Laborator nr. 7
Web Crawler secvenţial

an univ. 2018 – 2019

1 Aplicaţie propusă

Pornind de la pseudocod-ul de mai jos, implementaţi un Web Crawler secvenţial simplist. Coada iniţială de URL-uri va conţine o singură adresă: `http://riweb.tibeica.com/crawl/`

Condiţiile de stop sunt:

- coada de lucru este vidă sau
- au fost explorate şi salvate local 100 de pagini web.

Algoritm 1 Pseudocod generic pentru un robot WEB (adaptare după [4])

```
1: Initializeaza coada (Q) cu un set primar de URL-uri
2: while C1: Q contine pagini sau nu a fost atinsa limita maxima de pagini sau de timp do
3:    Fie L un URL din Q
4:    if L == vizitat then
5:       Re-evalueaza C1
6:    end if
7:    Descarca P = continutul lui L
8:    if P != NULL then
9:       Indexeaza P
10:      Extrage din P o lista noua de legaturi N
11:      Adauga N la Q
12:   end if
13: end while
```

Indicaţie: pentru simplitate, se recomandă utilizarea librăriei JSOUP pentru determinarea URL-urilor absolute necesare explorării automate [2]. De asemenea, pentru clientul HTTP necesar aplicaţiei se pot utiliza instanţe ale unor clase specilizate precum `URLConnection` (Java – [3]).

1.1 Aplicaţie bonus

Pornind de la modelul propus de Manning [1], adaptaţi pseudocodul inclus în Algoritmul 1 astfel încât să respecte modelul din Figura 1 (Anexa 1). Aplicaţia dezvoltată va trebui să respecte în mod obligatoriu protocolul REP (atât la nivel de domeniu – fişier `robots.txt`, cât şi la nivel de pagină HTML – tag-ul `meta` → `name="robots"`). Pentru modulele aferente componentelor DNS şi Fetcher se preferă implementări proprii.

Bibliografie

[1] Christopher D. Manning et. al. Introduction to Information Retrieval. `http://nlp.stanford.edu/IR-book/information-retrieval-book.html`, 2009.

[2] Jonathan Hedley. jsoup: Java HTML Parser. `http://jsoup.org`.

[3] Oracle. Reading from and Writing to a URLConnection. `http://docs.oracle.com/javase/tutorial/networking/urls/readingWriting.html`.

[4] Raymond J. Mooney. Information Retrieval and Web Search (note de curs). `http://www.cs.utexas.edu/~mooney/ir-course/`.

Anexa 1 – Modelul generic al unui web crawler secvenţial

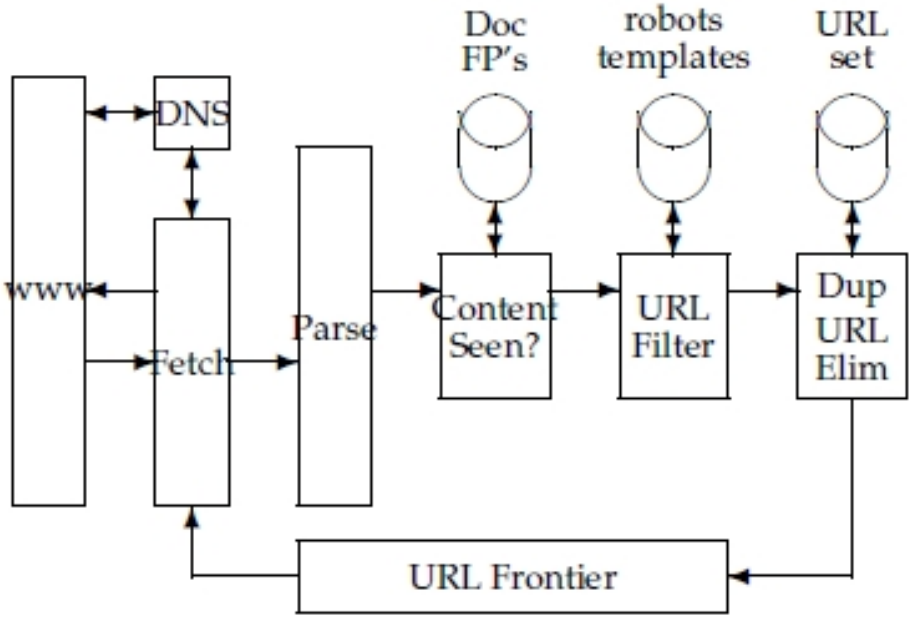

Figura 1: Componentele unui robot WEB (preluare din [1])