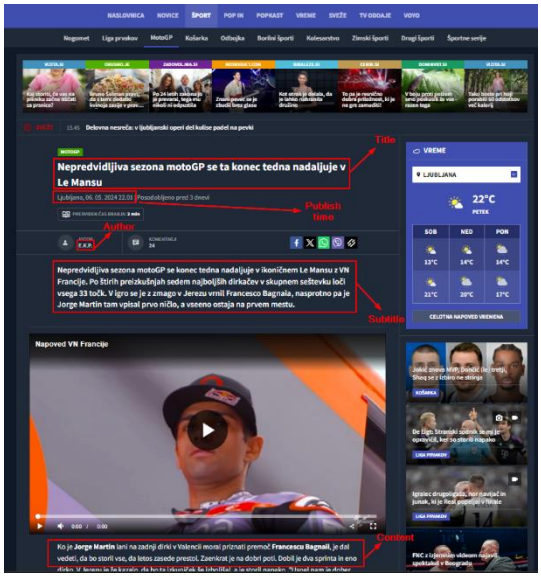
Nejc Rihter, Tadej Lenart in Tomaž Sagaj

Ekstrakcija podatkov s spleta

Projektna naloga pri predmetu Iskanje in ekstrakcija podatkov s spleta

MENTORJA: prof. Dr. Marko Bajec in Timotej Knez

1. Uvod



Slika 1: Prikaz elemetnov spletne strani

V tem poročilu obravnavamo implementacijo treh metod za izvlečenje strukturiranih podatkov iz spletnih strani: uporabo regularnih izrazov, uporabo XPath in avtomatizirano izvlečenje podatkov z algoritmom podobnim Webstemmer. Namen naloge je razviti učinkovite tehnike za pridobivanje podatkov iz treh različnih spletnih domen 24ur, Overstock in Rtvslo.si.

Dve novi spletni strani, ki smo jih dodali sta novici iz spletnega mesta 24ur. Za podatke, ki smo jih hoteli pridobiti smo si izbrali Title, Subtitle, Author, Published Time in Content. V spodnji sliki so ti elementi tudi vizualno prikazani.

2. Implementacija

2.1. Regularni izrazi

Za RTV:

- Title: r'(?<=<h1>)(.*?)(?=<\/h1>)'
- Author: r'(?<=\<div class="author-name"\>).*?(?=\<\/div\>)'
- Published Time: r'(?<=\<div class="publish-meta"\>).*?(?=\<\/div\>)'
- Subtitle: r'(?<=\<div class="subtitle"\>).*?(?=\<\/div\>)'
- Lead: r'(?<=\<p class="lead"\>).*?(?=\<\/p\>)'
- Content: <p(?: class="Body")?>(.*?)<\/p>

Za Overstock:

- Titles: r'<td valign="top">\s*(.*?)
'
- List Prices: r'<td align="right" nowrap="nowrap">List Price:</td>\s*<td align="left" nowrap="nowrap"><s>(.*?)</s></td>'
- Prices: r'<td align="right" nowrap="nowrap">Price:</td>\s*<td align="left" nowrap="nowrap">(.*?)</td>'
- Savings: r'<td align="right" nowrap="nowrap">You Save:</td>\s*<td align="left" nowrap="nowrap">(\$[\d,.]+)'
- Saving Percentages: r'\((\d+%)\)'
- Content: r'(.*?)
'

Za 24ur:

- Title: r'<h1 class=".*?">(.*?)</h1>'
- Author: r' (.*?) '
- Published Time: r'<p class="text-black/60 dark:text-white/60 mb-16 leading-caption">(.*?) \|'
- Subtitle: r' (.*?) '
- Content: r'<p.*?>(.*?)(?=<p|$)'

2.2. XPath

Za RTV:

- Title: //h1/text()
- Author: //div[@class="author"]/div[@class="author-name"]/text()
- Published Time: //div[@class="publish-meta"]/text()
- Subtitle: //div[@class="subtitle"]/text()
- Lead: //p[@class="lead"]/text()
- Content: //p[@class="Body"]/text()

