Crawler pentru indexare web

(Software Requirements Specifications)

26 Noiembrie, 2020

Echipa de dezvoltare(C-114D):

* Mirea Irina
* Călugăru Teodor
* Ghiță Cristian
* Albu Alexandru

Contents

[1. Scopul documentului 2](#_Toc57297351)

[2. Descrierea generală a produsului 2](#_Toc57297352)

[Misiunea produsului 3](#_Toc57297353)

[Situația curentă 3](#_Toc57297354)

[Contextul produsului 3](#_Toc57297355)

[3. Cerințe funcționale 4](#_Toc57297356)

[Actorii: 4](#_Toc57297357)

[Cazuri de utilizare: 4](#_Toc57297358)

[Dependințe: 6](#_Toc57297359)

[Interfețe: 6](#_Toc57297360)

[4. Cerințe nefuncționale 6](#_Toc57297361)

[Cerințe de interfață 6](#_Toc57297362)

[Cerințe de performanță 7](#_Toc57297363)

[Cerințe de securitate 7](#_Toc57297364)

# Scopul documentului

Documentul are că scop descrierea aplicației de tip crawler web dezvoltată în cadrul grupei C114D.

Documentul prezintă o descriere generală a crawlerului web, descrie cerințele funcționale și nefuncționale ale aplicației, cât și cazurile de utilizare ale aplicației.

Documentul se adresează beneficiarului aplicației, Academia Tehnică Militare.

# Descrierea generală a produsului

## Misiunea produsului

Aplicația va fi utilizată pentru a crea o copie a unor pagini web pentru procesarea ulterioară. Va trebui că la baza să îndeplinească funcția de web crawler, mai exact să descarce și indexeze pagini web. Acest lucru se va realiza recursiv plecând de la niște URL-uri preluate dintr-un fișier. La final, link-urile vor fi înlocuite cu cai către paginile descărcate. Aplicația va putea realiza următoarele operații (pentru procesarea ulterioară) asupra folderelor : Filtrare după tip, Căutare după cuvinte cheie, Limita dimensiune, Creare sitemap. De asemenea, software-ul va conține și un filtru pentru fișierele descărcate bazat pe niște restricții.

## Situația curentă

De-a lungul timpului au fost implementate mai multe soft-uri de tip web crawler, iar dintre acestea se remarcă următoarele:

*Screaming Frog* permite utilizatorilor să analizeze și să realizeze audit tehnic.Poate fi utilizat pentru a accesa gratuit până la 500 de adrese URL.

*Apify* automatizează fluxurile de lucru. Permite utilizatorilor săi accesarea de site-uri web arbitrare folosind browserul Chrome și extragerea de date folosind JavaScript.

*Deepcrawl* este un instrument bazat pe cloud care permite citirea și accesarea conținutului site-ului web. Va permite înțelegerea și monitorizarea problemelor tehnice ale site-ului web pentru a îmbunătăți performanța SEO.

*WildShark SEO Spider Tool* ajută la identificarea paginilor cu etichete de descriere duplicate. Utilizatorii îl pot folosi pentru a găsi titluri duplicate lipsa.

## Contextul produsului

Această aplicație de tip web crawler este utilizată, în principiu, pentru indexarea web. Practic, plecând de la un URL(adresa unei pagini web), se vor descărca în mod recursiv toate paginile web către care există link(se vor descărca astfel noi pagini până când se va ajunge la o pagină ce nu mai are niciun link către altele).

Unde va rula aplicația? Pentru acest lucru se va folosi un utilitar în linie de comandă(command prompt).De asemenea, se dorește ca aplicația să ruleze pe mai multe fire de execuție(thread-uri), scenariul fiind următorul:

* Utilitarul primește un fișier în care se vor afla mai multe URL-uri. Thread-ul X va prelua o parte din aceste URL-uri și va obține toate link-urile care există în cadrul acestora(în mod arborescent va căuta până nu mai exista niciun link).Apoi, pentru fiecare link descoperit se va descărca pagina pe care acesta o indică. .După ce va descărca un URL, thread-ul îl va pune într-o listă de link-uri vizitate. Această listă va fi consultată înaintea fiecărei descărcări(pentru a nu se descărca aceeași pagină de mai multe ori).
* În același timp, thread-ul Y va prelua o altă serie de URL-uri din fișierul de intrare, și așa mai departe. Se va face o împărțire relativ egală a numărului de URL-uri pe thread-uri.

