Preddiplomski studij

Računarstvo

Informacija, logika i jezici

2. Laboratorijska vježba

Ak. g. 2023./2024.

Tokom izrade ove laboratorijske vježbe svi primjeri s predavanja se slobodno koriste i modificiraju po potrebi.

U ovoj laboratorijskoj vježbi upoznati ćemo se sa parsiranjem velikih XML datoteka, te postavljanje upita nad njima. Na predavanju su obrađeni alati XQuery i XPath koji prilikom parsiranja XML-a stvaraju DOM stablo te nad njima izvršavaju zadane upite. Izrada DOM stabla može biti memorijski jako zahtjevan posao, što kod velikih datoteka stvara problem. Za 1 kB XML dokumenta tokom parsiranja stvara se DOM stablo veličine otprilike 10 kB, odnosno možemo reći da zauzima otprilike 10x više memorije. Stoga možemo zaključiti da spremanje DOM stabla u memoriju se može primjenjivati samo kod malih datoteka. Kod velikih datoteka, učitavanje se izvodi dinamički, najčešće korištenjem tokova podataka. Datoteka se predstavi kao tok podataka te se nad njime dinamički izvode upiti kako se ona učitava.

Za stvaranje velikih XML datoteka koje su potrebne za izradu ove laboratorijske vježbe koristi ćemo alat XMark (*https://projects.cwi.nl/xmark/index.html*). XMark je alat koji stvara testne XML datoteke koje sadrže veliku količinu podataka tj. zauzimaju puno mjesta na disku. Potrebno je skinuti alat XMark te kreirati veliku datoteku. Prilikom početka izrade vježbe preporuka je koristiti najmanju moguću stvorenu datoteku koju možete bez problema otvoriti i u tekstualnom pregledniku te vidjeti strukturu XML-a. Kod velikih datoteka, otvaranje u tekstualnom pregledniku nije preporučljivo. Upute za korištenje alata XMark nalaze se na slijedećoj poveznici *https://projects.cwi.nl/xmark/faq.txt*. Prilikom kreiranje većih datoteka pripazite na eksponencijalni rast datoteke ovisno o zastavici `-f` (`/f` na operacijskom sustavu Windows). Za korisnike operacijskog sustava MacOS i/ili nekih inačica operacijskog sustava Linux potrebno je skinuti *source* (`unix.c`) datoteku te ju kompajlirati koristeći npr. *gcc compiler*.

Vaš zadatak je u programskoj jeziku Java napraviti konzolnu aplikaciju za parsiranje XML datoteka i postavljanje upita za filtriranje sadržaja. Nakon pokretanja aplikacija od korisnika traži unos putanje do XML datoteke. Zatim korisnik upisuje naredbu za filtriranje te dobiva rezultat ispisan na ekran, nakon čega je moguće ponovo unijeti novu naredbu za filtriranje i tako sve dok se ne unese naredba za kraj. Datoteka se ne smije odjednom učitati i pohraniti u memoriju! Parsiranje XML datoteke potrebno je implementirati korištenjem StAX parsera (*https://docs.oracle.com/javase/tutorial/jaxp/stax/why.html*). Na slijedećoj poveznici prikazan je jednostavan primjer korištenja StAX parsera *https://www.tutorialspoint.com/java_xml/java_stax_query_document.htm*.

Korisnički upiti za filtriranje sastoje se od 3 parametra. Prvi parametar predstavlja dio XML-a koji želimo, a može biti sadržaj cijelog elementa, sadržaj atributa ili tekstualni sadržaj elementa. Drugi parametar predstavlja ime elementa/atributa a treći parametar N predstavlja prvih N ili manje pojavljivanja traženog uvjeta. Primjerice nakon unosa naredbe `ELEMENT <a> 3`, korisniku će se ispisati prva 3 (ili manje ukoliko ih ima manje od 3) sadržaja elementa <a> sa svim elementima unutar njega. Pokretanjem naredbe `ATTRIBUTE value *` korisniku će se ispisati sve vrijednosti atributa value od svih elemenata koji taj atribut sadrže. Pokretanjem naredbe `TEXT <b> 2` korisniku će se ispisati prva 2 (ili manje ukoliko ih ima manje od 2) teksta elementa <b>. Pokretanjem naredbe EXIT završava se sa radom aplikacije.