Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Beti Piščanec, Sebastjan Mevlja in Lan Zukanović

Ekstrakcija strukturiranih podatkov s spleta

Seminarska naloga pri predmetu Iskanje in ekstrakcija podatkov s spleta

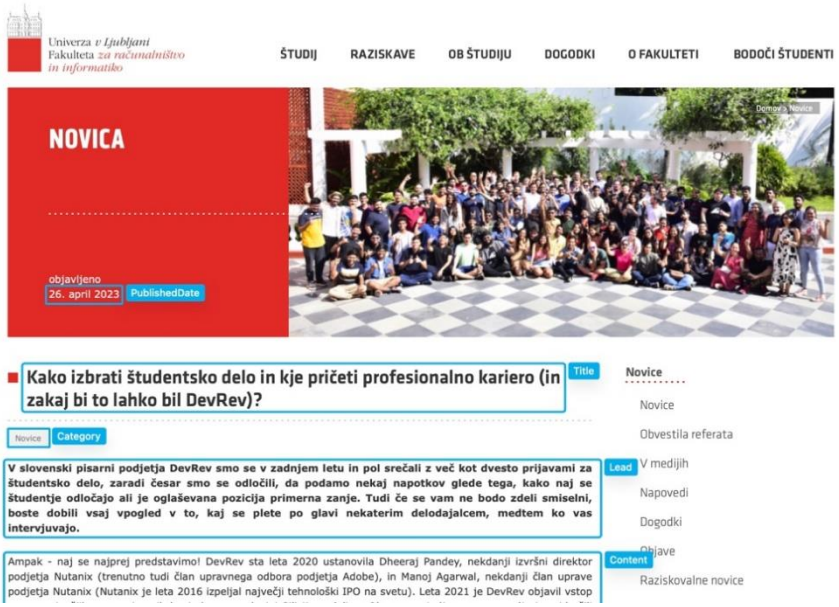
MENTOR: doc. dr. Slavko Žitnik

1. Uvod

Seminarska naloga opisuje implementacijo treh različnih algoritmov za ekstrakcijo strukturiranih podatkov s spleta.

2. Izbrane spletne strani

Kot lasten primer smo izbrali dve novici iz spletne strani univerze. Iz strani lahko razberemo naslov, datum objave, kategorijo, uvod in vsebino.



3. Regex

Overstock.com:

- **Title**: `/<td valign="top">[\n\s]*?<a\s.*?><b>(.*?)</b></a>/`
- **List price**: `/<td align="left"\s.*?>[\n\s]*?<s>(.*?)</s>/`
- **Price**: `/<td align="left"\s.*?>[\n\s]*?<span class="bigred"><b>(.*?)</b></span>/`
- **Saving and saving percent**: `/<td align="left"\s.*?>[\n\s]*?<span class="littleorange">(\$.*?)\s\((.*?)\)</span>/`
- **Content**: `/<span class="normal">([\S\s]*?)[\s]*<br>/`

Rtv.si:

- **Author**: `/<div class="author-name">(.*?)</div>/`
- **Published time**: `/<div class="publish-meta">[\n\s\t]*(.*?)<br>/`

- **Title**: `/<h1>(.*?)</h1>/`
- **Subtitle**: `/<div class="subtitle">(.*?)</div>/`
- **Lead**: `/<p class="lead">(.*?)</p>/`
- **Content**: `/^(?:.*)<div class="article-body">|<div class="gallery">(?:.*)$|<script\b[^<]*(?:(?!<\/script>)<[^<]*)*<\/script>|<[^>]*>/s`

Fri.uni-lj.si:

- **Published date**: `/<div class="text">[\n\s\S]*?<br>(.*?)[\n\s]*?<\/div>/`
- **Title**: `/<div class="heading-article\snovica-title">[\n\s]*<ul>[\n\s]*<li\s.*?>(.*)<\/li>/`
- **Category**: `/<span class="kategorija">(.*?)<\/span>/`
- **Lead**: `/<div class="novica-content">[\s\S]*?<p\s?(?:class="rtejustify")?>(?:<.*?>)*(.*?)(?:<.*?>)*<\/p>/`
- **Content**: `/^(?:.*)<div class="novica-content">[\n\s\S]*?<br>[\n\s]*|<\/div>(?:.*)$|<[^>]*>|.\n{2,}/s`

4. XPath

Overstock.com:

XPath za korenski element seznama izdelkov:
`/html/body/table[2]/tbody/tr[1]/td[5]/table/tbody/tr[2]/td/table/tbody/tr/td/table/tbody/tr[@bgcolor]`

