iskanje v širino.

Duplikati Da je bilo sledenje duplikatom lažje smo tabeli-strani v podatkovni bazi dodali polje $hash - content$, ki vsebuje zgoščeno vsebino spletne strani. Za vsak url, ki smo ga izluščili na html strani, smo najprej preverili v katero domeno spada, ali že obstaja stran z enakim urljem v podatkovni bazi, nato pa še obstoj strani z enako vsebino na podlagi zgoščene vsebine. Za zgoščevanje smo uporabili sha256 algoritem, ki ga zaradi pomanjkanja časa nismo natanko preučili. Najverjetneje je nekoliko prepotraten za naše namene.

Robots.txt Podatkovni pajek v prvem koraku pregleda osnovne spletne strani za vsako domeno. V tem koraku dopolni tabelo *crawldb.site* in url doda v tabelo-strani ter tako dobi začetne strani v Frontierju.

Table 1: Statistika strani gov.si domen. Opomba: html strani predstavljajo obdelane Html strani in Binary določene binarne strani.

	Vse	Html	Frontier	Binary	Duplicates	Robots
absolutno	17810	3050	10031	2967	1666	96
na stran				0.97	0.55	0.03

Table 2: Statistika binarnih strani državnih strani.

	Vse	PDF	PPT	PPTX	DOC	DOCX
absolutno	2967	1709	0	0	472	0
na stran	0.97	0.56	0	0	0.15	0

Nato preveri vsebino datoteke robots.txt in sitemap.xml (če ta obstaja). S pomočjo pridobljenih linkov ustrezno dopolni tabelo strani, pri tem zaradi večje preglednosti nedovoljenim linkom v tabeli strani priredi kodo strani ROBOTS (To smo dodali v tabelo tipov strani).

Iskanje slik in linkov Za izločanje vsebine strani smo uporabili knjižnico *selenium*. Z njeno pomočjo smo zbrali iz strani slike in linke. Ustrezno pa smo pri temo posodabljali tudi podatkovno bazo.

Binarne datoteke Binarne datoteke smo določali na podlagi MIME tipov (Multipurpose Internet Mail Extensions). Glede na tip vsebine smo s pomočjo razbijanja textov prišli do tipa strani. V primeru, da je imela stran html vsebino smo jo dodali v podatkovno bazo kot FRONTIER, če pa je ustrezala kateremu izmed izbranih binarnih tipov (npr. pdf, doc), pa smo jo dodali v bazo kot tip BINARY in dopolnili tudi tabelo $page - data$.

Večprocesno delovanje Zaradi narave pythona je večprocesno delovanje nekoliko bolj zapleteno implementirati. Ker se s tem do sedaj v ekipi še nismo srečali, smo za reševanje problemov pri tej nalogi porabili kar nekaj časa. Žal pa nam problemov ni uspelo rešiti do konca. Ta del kode ostaja zakomentiran v celotni kodi.

Statistika

Domene gov.si: evem.gov.si, e-uprava.gov.si, podatki.gov.si, e-prostor.gov.si, mizs.gov.si, mddsz.gov.si, mf.gov.si, www.mgrt.gov.si. Čas delovanja programa: 7 h

Poljubne domene: ucilnica.fmf.uni-lj.si, 21run.com, dibikibi.com. Čas delovanja programa: 7 h

Table 3: Statistika slik državnih strani.

Images	Images per page
767	0.18

Table 4: Statistika strani poljubnih domen.

	Vse	Html	Frontier	Binary	Duplicates	Robots
absolutno	24653	4243	7329	22	13022	2
na stran				0.05	3.07	0.00

Table 5: Statistika binarnih strani pri poljubnih domenah.

	Vse	PDF	PPT	PPTX	DOC	DOCX
absolutno	22	15	2	0	4	1
na stran	0.05	0.003				

Table 6: Statistika slik pri poljubnih domenah.

Images	Images per page
7	0,0016

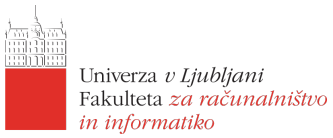

Figure 1: Primer prenešene slike.

Zaključek Izdelan spletni pajek ima še veliko pomankljivosti, tako da bo treba prebrati še precej literature in izvesti raznih testov, če ga bomo želeli vsaj do neke mere izpopolniti.