RIWebCrawler

Proiect RIW

Student: Condriea Stefan-Catalin

Prof. îndrumător: Alexandru Archip

Proiectarea unui crawler web

Proiectul de față este implementat în limbajul Java și implică următorii pași:

- Un url sau o listă de url-uri sunt introduse într-o coadă (se rulează **MainClass** din fisierul cu același nume, care apelează clasa **WebCrawler)**;

- Se preia primul url din coadă (obiectul **queueURLS**din clasa **WebCrawler**)

- Se identifică domeniul și în cazul în care este la prima explorare, se face cerere pentru IP-ul acestuia(Se folosește clasa **InnetAddres** din java), iar dacă a mai fost deja explorat se reține într-un cacheDNS(obiectul **dnsCache** din clasa WebCrawler;)
- Cerere **robots.txt**și se preiau permisiunile; Toate aceste permisiuni, precum și ip-ul domeniului sunt reținute într-un obiect de tip **Domain**); S-a creat un HashMap<String, Domain> în care pentru fiecare domeniu se reține ip-ul și toate permisiunile din robots.txt. În cazul în care avem permisiuni totale sau nu avem permisiuni deloc pe domeniu, s-au creat două câmpuri de tip boolean în clasa **Domain** în care se reține acest lucru. Mai întâi se verifică aceste câmpuri și apoi se trece la parcurgerea listei de permisiuni dissalow.
- Se verifică secțiune **Dissalow** pentru url-ul curent și, în cazul în care avem permisiuni, se descarcă pagina, se preiau link-urile sub formă absolută și se adaugă în coada de explorare, apoi se creează un fișier stocat într-o structură de directoare conform url-ului.(se poate seta folderul rădăcină, însă implicitse creează *D://RIWEB_CRAWLER_DATASET/* un folder ca rădăcină pentru stocarea folderelor/fișierelor)
- Se trece la următorul url din coadă
- Execuția se termină când coada este goală sau după un număr de iterații dat ca parametru(câmpul LIMIT_QUEUE din WebCrawler)

  Link-urile deja vizitate se stochează într-un HashSet
  - Pentru fiecare link reținut din pagină, se verifică să nu fie în lista de link-uri deja explorate sau să fie deja introdus în coada de explorare;
  - La fiecare aproximativ un minut se afișează timpul

- La fiecare iterație se afișează link-ul explorat și dimensiunea cozii de explorare
- Preluarea paginii se face folosind librăria Jsoup*, prin urmare nu există o implementare proprie a unei cereri HTTP.