- **Title**: `./td[2]/a/b/text()`
- **List price**: `./td[2]/table/tbody/tr/td[1]/table/tbody/tr[1]/td[2]/s/text()`
- **Price**: `./td[2]/table/tbody/tr/td[1]/table/tbody/tr[2]/td[2]/span/b/text()`
- **Saving and saving percent**: `./td[2]/table/tbody/tr/td[1]/table/tbody/tr[3]/td[2]/span/text()`
- **Content**: `./td[2]/table/tbody/tr/td[2]/span/text()`

Rtv.si:

- **Author**: `//*[@id="main-container"]/div[3]/div/div[1]/div[1]/div`
- **Published time**: `//*[@id="main-container"]/div[3]/div/div[1]/div[2]/text()[1]`
- **Title**: `//*[@id="main-container"]/div[3]/div/header/h1`
- **Subtitle**: `//*[@id="main-container"]/div[3]/div/header/div[2]`
- **Lead**: `//*[@id="main-container"]/div[3]/div/header/p`
- **Content**: `//*[@id="main-container"]/div[3]/div/div[2]/descendant::*[not(name() = "script")]/text()`

Fri.uni-lj.si:

- **Published date:** `//*[@id="banner-header"]/div[2]/div/div[3]/text()[2]`
- **Title:** `//*[@id="katedre-container"]/div[2]/span`
- **Category:** `//*[@id="katedre-container"]/div[1]/div[1]/ul/li`
- **Lead:** `//*[@id="katedre-container"]/div[3]/p[1]/strong/span/span/span` **ali** `//*[@id="katedre-container"]/div[3]/p[1]/strong`
- **Content:** `//*[@id="katedre-container"]/div[3]/*[position() > 2]`

5. RoadRunner

Pri implementaciji algoritma RoadRunner smo si pomagali s predlaganim člankom [1]. Pred zagonom algoritma smo iz vhodnih HTML dokumentov odstranili odvečne značke in komentarje, ter iz HTML elementov ustvarili žetone. Generirane ovojnice vseh treh strani so smiselne. Delovanje smo preverili tudi z dvema avtomatskima testoma, ki primerjata naš rezultat s primeroma iz članka.

5.1 Psevdokoda delovanja algoritma

If the length of both pages is the same or one of the indexes is greater or equal length of the page:

    Return current wrapper

If the first tokens of both pages are equal:

    Add it to the wrapper, increment both indexes and recursively call the function again

If both tokens are database field:

    Add "#PCDATA" to the wrapper, increment both indexes and recursively call the function again

Check for iterators:

    Fix the wrapper, add iterator tags , increment page index and recursively call the function again

Check for optionals:

