Preddiplomski studij

Računarstvo

Informacija, logika i jezici

1. Domaća zadaća

Ak. g. 2023./2024.

Cilj ove domaće zadaće upoznavanje je s regularnim izrazima. Kroz zadaću potrebno je samostalno napraviti Java konzolnu aplikaciju za parsiranje HTML datoteka.

Tokom izrade ove domaće zadaće svi primjeri s predavanja se slobodno koriste i modificiraju po potrebi.

U programskom jeziku Java potrebno je napraviti konzolnu aplikaciju za parsiranje HTML datoteka te brzo dohvaćanje sadržaja unutar određenih *tagova*. Sve klase potrebno je smjestiti u programski paket `hr.tel.fer.dz1.htmlregex`. Prilikom pokretanja aplikacije korisnik kao argument unosi putanju do željene HTML datoteke koju je potrebno parsirati. Potrebno je provjeriti je li zadana putanja ispravna te potom učitati sadržaj datoteke. Aplikacija čeka korisnički unos te nakon ispravnog unosa vraća traženi sadržaj te ponovno čeka novi unos sve dok se ne unese naredba za kraj. Za uvjete koji su zadani kao naredbe potrebno je izvući sadržaj HTML datoteke koji odgovara uvjetima koristeći regularne izraze. Popis naredbi i njihovi opisi dani su u nastavku:

Naredba	**Opis**
ALL	Vraća cijelu HTML datoteku.
ALL *<tag>*	Vraća sadržaj svih zadanih tagova. Npr. ALL <b> vratit će sadržaj svih <b> tagova. Korisnik sam upisuje tag koji je potrebno vratiti.
ALL email	Vraća popis svih email adresa u formatu: *username@host.domain* (korisničko ime *username* može sadržavati slova i brojke, paziti na valjano ime hosta i domene)
ALL IP	Vraća popis svih IPv4 adresa u formatu: *x.x.x.x* (paziti na dozvoljeni raspon adresa)
ALL date	Vraća popis svih datuma u formatu: *dd/mm/yyyy* (paziti na broj dana i mjeseci, ali nije nužno implementirati prijestupne godine unutar regularnog izraza)
ALL tel	Vraća popis svih telefonskih brojeva u formatu: *385 34 123 4567* (pozivni broj za zemlju je opcionalan, a kao znak za odvajanje brojeva valja prihvatiti – ili razmak)
ONLY *<tag> broj*	Vraća prvih *broj* sadržaja određenog taga <tag>. Npr. ONLY <b> 5 će vratiti sadržaj prvih 5 (ili manje ukoliko ih nema 5) <b> tagova.
ONLY email *broj*	Vraća prvih *broj* email adresa. Npr. ONLY email 5 će vratiti prvih 5 (ili manje ukoliko ih nema 5) email adresa koje se pojavljuju.
ONLY IP *broj*	Vraća prvih *broj* IPv4 adresa. Npr. ONLY IP 5 će vratiti prvih 5 (ili manje ukoliko ih nema 5) IPv4 adresa koje se pojavljuju.
ONLY date *broj*	Vraća prvih *broj* datuma. Npr. ONLY date 5 će vratiti prvih 5 (ili manje ukoliko ih nema 5) datuma koji se pojavljuju.
ONLY tel *broj*	Vraća prvih *broj* telefonskih brojeva. Npr. ONLY tel 5 će vratiti prvih 5 (ili manje ukoliko ih nema 5) telefonskih brojeva koji se pojavljuju.
HELP	Vraća popis opcija koje su implementirane, u formatu sličnom ovoj tablici. Svaka opcija treba imati svoj ID opcije.
REGEX ID *broj*	Vraća izgled regularnog izraza za opciju čiji je ID jednak zadanom argumentu *broj*. Npr. Ako opcija ALL IP ima ID = 5, REGEX ID 5 će vratiti izgled regularnog izraza za opciju ALL IP.
EXIT	Prekida rad programa.

Potrebno je voditi računa kako se u HTML-u tagovi mogu dodatno proširiti s atributima npr: `<input type="hidden" name="_v1param" value="">`. Atribute unutar tagova potrebno je ignorirati te bi navedeni tag trebalo smatrati kao `<input>` tag. Također pretpostavite da se neće dogoditi ispreplitanje HTML tagova npr. `<b>tekst1<i>tekst2</b>tekst3</i>` koji predstavlja **tekst1*tekst2*tekst3**, kao ni ugniježđivanje tagova kao npr. `<p>tekst1<p>tekst2</p>tekst3</p>`. Tagovi koji su zapisani u skraćenom obliku ne sadrže ništa osim atributa te se oni također mogu ignorirati.

Primjer jednostavne HTML datoteke:

```
<html>
	<head>
		<title>Sample web application</title>
	</head>
	<body bgcolor=white>
		<table border="0" cellpadding="10">
			<tr>
				<td>
					<h2>Person 1</h2>
					<ul>
						<li>email: marko.maric@fer.hr</li>
						<li>IP: 252.255.255.255</li>
						<li>DOB: 11/9/2001</li>
						<li>tel: 385-1-222-2222</li>
					</ul>
				</td>
			</tr>
		</table>
		<p>This is the home page for the web application. </p>
	</body>
</html>
```