Za Overstock:

- Titles: //td[@valign="top"]/a/b/text()
- List Prices: //td[.//b[contains(text(), 'List Price:')]]/following-sibling::td[1]/s/text()
- Prices: //td[.//b[contains(text(), 'Price:')]]/following-sibling::td[1]/span/b/text()
- Savings and Saving percents: //td/span[@class="littleorange"]/text()
- Content: //td/span[@class="normal"]/text()

Za 24ur:

- Title: //h1/text()
- Author: //div[@class="text-14 font-semibold text-black/80 dark:text-white/80"]/text()
- Published Time: //p[@class="text-black/60 dark:text-white/60 mb-16 leading-caption"]/text()
- Subtitle: normalize-space(//p[@class="text-article-summary font-semibold leading-tight text-black dark:text-white"])
- Content: normalize-space(string(//div[@class="contextual"]//p))

2.3. Webstemmer

Psevdokoda

```
FUNCTION webstemmer(file1, encoding1, file2, encoding2)
   PARSE HTML content from file1 and file2
   EXTRACT layout blocks from both HTML contents
   COMBINE blocks from both files
   CLUSTER combined blocks
   GENERATE CSS selectors for clusters
   PRINT selectors in JSON format
END FUNCTION
```

Mejo za algoritem smo izbrali na podlagi testiranja in tako, da bi rezultati, (ovojnica), ki jo ta generira lahko bili predstavljeni na eni A4 strani. Vrednost praga pri webstemmer algoritma smo prilagajali na osnovi ročnih testiranj. Za spletne strani RTV smo izbrali prag 0.05, za Overstock ter za 24ur pa 0.003. To smo storili v namen temu, da bi iz algoritma izluščili čim boljše rezulate.

3. Zaključek

V našem poročilu smo predstavili in analizirali tri različne metode za ekstrakcijo podatkov s spletnih strani: regularne izraze, XPath in avtomatiziran pristop z uporabo algoritma, podobnega Webstemmerju. Vsaka metoda ima svoje prednosti in slabosti, ki so odvisne od specifičnih zahtev in konteksta uporabe.

4. Rezultati webstemmer algoritma

Za RTV:


```
[
   "div[class=\"['pswp--caption--center']\"]",
   "div[class=\"['dropdown-menu']\"]",
   "div[class=\"['advert-title']\"]",
   "div[class=\"['modal-body']\"]",
   "div[class=\"['modal-header']\"]",
   "div[class=\"['col-12']\"]",
   "div[class=\"['swiper-wrapper']\"][style=\"transform: translate3d(-10896px, 0px, 0px);
transition-duration: 0ms;\"]",
   "div[class=\"['modal-header']\"]",
   "div[class=\"['publish-meta']\"]",
   "div[class=\"['md-news']\"]",
   "div[class=\"['footer-bottom-privacy']\"]",
   "div[class=\"['md-news']\"]",
   "div[class=\"['collapse', 'navbar-collapse']\"][id=\"adminMenuCollapse\"]",
   "div[class=\"['form-group', 'row']\"]",
   "div[class=\"['collapse', 'navbar-collapse']\"][id=\"adminMenuCollapse\"]",
   "p[class=\"['Body']\"]",
   "p[class=\"['published', 'template-published']\"]",
   "p[class=\"['Body']\"]",
   "p",
   "p[class=\"['lead']\"]",
   "title",
   "h1",
   "h2[id=\"user_loggend_in_name_mobile\"]",
   "h3[class=\"['search-section-title']\"]",
   "h3[class=\"['section-title', 'animated-circles-onhover']\"]",
   "h4[class=\"['block-title', 'green']\"]",
   "h5",
   "h5[class=\"['modal-title']\"][id=\"exampleModalLongTitle\"]",
   "div[class=\"['section-heading', 'green']\"]",
   "div[class=\"['swiper-wrapper']\"][style=\"transform: translate3d(-10215px, 0px, 0px);
transition-duration: 0ms;\"]",
   "div[class=\"['right-block', 'similar-articles']\"]",
   "div[id=\"article-comments-anchor\"]",
   "p[class=\"['hidden-comments-notice']\"]",
   "h3[class=\"['section-title', 'animated-circles-onhover']\"]"
]
```