În ceea ce privește fișierele descărcate, vor exista totuși anumite restricții(se vor putea descărca doar anumite tipuri de fișiere, nu orice tip).Pentru a nu descărca decât tipurile de fișiere dorite, se va putea consulta conținutul fișierului robots.txt(în care apar instrucțiuni de execuție pentru crawler-ul web) sau se pot preciza prin intermediul unui fișier de configurare al aplicației tipurile de fișier care trebuie descărcate.

Paginile descărcate vor fi salvate în mai multe foldere. În final, se va putea accesa un astfel de folder și se vor putea face anumite operații asupra fișierelor descărcate, precum filtrarea după tip a fișierelor, căutarea după cuvinte cheie sau limitarea dimensiunii fișierelor.

Se va implementa un mecanism de tratare a tuturor excepțiilor apărute pe parcursul rulării proiectului(FileNotFound, ConnectionFailed etc).

# Cerințe funcționale

## Actorii:

Actorii aplicației vor fi userii crawlerului care doresc să utilizeze funcționalitățile acestuia. Aceștia vor interacționa cu aplicația prin linie de comandă și printr-un fișier .txt dat că parametru, care va conține url-ruile tuturor site-urilor pe care crawlerul trebuie să le viziteze.

## Cazuri de utilizare:

Userii vor putea să:

-Descarce paginile respective, și recursiv până la o anumită adâncime, toate paginile spre care există un link.

-Descarce doar anumite tipuri de fișiere(.docx, .pdf, etc)

-Precizeze dacă crawlerul va verifică conținutul fișierului robots.txt și dacă va ține cont de acesta în căutarea să sau nu.

-Precizeze numărul de threaduri pe care va rula aplicația.

-Precizeze fișierul de log

-Precizeze directorul de download.

Pentru folderul unde se află un site deja descărcat crawlerul va putea:

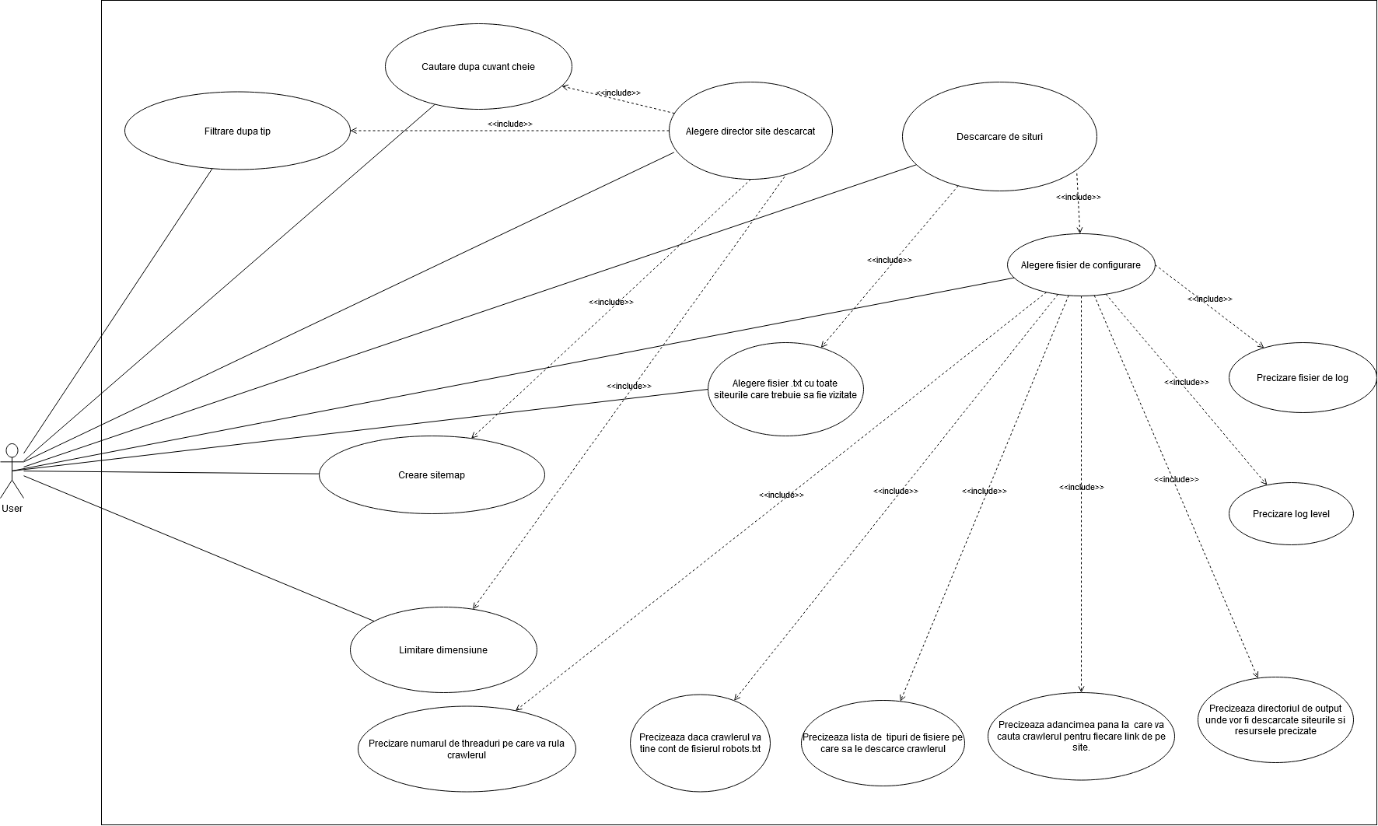
-crea un sitemap

-filtra după tipul fișierelor.

-limita dimensiunea fișierului.

-Caut după cuvinte cheie.

Mai jos este prezentată diagramă de cazuri de utilizare a sistemului:



Aplicația va primi ca parametrii fișierul cu site-urile care trebuie vizitate și un fișier de configurare.

* Fișierul de configurare va avea următorul format:

1. *Nthreads*: (numărul de threaduri) int intre 1 si 50
2. *File\_types*: (tipul fișierelor care vor fi descarcate) stringuri separate prin virgula
3. *Robots.txt*: (daca se va tine cont de robots.txt sau nu) bool
4. *Log\_file*: (Fișierul unde vor fi stocate toate locurile) string
5. *Log\_level*: (Nivelul de printare pentru loguri) int intre 0 si 3

* Intrările din fișierul cu situri vor avea următorul format:

Nume\_site(string):url\_site(string)

* Fișierul de log va conține mesaje 3 tipuri de mesaje:

1. *INFO*: mesaj de informare pentru utilizator
2. *ERROR*: mesaje de eroare care împiedica execuția programului
3. *WARNING*: mesaje de avertisment care semnalează probleme in timpul execuției dar nu opresc execuția.

* Toate mesaje de log vor conține data si ora la care au fost afișate.

1. Crawlerul va scrie in fișierul de log in funcție de nivelul de log:
2. Log\_level 0: Crawlerul nu va scrie nimic.
3. Log\_level 1: Crawlerul va scrie doar mesajele INFO de succes si eroare
4. Log\_level 2: Crawlerul va scrie mesajele INFO de succes, eroare si warning
5. Log\_level 3: Crawlerul va scrie toate mesajele, inclusiv mesaje de tip INFO despre stadiul la care se afla crawlerul.

## Dependințe:

Crawlerul va avea nevoie de o conexiune la internet pentru a vizita site-uri, de acea va afișa un mesaj de eroare in log in momentul in care nu exista o conexiune la internet.

## Interfețe:

Datorită faptului că aplicația va fi folosită din comand line, această poate fi integrată foarte ușor cu alte aplicații.

# Cerințe nefuncționale

## Cerințe de interfață

Aplicația va trebui rulată pe unul dintre următoarele sisteme de operare:

⦁ Linux

⦁ Windows 10

⦁ Windows 8

⦁ Windows 7

Pentru a putea fi rulată este nevoie de a instala Java Virtual Machine (JVM) pe același sistem de operare pe care se regăsește și aplicația.

## Cerințe de performanță

Pentru obținerea unei capabilități mărite de funcționare, aplicația ar trebui rulată pe un sistem ce dispune de o conexiune stabilă și constantă la Internet.

Deoarece produsul va lucra cu site-uri web ce pot avea , va fii nevoie ca mașina pe care este rulat să dispună de suficientă putere de procesare.

## Cerințe de securitate

Produsul nu necesită cerințe suplimentare referitoare la securitate inafară de cele aferente sistemului de operare pe care este rulat.