    Add optional tags , increment page index and recursively call the function again

Return empty wrapper

-

5.2 Ovojnica Overstock

<html> <head> <title> Overstock.com, save up to 80% every day! </title> <link> </link> <link> </link> </head> <body> <input> </input> <table> <tbody> <tr> <td> </td> <td> <table> <tbody> <tr> <td> <table> <tbody> <tr> <td> <a> <img> </td> <td> <img> </td> <td> <a> <img> </td> </tr> </tbody> </table> <map> <area> </area> <area> </area> <area> </area> <area> </area> <area> </area> </map> </td> </tr> <tr> <td> <a> <img> <a> <img> <a> <img> <a> <img> <a> <img> <a> <img> <a> <img> <a> <img> </td> </tr> </tbody> </table> </td> </tr> <tr> <td> <img> </td> </tr> <form> </form> <tr> <td> <table> <tbody> <tr> <td> <table> <tbody> <tr> <td> Search: <select> <option> All Stores </option> <option> Home & Garden </option> <option> Electronics & Computers </option> <option> Books, Movies, CDs, Games </option> <option> Jewelry, Watches & Gifts </option> <option> Sports, Travel & Toys </option> <option> Worldstock </option> <option> Apparel, Shoes & Access. </option> </select> </td> <td> </td> <td> <input> </input> </td> <td> <input> </input> </td> </tr> </tbody> </table> </td> <td> <img>
 </br> <a> <img> </td> <td> <img> </td> </tr> </tbody> </table> </td> </tr> <tr> <td> <img> </td> </tr> </tbody> </table> <table> <tbody> <tr> <td> <img> </td> <td>
 </br> <table> <tbody> <tr> <td> <table> <tbody> (<tr><td></td><td><a>#PCDATA</td></tr>)+ </tbody> </table> </td> </tr> </tbody> </table>
 </br>   Stores
 </br> <table> <tbody> <tr> <td> <a> Apparel, Shoes & Access. </td> </tr> <tr> <td> <a> Books, Movies, CDs, Games </td> </tr> <tr> <td> <a> Electronics & Computers </td> </tr> <tr> <td> <a> Home & Garden </td> </tr> <tr> <td> <a> Jewelry, Watches & Gifts </td> </tr> <tr> <td> <a> Sports, Travel & Toys </td> </tr> <tr> <td> <a> Worldstock </td> </tr> </tbody> </table>
 </br>   New Stock
 </br> <table> <tbody> <tr> <td> <a> Ralph Lauren $29.95 </td> </tr> <tr> <td> <a> Ben Sherman 53% off </td> </tr> <tr> <td> <a> Pre-order Harry Potter DVD </td> </tr> <tr> <td> <a> HP 2GHz System $499 </td> </tr> <tr> <td> <a> New Items within 7 Days </td> </tr> </tbody> </table>
 </br>   Customer Service
 </br> <table> <tbody> <tr> <td> <a> Shopping Cart & Checkout </td> </tr> <tr> <td> <a> Track Your Order </td> </tr> <tr> <td> <a> Your Account </td> </tr> <tr> <td> <a> Help & FAQ </td> </tr> <tr> <td> <a> Best Price Guarantee </td> </tr> </tbody> </table>
 </br>   About Us
 </br> <table> <tbody> <tr> <td> <a> About Us </td> </tr> <tr> <td> <a> Privacy & Security </td> </tr> <tr> <td> <a> Terms & Conditions </td> </tr> <tr> <td> <a> Become An Affiliate </td> </tr> <tr> <td> <a> Business Purchases </td> </tr> <tr> <td> <a> Have Products to Sell? </td> </tr> <tr> <td> <a> Investor Relations </td> </tr> </tbody> </table> <img>
 </br> </td> <td> <img> </td> <td> <img> </td> <td>
 </br> <a> Jewelry, Watches & Gifts > <a> Jewelry > <a> #PCDATA > View All
 </br> <table> <tbody> <tr> <td>
 </br> <table> <tbody> <tr> <td> <table> <tbody> (<tr><td>#PCDATA</td><td>#PCDATA#PCDATA(30of296|)? <a>#PCDATA|<a>#PCDATA|<a>#PCDATA(|<a>Quantity<a/>|<a>Markdowns<a/>)? </td></tr>)+ </tbody> </table> </td> </tr> </tbody> </table> </td> </tr> <tr> <td> <table> <tbody> <tr> <td> <table> <tbody> <tr> (<td>#PCDATA(

<a>Click here to purchase.)? (<a/>

)? </td>)+ </tr> </tbody> </table> </td> </tr> <tr> <td> <img> </td> </tr> <tr> <td> <table> <tbody> <tr> <td> <a> <img> </td> </tr> <tr> <td> <a> More Info... </td> </tr> </tbody> </table> </td> <td> <a> #PCDATA
 </br> <table> <tbody> <tr> <td> <table> <tbody> <tr> <td> List Price: </td> <td> <s> #PCDATA </s> </td> </tr> <tr> <td> Price: </td> <td>   #PCDATA   </td> </tr> <tr> <td> You Save: </td> <td> #PCDATA </td> </tr> </tbody> </table> </td> <td> #PCDATA
 </br> <a>   Click here to purchase.
 </br> </td> </tr> </tbody> </table> </td> </tr> <tr> <td> <img> </td> </tr> <tr> <td> <table> <tbody> <tr> <td> <a> <img> </td> </tr> <tr> (<td>#PCDATA(

<a>Click here to purchase.)? (<a/>

)? </td>)+ </tr> </tbody> </table> </td> </tr> <tr> <td> <img> </td> </tr> <tr> <td> <table> <tbody> <tr> <td> <a> <img> </td> </tr> <tr> ()+ </tr> <tr> ()+ </tr> <tr> ()+ </tr> </tbody> </table> (<td/><tr/><tbody/><table/><td/><tr/><tbody/><table/><td/><tr/><tbody/><table/><td/><tr/><tbody/><table/><td/><tr/><tbody/><table/><td/><tr/><tbody/><table/><td/><tr/><tbody/><table/><td/><tr/><tbody/><table/><td/><tr/><tbody/><table/><td/><tr/><tbody/><table/><td/><tr/><tbody/><table/><td/><tr/><tbody/><table/><td/><tr/><tbody/><table/><td/><tr/><tbody/><table/><td/><tr/><tbody/><table/><td/><tr/><tbody/><table/><td/><tr/><tbody/><table/><td/><tr/><tbody/><table/>)? <map> <area> </area> <area> </area> <area> </area> </map> </body> </html>

