

UNIVERSITATEA TEHNICĂ „Gheorghe Asachi” din IAȘI  
FACULTATEA DE AUTOMATICĂ ȘI CALCULATOARE  
DOMENIUL: Calculatoare și Tehnologia Informației  
SPECIALIZAREA: Tehnologia Informației

# **Crawler WEB**

PROIECT LA DISCIPLINA  
**REGĂSIREA INFORMAȚIILOR PE WEB**

Studentă: Frentescu Maria

Grupa: 1409A

Profesor îndrumător: Ș.l. dr. ing. Alexandru Archip

Iași, 2020

## Cuprins

Capitolul 1. Prezentarea generală.....	II
1.1. Executia programului.....	II
Capitolul 2. Exemple de execuției.....	III

## Capitolul 1. Prezentarea generală

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

1. se va prelua următorul URL din coada de explorare și se procesează astfel încât să se extragă numele domeniului explorat și 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 resursa, atunci se trece la pasul 4, dacă nu, se trece la următorul URL din coada de explorare;
4. se preia resursa indicată de URL și 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 primește un cod 301 *Moved Permanently*, atunci se va reface cererea pentru noua locație și se vor actualiza datele deja salvate; orice alt tip de redirect va implica numai reluarea cererii pe noua locație 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.

### 1.1. Executia programului

Pentru a începe acțiunea de *crawling*, s-a adăugat pentru început în coada de url-uri, url-ul „http://riweb.tibeica.com/crawl”, și s-a folosit ca user-agent numele RIWEB\_CRAWLER. La executare va începe acțiunea de crawling, pornind de la url-ul inițial. Aplicația folosește resurse de tip robots. S-a implementat o variantă proprie pentru Cache DNS pentru popularea dicționarului cu adrese de domeniu. Aplicația verifică parcurgerea paginilor html, după fiecare domeniu se aplică un delay de o secundă. Programul rulează atâta timp cât încă sunt url-uri în coadă și cât timp nu s-a ajuns la limita de 100 de pagini descărcate.

## Capitolul 2. Exemple de execuției

Exemplu de execuție cu rezultatul ratei de transfer 100pagini/minut:

```
Start
Got: 100 pages in 105 sec
Stop
```

Figura 2.1: Performanțe

Fișierele html salvate în structura de directoare:





























Name	Date modified	Type	Size
 about.html	6/12/2020 11:53 PM	Chrome HTML Document	5 KB
 app-changes.html	6/12/2020 11:53 PM	Chrome HTML Document	5 KB
 contents.html	6/12/2020 11:52 PM	Chrome HTML Document	10 KB
 dir-conn.html	6/12/2020 11:53 PM	Chrome HTML Document	5 KB
 dir-conn-ch.html	6/12/2020 11:53 PM	Chrome HTML Document	5 KB
 directives.html	6/12/2020 11:53 PM	Chrome HTML Document	7 KB
 dir-filter.html	6/12/2020 11:53 PM	Chrome HTML Document	5 KB
 dir-filter-if.html	6/12/2020 11:53 PM	Chrome HTML Document	5 KB
 dir-filter-of.html	6/12/2020 11:53 PM	Chrome HTML Document	5 KB
 dir-handlers.html	6/12/2020 11:53 PM	Chrome HTML Document	6 KB
 dir-handlers-ach.html	6/12/2020 11:53 PM	Chrome HTML Document	6 KB
 dir-handlers-auh.html	6/12/2020 11:53 PM	Chrome HTML Document	7 KB
 dir-handlers-auzh.html	6/12/2020 11:53 PM	Chrome HTML Document	5 KB
 dir-handlers-fuh.html	6/12/2020 11:53 PM	Chrome HTML Document	5 KB
 dir-handlers-hph.html	6/12/2020 11:53 PM	Chrome HTML Document	5 KB
 dir-handlers-pch.html	6/12/2020 11:53 PM	Chrome HTML Document	6 KB
 dir-handlers-ph.html	6/12/2020 11:53 PM	Chrome HTML Document	5 KB
 dir-handlers-pih.html	6/12/2020 11:53 PM	Chrome HTML Document	6 KB
 dir-handlers-plh.html	6/12/2020 11:53 PM	Chrome HTML Document	5 KB
 dir-handlers-prrh.html	6/12/2020 11:53 PM	Chrome HTML Document	7 KB
 dir-handlers-syn.html	6/12/2020 11:53 PM	Chrome HTML Document	7 KB
 dir-handlers-th.html	6/12/2020 11:53 PM	Chrome HTML Document	6 KB
 dir-handlers-tph.html	6/12/2020 11:53 PM	Chrome HTML Document	5 KB
 dir-other.html	6/12/2020 11:53 PM	Chrome HTML Document	5 KB
 dir-other-epd.html	6/12/2020 11:53 PM	Chrome HTML Document	6 KB
 dir-other-ipd.html	6/12/2020 11:53 PM	Chrome HTML Document	7 KB
 dir-other-ipdv.html	6/12/2020 11:53 PM	Chrome HTML Document	7 KB
 dir-other-ipdv.html	6/12/2020 11:53 PM	Chrome HTML Document	6 KB

Figura 2.2: Salvarea paginilor html

### **Auto-evaluare:**

- Baza – **1 punct**
- Realizarea corectă a cererii pentru a prelua o resursă HTML  
componenta HTTP – implementare proprie **2 puncte**  
componenta DNS – implementare proprie **2 puncte**
- Salvarea completă și corectă a paginii HTML în cadrul unei structuri de directoare, ținând cont de structura URL-urilor **1 puncte**
- Respectarea pseudo-protocolului REP **2 puncte**
- Gestionarea corectă a structurilor de tip URL Frontier **2 puncte**
- Bonus: Implementare și gestionarea cache DNS +Rata de transfer secvențială – 100 pag./minut **2-3 puncte**

**Total: 10 (2-3 puncte bonus)**