Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Jakob Dekleva, Matjaž Bevc in Urban Poljšak

Poročilo za drugo domačo nalogo (Data extraction from the Web)

Domača naloga pri predmetu Iskanje in ekstrakcija podatkov s spleta

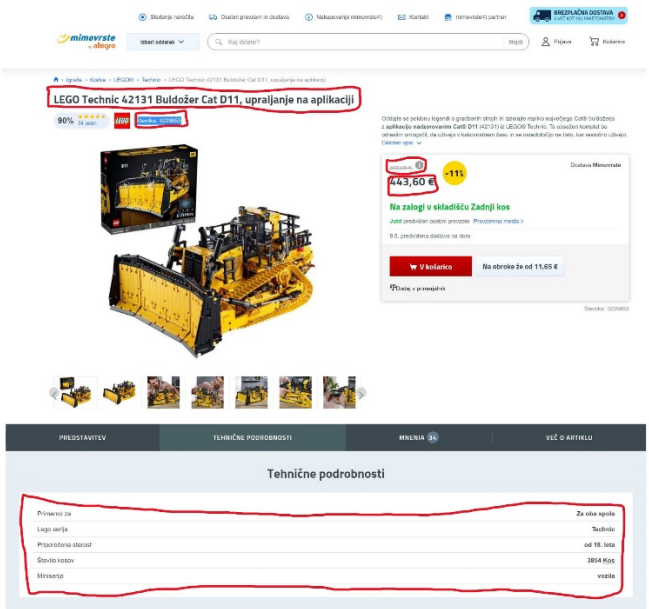
MENTORJA: prof. dr. Marko Bajec, doc. dr. Slavko Žitnik

1. Uvod

V tem poročilu bomo opisali izdelavo treh načinov ekstrakcije podatkov iz spleta: regularni izrazi, xpath in implementacija splošnega algoritma Webstemmer. Izbrane spletne strani so overstock.com, rtv.si in mimovrste.si.

2. Opis izbranih spletnih strani

Za dodatne dve strani smo izbrali produkte iz spletne strani mimovrste.si. Iz spletne stani smo ekstrahirali: naslov, številko izdelka, ceno, znižano ceno, dostopnost in tehnične lastnosti izdelka. Parametri so označeni na sliki 1.



Slika 1: ekstrahirani podatki

3. Implementacija regularnih izrazov in xpath-a

Regular Expressions (regex) omogočajo relativno učinkovito iskanje želenih elementov v HTML kodi, čeprav orodje ni zasnovano za to uporabo. XPath je orodje, posebej zasnovano za iskanje elementov v XML in HTML strukturah. Omogoča enostavno definiranje lokacije elementa v drevesni strukturi in po potrebi iskanje njegovih pod-elementov. Iskanje običajno poteka tako, da definiramo tip in prepoznavno lastnost želenega elementa.

Slika 2: Implementacija regularnih izrazov in xpath

4. Implementacija algoritma Webstemmer

Funkcija webstemmer deluje tako, da sprejme dve HTML strani ter pragove za; merjenje podobnosti, odkrivanje naslova in glavnega besedila. Razčleni strani, izračuna matriko podobnosti in združi podobne strani v grozde. Nato odstrani statične bloke in odkrije naslov ter glavno besedilo znotraj grozdov. Rezultate pretvori v JSON objekt in jih vrne kot končni rezultat.

Psevdokoda rešitve

1. webStemmer(html_page1, html_page2, sim_threshold, diff_threshold, title_threshold, main_text_threshold)
   - wrapper ← []
2. Razčlenjevanje strani:
   - parsed_pages ← parse_page(html) za vsako html stran
3. Izračun podobnosti:
   - sim_matrix ← compute_similarity_matrix(parsed_pages)
4. Grupiranje in ustvarjanje vzorca postavitve:
   - clusters ← cluster_pages(sim_matrix, sim_threshold)
5. Obdelava vsakega grozda:
   - Za vsak grozd:
     - Izračunaj diff_scores za bloke v grozdu
     - Odstrani statične bloke (diff_scores > diff_threshold)
     - Poišči naslov in glavno besedilo (upoštevaj title_threshold, main_text_threshold)
     - Dodaj rezultate v wrapper

6. Vrni wrapper kot JSON niz

Slika 3: izhodni wrapper za Overstock, RTV in Mimovrste