5.3 Ovojnica Rtvslo

```
<html> <head> <link> </link> <link> </link> <link> </link> <link> </link> <link> </link> <link> </link> <title> #PCDATA </title> <link>
</link> <link> </link> <link> </link> <link> </link> <link> </link> <link> </link> <link> </link> <link> </link> <link> </link> <link>
</link> <link> </link> <link> </link> <link> </link> <link> </link> <link> </link> <link> </link> <link> </link> <link> </link> <link>
</link> <link> </link> <link> </link> </head> <body> <div> <div> <div> <iframe> </iframe> </div> <div> </div> </div> </div> <div> <div>
<div> <div> <div> <a> RTV SLO </a> </div> <ul> <li> <a> Radio <span> </span> </a> <div> <div> <a> <img> </a> <a> <img> </a> <a> <img>
</a> <a> <img> </a> <a> <img> </a> <a> <img> </a> <a> <img> </a> <a> <img> </a> </div> </div> </li> <li> <a> Televizija </a> </li> <li>
<a> O RTV </a> </li> </ul> </div> <div> <div> <a> 4D </a> <ul> <li> <a> Spored </a> </li> <li> <a> V živo </a> <div> <div> <div> <a>
<img> </a> </div> <div> <a> <img> </a> </div> <div> <a> <img> </a> </div> <div> <a> <img> </a> </div> <div> <a> <img> </a> </div> <div>
<a> <img> </a> </div> <div> <a> <img> </a> </div> <div> <a> <img> </a> </div> <div> <a> <img> </a> </div> <div> <a> <img> </a> </div>
<div> <a> <img> </a> </div> <div> <a> <img> </a> </div> <div> <a> <img> </a> </div> <div> <a> <img> </a> </div> </div> </div> </li> <li>
<a> Arhiv </a> </li> </ul> </div> </div> <div> <ul> <li> <a> <img> <img> </a> </li> <li> <a> <img> <img> </a> </li> <li> <a> <img> <img>
</a> </li> <li> <a> <img> <img> </a> </li> <li> <a> <img> <img> </a> </li> </ul> <form> <input> </input> </form> <div> <a> <span> Prijava
</span> </a> <div> <form> <input> </input> <div> <label> Uporabniško ime: </label> <input> </input> </div> <div> <label> Geslo: </label>
<input> </input> </div> <div> <button> Prijavi </button> <a> Registracija </a> </div> </form> <a> Pozabljeno geslo? </a> <div> <div>
<span> Temni način <span> BETA </span> </span> <label> <input> </input> <span> </span> </label> </div> </div> </div> </div> <div> <a>
</a> <div> <h2> </h2> <a> Odjava </a> <a> Uporabniški račun </a> <div> <div> <span> Temni način <span> BETA </span> </span> <label>
<input> </input> <span> </span> </label> </div> </div> </div> </div> </div> </div> </div> </div> <div> <div> <div> <div> <div> <div>
</div> <div> </div> <div> </div> </div> <div> <a> RTVSLO.si </a> </div> <ul> <li> <a> Slovenija </a> </li> <li> <a> Svet </a> </li> <li>
<a> Šport </a> </li> <li> <a> Kultura </a> </li> <li> <a> Življenjski slog </a> </li> <li> <a> Svet zabave </a> </li> </ul> <div> <div>
Iskanje </div> <form> <input> </input> <button> </button> <div> </div> </form> </div> <div> <div> <span> </span> </div> <div> <a> </a>
</div> <div> <a> </a> </div> <div> <div> <span> </span> <span> </span> <span> </span> <span> </span> </div> <span> Kazalo </span> </div>
</div> </div> </div> </div> <div> <div> <div> <div> <div> </div> <div> <div> <div> <h3> Predlogi </h3> <div> <div> Odbojka </div> <div>
Hokej </div> <div> Tenis </div> </div> </div> <div> <h3> Rezultati iskanja </h3> <div> </div> </div> </div> <p> Zaradi testiranja je
dodanih umetnih 2 sekunde delaya. </p> </div> </div> </div> </div> </div> <div> <div> <div> <div> <a> <img> </a> </div> <div> <h5> <span>
</span> <span> <a> </a> </span> <a> </a> </h5> <p> </p> </div> </div> </div> </div> </div> <div> <div> (<span><span/><div/><div/>)? <div>
(<div/>)? <div> (<div/><div/>)? <div> <div> (<div/><div/>)?(<div>(<div>)? </div>)+
(<article><figure><a><img><a/><figcaption><span><span/>XC40 se v mestu odlično zlije z okolico. Foto: David
Šavli<figcaption/><figure/><div/><div/><h1>Audi A6 50 TDI quattro: nemir v premijskem razredu<h1/><div>Test nove