Za Overstock:


```
[ "title", "tr", "td"]
```

Za 24ur:


```
[
  "div[class=\"['flex-wrap', 'flex', 'mx-auto', 'h-full', 'justify-center']\"]",
  "div[id=\"breaking_news_placeholder\"]",
  "div[class=\"['hide-at-start', 'if-user-loggedin', 'flex-col']\"]",
  "div[class=\"['flex', 'justify-between', 'pb-8', 'text-14', 'text-black/80', 'dark:text-
white']\"]",
  "div[class=\"['leading-normal', 'ml-8', 'text-14', 'mb-16', 'font-normal', 'text-primary',
'dark:text-primary-400']\"]",
  "div[class=\"['border-t-2', 'border-white/20', 'flex', 'flex-col', 'pt-2']\"]",
  "div[class=\"['text-white/70', 'hover:text-white/90', 'py-2', 'mb-8', 'default-
transition']\"]",
  "div[class=\"['flex', 'justify-between', 'pb-8', 'text-14', 'text-black/80', 'dark:text-
white/80']\"]",
  "div[class=\"['comment--positive']\"]",
  "div[class=\"['poll-answers-btns', 'pt-16', 'mt-16', 'border-gray-200', 'dark:border-
white/10', 'border-t']\"]",
  "div[class=\"['cursor-pointer', 'bg-black/40', 'rounded-t-8', 'w-32', 'h-32', 'text-white',
'flex', 'justify-center', 'items-center', 'ml-auto', 'transition', 'hover:bg-black/50', 'absolute',
'left-0', '-top-32']\"]",
  "div[class=\"['flex', 'justify-between', 'pb-8', 'text-14', 'text-black/80', 'dark:text-
white/80']\"]",
  "div[class=\"['arrow-up']\"]",
  "div[class=\"['flex', 'justify-between', 'pb-8', 'text-14', 'text-black/80', 'dark:text-
white/80']\"]",
  "p[class=\"['leading-tight', 'text-14', 'pb-16', 'text-black', 'dark:text-white']\"]",
  "p[class=\"['text-14', 'mb-16', 'text-black', 'dark:text-white']\"]",
  "title",
  "h1[class=\"['text-black', 'dark:text-white']\"]",
  "h2[class=\"['text-20', 'font-bold', 'text-black', 'dark:text-white', 'pb-8', 'relative', 'top-
2']\"]",
  "h3[class=\"['font-bold', 'text-20', 'text-black', 'dark:text-white', 'py-8', 'ml-8']\"]",
  "h4[class=\"['h4', 'text-16', 'drop-shadow-text']\"]",
  "h5[class=\"['h5', 'absolute', 'text-left', 'p-16', 'bottom-0', 'drop-shadow-text', 'text-
white', 'z-10', 'font-bold']\"]",
  "h6[class=\"['absolute', 'bottom-0', 'z-10', 'text-white', 'm-10', 'pb-1', 'font-semibold',
'leading-none', 'line-clamp-3', 'text-14', 'drop-shadow-text']\"]",
  "div[class=\"['flex', 'menu__item-inactive']\"]",
  "div[class=\"['text-14', 'font-semibold', 'text-black/80', 'dark:text-white/80']\"]",
  "div[class=\"['flex', 'wrap', 'flex-col', 'xl:flex-row', 'border-b', 'border-black/10',
'dark:border-white/10', 'mb-16']\"]",
  "div[class=\"['label', 'text-white', 'font-bold', 'mb-8', 'mr-2', 'bg-brand-1', 'bg-brand-
sport']\"]",
  "div[class=\"['text-14', 'font-semibold', 'text-black/80', 'dark:text-white/80']\"]",
  "div[class=\"['flex', 'justify-between', 'pb-8', 'text-14', 'text-black/80', 'dark:text-
white/80']\"]"
```