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

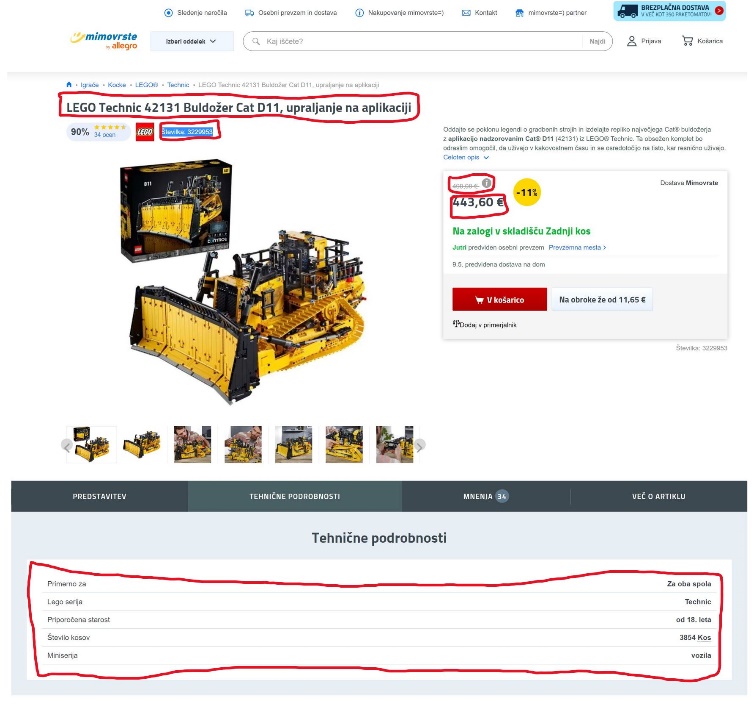
MENTORJA: prof. dr. Marko Bajec, doc. dr. Slavko Žitnik

# 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.

# Opis izbranih spletnih strani

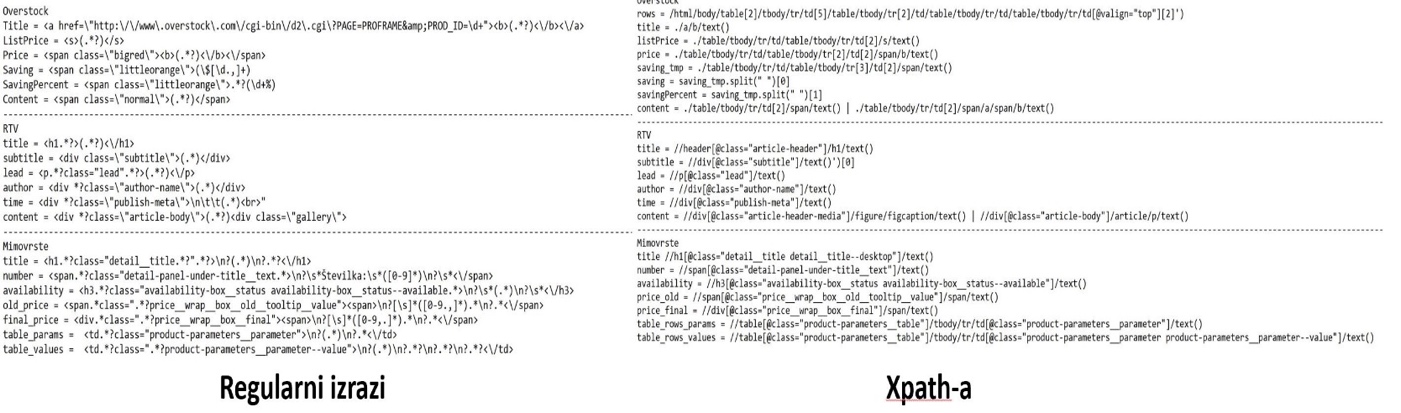
Za dodatne dve strani smo izbrali produkte iz spletne strani mimovrste.si. Iz spletne stani smo ekstrahirali: naslov, številko izdelka, ceno, znižano ceno, dostopnost in tehnične lastnosti izdelka. Parametri so označeni na sliki 1.