generacije<div/><div/><div/>)?(<p>#PCDATA</p>)+ (<header/><div><div><div>Miha Merljak<div/><div/><figure>)? <a> (<img>)? </a>
(<div/><figure><h5>Poudarki<h5/><ul><li>Zaradi videza se mu na avtocesti spoštljivo umikajo<li/><ul/><figcaption><span><span/>Foto: David
Šavli<figcaption/>)? </figure> (<div/><article><figure><div>Oglas<div/><div/><figure/>)? <p> (XC 40 je od tal odmaknjen konkretnih 21 cm,
a sta vzmetenje in krmilni mehanizem tako nastavljena, da ponuja tudi v hitro odpeljanih ovinkih zelo dolgo nevtralno in predvidljivo
lego. V premeru preskušanega modela, ki je imel v paketu R design vzmetenje še nekoliko bolj trdo, se je to samo še bolj potrdilo, a je v
tem primeru treba računati na manj udobno vožnjo čez različne asfaltne grbine. Podoben razmislek velja opraviti tudi pri izbiri
motorja.)? </p> <p> #PCDATA </p> <p> (XC 40 je s čvrsto gradnjo, funkcionalno in udobno kabino ter številnimi asistenčnimi sistemi in
izstopajočim skandinavskim dizajnom v premišljenem trenutku vstopil na trg modnih mestnih terencev, v katerem se brez ene same sence
dvoma suvereno postavi med najdražje in najbolj premijske v mestu.<p/><strong>Novo poglavje<br><br/><strong/>Vozniški delovni prostor je
novo poglavje digitalne dobe, z dvema ogromnima zaslonoma, ki tako kot naprednejši telefoni dregnejo blazinice vaših prstov, kot se
sprehajate po steklu. A še bolj se nam zdi pomembno, da so osnovna stikala tam, kjer jih pričakujete. Najprej so torej zagotovili
enostavno osnovo, tisti bolj "advanced" vozniki pa si lahko nato vse skupaj še veliko bolj prilagodijo. Velik korak naprej pri kabinskem
udobju zaznavajo tudi na zadnji klopi, tam je prostora v vseh smereh precej več.<p/><figure>)? <div> (<iframe><iframe/><div/><figure/>)?
<div> <div> (<div>)?(<figure><a><img></a><figcaption>MMC RTV SLO</figcaption></figure>)+ </div> <div> </div> <div> </div> </div> <div>
<figure> <a> <img> </a> <figcaption> #PCDATA </figcaption> </figure> <figure> <a> <img> </a> <figcaption> #PCDATA </figcaption> </figure>
<figure> <a> <img> </a> <figcaption> #PCDATA </figcaption> </figure> <figure> <a> <img> </a> <figcaption> #PCDATA </figcaption> </figure>
</div> </div> </article> </div> <div> <div> <div> <div> </div> </div> <div> <h4> Zadnje iz sekcije </h4> <div> <div> <div> <a> <img> </a>
</div> <div> <a> #PCDATA </a> </div> </div> </div> <div> <div> <div> <a> <img> </a> </div> <div> <a> #PCDATA </a> </div> </div> </div>
<div> <div> <div> <a> <img> </a> </div> <div> <a> #PCDATA </a> </div> </div> </div> <div> <div> <div> <a> <img> </a> </div> <div> <a>
#PCDATA </a> </div> </div> </div> <div> <div> <div> <a> <img> </a> </div> <div> <a> #PCDATA </a> </div> </div> </div> </div> </div>
</div> <div> <div> (<a>#PCDATA</a>)+ </div> <div> <a> Prijavi napako </a> <div> </div> <div> </div> <div> </div> </div> <figure> <div>
Oglas </div> <div> </div> </figure> </div> </div> </div> <div> (<div/>)?()+ </div> <div> <div> <a> <img> </a> <h3> <span> <a> #PCDATA
</a> </span> <a> #PCDATA </a> </h3> <p> #PCDATA </p> </div> </div> <div> <div> <a> <img> </a> <h3> <span> (<span/>)? <a> #PCDATA </a>
(<span/><a>Bentley se je ob stoletnici spomnil blowerja, enega najbolj ikonskih avtomobilov iz dirkaške preteklosti<a/>)? </h3> <p>
#PCDATA </p> </div> </div> <div> <div> <a> <img> </a> <h3> <span> (<span/>)? <a> #PCDATA </a> (<span/><a>Rekorderska steza na slani
ravnici Bonneville potrebuje rešitev<a/>)? </h3> <p> #PCDATA </p> </div> </div> </div> <a> Več novic </a> </div> <div> <div> <div>
<figure> <div> Oglas </div> <div> <div> <iframe> </iframe> </div> <div> <img> </div> </div> </figure> <figure> <div> Oglas </div> <div>
