Proiect 2
Crawler WEB

an univ. 2018 – 2019

Aplicaţia corespunzătoare celei de a doua componenete de proiect trebuie să implementeze un **Crawler WEB**. Acest modul trebuie să realizeze corect cereri HTTP utilizând versiunea 1.1 a protocolului şi să salveze conţinutul HTML al resursei indicate. Aplicaţia trebuie proiectată şi implementată astfel încât să permită rularea în continuu. În acest sens, modulul *URL Frontier* (coada de explorare) aferent va include iniţial un set de două/trei URL-uri. Rularea în continuu implică următorii paşi:

1. se va prelua următorul URL din coada de explorare şi se procesează astfel încât să se extragă numele domeniului explorat şi URL-ul relativ al resursei dorite;
2. dacă domeniul este la **prima explorare**, atunci se va solicita resursa `/robots.txt`; dacă aceasta există, se trece la pasul 3, dacă nu se continuă cu pasul 4;
3. (*dacă există* `robots.txt`) se verifică clauzele `Disallow` pentru URL-ul relativ curent; dacă REP permite accesul pe resursă, atunci se trece la pasul 4, dacă nu, se trece la următorul URL din coada de explorare;
4. se preia resursa indicată de URL şi se salvează local – pentru fiecare domeniu se va crea un director, apoi, în cadrul acestui director, se va urma structura de directoare din cadrul URL-ului;
5. (*dacă este cazul*) dacă se primeşte un cod `301 Moved Permanently`, atunci se va reface cererea pentru noua locaţie şi se vor actualiza datele deja salvate; orice alt tip de redirect va implica numai reluarea cererii pe noua locaţie a resursei, fără alte actualizări de date;
6. se analizează tag-ul HTML `meta`, `name="robots"`; în cazul în care este permisă extragerea link-urilor incluse în document, se vor extrage aceste link-uri sub forma unui set de URL-uri absolute; din cadrul acestui set se elimină link-urile care nu respectă REP sau care se află deja în coada de explorare;
7. se reia pasul 1.

Cerinţă bonus
Aplicaţia poate respecta următoarele cerinţe bonus:

1. **Cache DNS**: aplicaţia implementează propriul mecanism de *caching* al înregistrărilor DNS.
2. **performanţe**: în mod secvenţial, în cazul în care REP nu impune restricţii de viteză, se doreşte o rată medie de transfer de aproximativ 100 pagini/minut;
3. Crawler-ul WEB va fi **distribuit/paralelizat** pentru a spori performanţele; dacă, de exemplu, se vor utiliza două module de tip *fetcher*, atunci rata medie de transfer dorită va fi de cel puţin 200 pagini/minut.

Predarea celei de a doua componente de proiect se va realiza prin intermediul platformei MOODLE, similar primei componente.

Barem evaluare proiect

Proiectele pornesc ca bază de notare de la **1 punct**.

Criteriu	**Punctaj**
1. Realizarea corectă a cererii pentru a prelua o resursă HTML - *obligatoriu*	
componentă DNS – client „3rd party”	1 punct
sau	
componentă DNS – implementare proprie	**2 puncte**
componentă HTTP – client „3rd party”	1 punct
sau	
componentă HTTP – implementare proprie	**2 puncte**
2. Salvarea completă şi corectă a paginii HTML	
în cadrul unui singur director de lucru, fără a ţine cont de structura URL-urilor	0,5 puncte
în cadrul unei structuri de directoare, ţinând cont de structura URL-urilor	**1 punct**
3. Respectarea pseudo-protocolului REP - *obligatoriu*	
la nivel de pagină	**1 punct**
la nivel de domeniu	**1 punct**
4. Gestionarea corectă a structurilor de tip URL Frontier - *obligatoriu*	**2 puncte**
reconstruirea link-urile care trebuie vizitate	
marcarea corectă a link-urilor care pot fi explorate	
marcarea corectă a link-urilor care au fost deja explorate	

Observaţii

Studenţii vor fi rugaţi să completeze o auto-evaluare a proiectului dezvoltat. Auto-evaluările care nu diferă de nota acordată de titularii de laborator vor primi **suplimentar 0.5 puncte**. În plus, implementările pot atrage punctaj suplimentar (puncte bonus) pentru media finală a disciplinei, astfel:

Implementare şi gestiunea cache DNS – 2 puncte;

Rată de transfer secvenţială – 100 pag./minut – 1 punct;

Paralelizare/distribuire eficientă – 2 puncte.