Slika : ekstrahirani podatki

# 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 : Implementacija regularnih izrazov in xpath

# 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 ← []

1. Razčlenjevanje strani:

* parsed\_pages ← parse\_page(html) za vsako html stran

1. Izračun podobnosti:

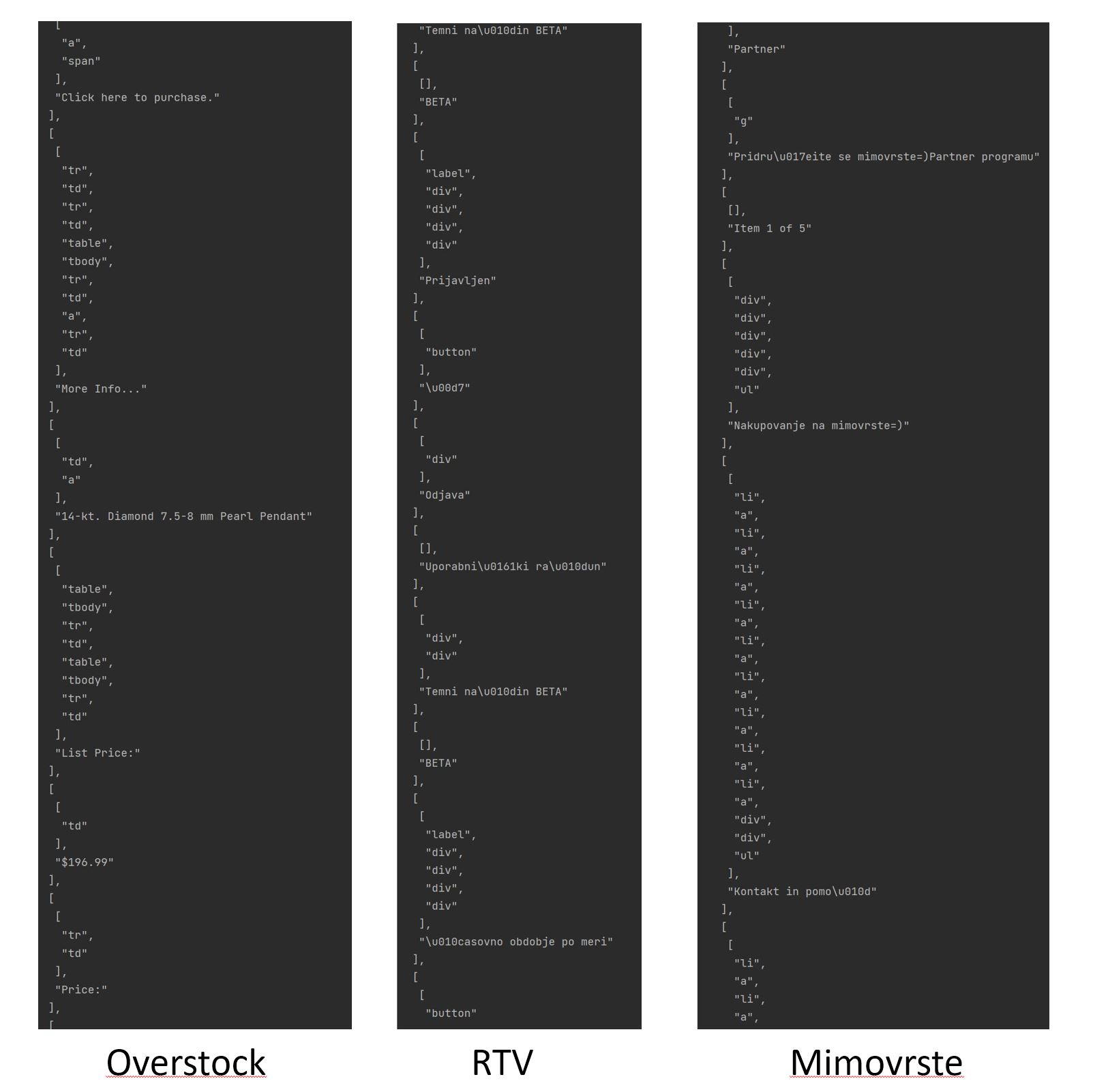
* sim\_matrix ← compute\_similarity\_matrix(parsed\_pages)

1. Grupiranje in ustvarjanje vzorca postavitve:

* clusters ← cluster\_pages(sim\_matrix, sim\_threshold)

1. 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

1. Vrni wrapper kot JSON niz

Slika : izhodni wrapper za Overstock, RTV in Mimovrste