</div> </figure> </div> </div> </div> <div> <div> <h3> <svg> <g> <circle> </circle> <circle> </circle> <circle> </circle> <circle>
</circle> </g> <g> <circle> </circle> <circle> </circle> <circle> </circle> <circle> </circle> </g> <circle> </circle> <circle> </circle>
</svg> <a> Komentarji <span> #PCDATA </span> </a> <div> <div> Prikazuj komentarje <span> </span> </div> <div> </div> </div> </h3> </div>
<div> <a> Uredi </a> </div> <div> <p> Komentarji so trenutno skriti. <a> Prikaži komentarje </a> . </p> <div> <div> <div> </div> </div>
</div> </div> <link> </link> <link> </link> </div> </div> <footer> <div> <div> <ul> <li> <a> Stari portal </a> </li> <li> <a> Impresum
</a> </li> <li> <a> O RTV Slovenija </a> </li> <li> <a> Varuhinja </a> </li> <li> <a> Kontakti </a> </li> <li> <a> RTV prispevek </a>
</li> <li> <a> DOSTOPNO </a> </li> <li> <a> • </a> </li> <li> <a> About RTV Slovenia </a> </li> </ul> <div> <a> </a> <a> </a> <a> </a>
<a> </a> <a> </a> </div> </div> <div> <div> <a> Nastavitve zasebnosti </a> | <a> Spletni piškotki </a> | <a> Varstvo osebnih podatkov
</a> </div> © MMC RTV Slovenija (ISSN 1581-372X) | 1995 - 2019 | Vse pravice pridržane </div> </div> </footer> <div> <div> <div> <div>
<h5> Prijava </h5> <button> <span> × </span> </button> </div> <div> <form> <input> </input> <div> <label> Uporabniško ime: </label>
<input> </input> </div> <div> <label> Geslo: </label> <input> </input> </div> <div> <button> Prijava </button> <a> Registracija </a>
</div> </form> <a> Pozabljeno geslo? </a> <div> <div> <span> Temni način <span> BETA </span> </span> <label> <input> </input> <span>
</span> </label> </div> </div> </div> </div> </div> </div> <div> <div> <div> <div> <h5> Prijavljen </h5> <button> <span> × </span>
</button> </div> <div> <h2> </h2> <a> Odjava </a> <a> Uporabniški račun </a> <div> <div> <span> Temni način <span> BETA </span> </span>
<label> <input> </input> <span> </span> </label> </div> </div> </div> </div> </div> </div> <div> <div> <div> <div> <h5> Časovno obdobje
po meri </h5> <button> <span> × </span> </button> </div> <div> <div> <label> Od </label> <div> <input> </input> </div> </div> <div>
<label> Do </label> <div> <input> </input> </div> </div> </div> <div> <button> Prikaži rezultate </button> </div> </div> </div> </div>
<div> <div> </div> <div> <div> <div> </div> <div> </div> <div> </div> </div> <div> <div> <div> </div> <button> </button> <button>
</button> <button> </button> <button> </button> <div> <div> <div> <div> </div> </div> </div> </div> </div> <div> <div> </div> </div>
<button> </button> <button> </button> <div> <div> </div> </div> </div> </div> </div> <link> </link> <nav> <a> <span> </span> </a>
<button> <span> </span> </button> <div> <ul> <li> <a> Preusmeritve / aliasi </a> </li> <li> <a> Meniji </a> </li> <li> <a> Uporabniki
</a> <div> <a> Skupine </a> <a> Iskalnik uporabnikov </a> <a> Pravice uporabnikov </a> <a> Moderirani uporabniki </a> </div> </li> <li>
<a> Vsebine </a> <div> <a> Administracija novic </a> <a> Administracija sekcij </a> <a> Administracija slike & video dneva </a> <a>
Administracija anket </a> <a> Administracija nujnih informacij </a> <a> Administracija mailerja </a> </div> </li> <li> <a> File upload
</a> </li> </ul> </div> </nav> <nav> <a> <span> </span> </a> <button> <span> </span> </button> <div> <ul> <li> <a> NOVA novica </a> </li>
<li> <a> STARA novica </a> </li> <li> <a> Uteži FP </a> </li> <li> <a> Admin. novic NOVA </a> </li> <li> <a> Admin. novic STARA </a>
</li> <li> <a> Neaktivne novice </a> </li> <li> <a> Videonovice </a> </li> <li> <a> Upload slik </a> </li> <li> <a> Upload datotek </a>
</li> </ul> </div> </nav> <iframe> </iframe> <img> <iframe> </iframe> <img> </body> <iframe> </iframe> </html>
```

5.4 Ovojnica Fri

<html> <head> <link> </link> <link> </link> <link> </link> <link> </link> <link> </link> <link> </link> <link> </link> <link> </link> <link> </link> <link> </link> <link> </link> <link> </link> <link> </link> <link> </link> <link> </link> <link> </link> <link> </link> <title> #PCDATA </title> </head> <body>   <header>   <a> Facebook   <a> Youtube   <a> Knjižnica   <a> Studis   <a> E-učilnica   <a> Urniki   <a> Intranet   <a> Kontakt   <a> EN   <a> <form> <input> </input> </form>   <a> <img>   <a> Študij   Nazaj   <a> Virtualni informativni dan   <a> Dodiplomski študij   <a>   Računalništvo in informatika   Računalništvo in informatika   <a>   Računalništvo in informatika   Računalništvo in informatika   <a>   Računalništvo in matematika   Računalništvo in matematika   <a>   Multimedija   Multimedija   <a>   Upravna informatika   Upravna informatika   <a>   Delovna praksa   Delovna praksa   <a>   Dvojni študijski program   Dvojni študijski program   <a>   Diplomsko delo   Diplomsko delo   <a>   Prestopanje med programi BUN-RI na BVS-RI   Prestopanje med programi BUN-RI na BVS-RI   <a>   Prehod iz starega v novi program   Prehod iz starega v novi program   <a>   Dodatno leto   Dodatno leto   <a> Magistrski študij   <a>   Računalništvo in informatika   Računalništvo in informatika   <a>   Podatkovne vede   Podatkovne vede   <a>   Multimedija   Multimedija   <a>   Računalništvo in matematika   Računalništvo in matematika   <a>   Kognitivna znanost   Kognitivna znanost   <a>   Uporabna statistika   Uporabna statistika   <a>   Digitalno jezikoslovje   Digitalno jezikoslovje   <a>   Umetna inteligenca   Umetna inteligenca   <a>   Dvojni študijski program   Dvojni študijski program   <a>   Zagovor magistrskega dela   Zagovor magistrskega dela   <a> Doktorski študij   <a>   Računalništvo in informatika   Računalništvo in informatika   <a>   Bioznanosti   Bioznanosti   <a>   Mentorji   Mentorji   <a>   Pogosta vprašanja   Pogosta vprašanja   <a>   Predmetnik pred š.l. 2015/2016   Predmetnik pred š.l. 2015/2016   <a>   Dvojni študijski program   Dvojni študijski program   <a> Mednarodne izmenjave   <a> Študentski status   <a> Študijski koledar   <a> Pravilniki, vloge, ceniki   <a> Prijava na izpite   <a> Vpis   <a> Pogosta vprašanja   <a> Splošni izbirni predmeti   <a> Napotki za pisanje poročila   <a> Štipendije   <a> Študentski referat   <a> Raziskave   Nazaj   <a> Laboratoriji   <a> Raziskovalni projekti   <a> Seminarji, konference in zagovori   <a> Sodelovanje z gospodarstvom   <a> Mednarodna vpetost   <a> Fakulteta in študenti   <a> Raziskovalna infrastruktura   <a> CERN   <a> JRC   <a> Ob študiju   Nazaj   <a> Poletne šole in delavnice   <a> Študentski svet FRI   <a> Karierni centri UL   <a> Mednarodna tekmovanja   <a> Šport   <a> Banda FRIdom   <a> Robo Liga FRI   <a> Študentski izzivi   <a> Alumni   <a> Dogodki <a> O fakulteti   Nazaj   <a> Poslanstvo in vizija   <a> Vodstvo   <a> Organi fakultete   <a> Osebje   <a> Kontakti   <a> Kako do nas   <a> Kje je kaj   <a> Pravilniki, programi, poročila   <a> Katalog informacij javnega značaja   <a> Javna naročila   <a> Kariera na FRI   <a> Medijsko središče   <a> Najem prostora   <a> Založba UL FRI   <a> Bodoči študenti   <a> Knjižnica   <a> Studis   <a> E-učilnica   <a> Urniki   <a> Intranet   <a> Kontakt   <a> Facebook   <a> Youtube   <a> EN   <nav>   <a> Študij   <a> Virtualni informativni dan   <a> Dodiplomski študij <a>   Računalništvo in informatika   univerzitetni program   <a>   Računalništvo in informatika   visokošolski program   <a>   Računalništvo in matematika   univerzitetni interdisciplinarni program   <a>   Multimedija   univerzitetni interdisciplinarni program   <a>   Upravna informatika   univerzitetni interdisciplinarni program   <a>   Delovna praksa   <a>   Dvojni študijski program   <a> Magistrski študij <a>   Računalništvo in informatika   magistrski program   <a>   Podatkovne vede   smer magistrskega programa   <a>   Multimedija   interdisciplinarni magistrski program   <a>   Računalništvo in matematika   interdisciplinarni magistrski program   <a>   Kognitivna znanost   interdisciplinarni magistrski program   <a>   Uporabna statistika   interdisciplinarni magistrski program   <a>   Digitalno jezikoslovje   mednarodni skupni magistrski program   <a>   Umetna inteligenca   mednarodni skupni magistrski program   <a> Doktorski študij <a>   Računalništvo in informatika   Doktorski program   <a>   Bioznanosti   interdisciplinarni doktorski program   <a>   Dvojni študijski program   <a> Mednarodne izmenjave   <a> Študentski status   <a> Študijski koledar   <a> Pravilniki, vloge, ceniki   <a> Prijava na izpite   <a> Vpis   <a> Pogosta vprašanja   <a> Splošni izbirni predmeti   <a> Napotki za pisanje poročila   <a> Študentski referat   <a> Raziskave   <a> Laboratoriji   <a> Raziskovalni projekti   <a> Seminarji, konference in zagovori   <a> Sodelovanje z gospodarstvom   <a> Mednarodna vpetost   <a> Fakulteta in študenti   <a> Raziskovalna infrastruktura   <a> CERN   <a> JRC   <a> Ob študiju   <a> Poletne šole in delavnice   <a> Študentski svet FRI   <a> Karierni centri UL   <a> Mednarodna tekmovanja   <a> Šport   <a> Banda FRIdom   <a> Robo Liga FRI   <a> Študentski izzivi   <a> Alumni   <a> Dogodki   <a> O fakulteti   <a> Poslanstvo in vizija   <a> Vodstvo   <a> Organi fakultete   <a> Osebje   <a> Kontakti   <a> Kako do nas   <a> Kje je kaj   <a> Pravilniki, programi, poročila   <a> Katalog informacij javnega značaja   <a> Javna naročila   <a> Kariera na FRI   <a> Medijsko središče   <a> Najem prostora   <a> Založba UL FRI   <a> Bodoči študenti   </nav> ▯   <a> <form> <input> </input> </form>   </header>   <a> Domov > Novice   novica   objavljeno
 </br> #PCDATA   ((Obvestila referata)?)+ (<a/>)?   (<a>)?((<footer> <a>Študij</a/>)?)+ ( )?   <a> Virtualni informativni dan   <a> Dodiplomski študij   <a> Magistrski študij   <a> Doktorski študij   <a> Mednarodne izmenjave   <a> Študentski status   <a> Študijski koledar   <a> Pravilniki, vloge, ceniki   <a> Prijava na izpite   <a> Vpis   <a> Pogosta vprašanja   <a> Splošni izbirni predmeti   <a> Napotki za pisanje poročila   <a> Štipendije   <a> Študentski referat   <a> Raziskave   <a> Laboratoriji   <a> Raziskovalni projekti   <a> Seminarji, konference in zagovori   <a> Raziskovalna infrastruktura   <a> CERN   <a> JRC   <a> Ob študiju   <a> Poletne šole in delavnice   <a> Študentski svet FRI   <a> Karierni centri UL   <a> Šport   <a> Banda FRIdom   <a> Robo Liga FRI   <a> Študentski izzivi   <a> Alumni   <a> Dogodki   <a> Akademija FRI   <a> Poletna šola FRI   <a> Teden programiranja   <a> Frižider   <a> Izobraževanje za učitelje   <a> Do nosu v blatu   <a> FRI Piškot   <a> Garaža FRI   <a> Koronavirus   <a> Novice   <a> O fakulteti   <a> Bodoči študenti   <a> Kariera na FRI   <a> Prijavi napako   <a> Informacijske storitve   <a> Varstvo oseb. podatkov   <a> Zaupne osebe   <a> Študentski referat   Uradne ure   Ponedeljek, torek, sreda, četrtek: od 10:30   do   12:30   E: <a> studinfo@fri.uni-lj.si   T: 01 / 479 8118   <img>   UNIVERZA V LJUBLJANI   Fakulteta za računalništvo in informatiko   Večna pot 113, 1000 Ljubljana   <a> T: 01 / 479 8000 <a> M: 041 / 309 751   <a> E: <a> dekanat@fri.uni-lj.si   FRI na družbenih omrežjih   <a> <img> <a> <img> <a> <img> <a> <img> <a> <img> <a> <img>   <a> <img>   Vse pravice pridržane © Fakulteta za računalništvo in informatiko (ISSN 2820-2961), 2017-2022   </footer>   Ta stran uporablja piškotke z namenom zagotavljanja najboljše spletne storitve in funkcionalnosti. Soglašate s piškotki? <button> Več informacij </button>   <button> Soglašam </button> <button> Ne soglašam </button>   </body> </html>

5 Literatura

[1] Crescenzi, Valter & Mecca, Giansalvatore & Merialdo, Paolo. (2001). RoadRunner: Towards Automatic Data Extraction from Large Web Sites. https://www.vldb.org/conf/2001/P109.pdf