Specificarea Cerințelor Software pentru Proiectul **Aranea**

Apostolescu Ștefan
Băjan Ionuț-Mihăiță
Brumă Ionuț-Cosmin
Iosif George-Andrei
Stanciu Răzvan-Daniel

30/11/2020

Tabelă de Conținut

Listă de Imagini

Listă de Tabele

1 Detalii despre Document

1.1 Scop

Scopul acestui document este de a descrie cu lux de amănunte modul în care membrii echipei, enumerați anterior, vor proiecta, implementa și testa sistemul Aranea pe baza cazurile de utilizare și a cerințele funcționale și nefuncționale.

1.2 Conținut

Documentul este împărțit în trei capitole. Primul oferă o descriere generală a proiectului, plecând de la problemele întâlnite și ajungând până la modul în care soluția dezvoltată le va rezolva. Al doilea capitol prezintă cerințele funcționale pe care proiectul trebuie să le îndeplinească, cuprinzând aici actorii implicați, limitele sistemului și cazurile de utilizare. Al treilea descrie cerințele nefuncționale, precum interfața cu utilizatorii, performanță, fiabilitate și securitate.

2 Descriere Generală a Proiectului

2.1 Situație Curentă

Acum 25 de ani, numai 0.4% din populația lumii era conectată la Internet. Datorită avantajelor evidente pe care acesta le oferă, numărul persoanelor care îl folosesc a crescut considerabil, în prezent fiind de peste 60%[1].
Această creștere s-a reflectat și în conținutul disponibil online. Deoarece devenise din ce în ce mai dificilă găsirea unei informații particulare, au fost introduse motoare software cu rolul de a căuta sistematic World Wide Web-ul. Acestea folosesc pentru indexarea informației un tip special de roboți, numit crawler sau spider.
Pe lângă acest caz de utilizare, un program de tip crawler mai este folosit pentru colectarea din World Wide Web a informațiilor care corespund anumitor criterii. De exemplu, o companie se poate folosi de un astfel de robot pentru a căuta pe Internet prețurile companiilor asemănătoare, din aceeași localitate, cu scopul de a obține un avantaj concurențial prin practicarea unor prețuri mai mici.

2.2 Misiune

Prin implementarea proiectul, se dorește oferirea unei soluții open-source pentru crawling web. Aceasta va putea fi folosită de către oricine, pentru a-și îndeplini sarcinile specifice, putând contribui în același timp la rezolvarea de erori, respectiv la dezvoltarea ulterioară a soluției, prin intermediul platformei GitHub.

2.3 Context

Aranea pune la dispoziția utilizatorilor, printr-o interfață în linie de comandă, o serie de operații utile descărcării locale a paginilor web și a procesării lor.
Crawling-ul constă în vizitarea recursivă a paginilor unor website-uri listate într-un fișier dat ca intrare programului. Printr-un alt fișier, de configurare, se va putea personaliza tot procesul de descărcare al paginilor. Acesta profită de puterea procesoarelor moderne, utilizând mai multe fire de execuție. Scopul final este de a salva local paginile care corespund anumitor criterii dictate de către utilizator și, opțional, de a genera hărți ale website-urilor.
Pe lângă crawling proiectul oferă posibilitatea de a filtra după o anumită extensie fișierele salvate local, de a căuta după un anumit șablon și de a solicita ajutor.
Soluția oferă avantajul de a fi rulată pe orice calculator. Nu necesită resurse superioare, însă, în cazul în care se dorește salvarea unui website publicat online (deci care nu aparține rețelei locale), este necesară o conexiune stabilă la Internet.

2.4 Beneficii

Beneficiile pe care proiectul le aduce constau, în primul rând, în funcționalitățile menționate anterior. Pe lângă acestea, mai oferă avantajele de a fi ușor de utilizat, open-source și de a nu necesita resurse superioare de la calculatorul pe care este rulat.

[1] Conform Internet Growth Statistics.

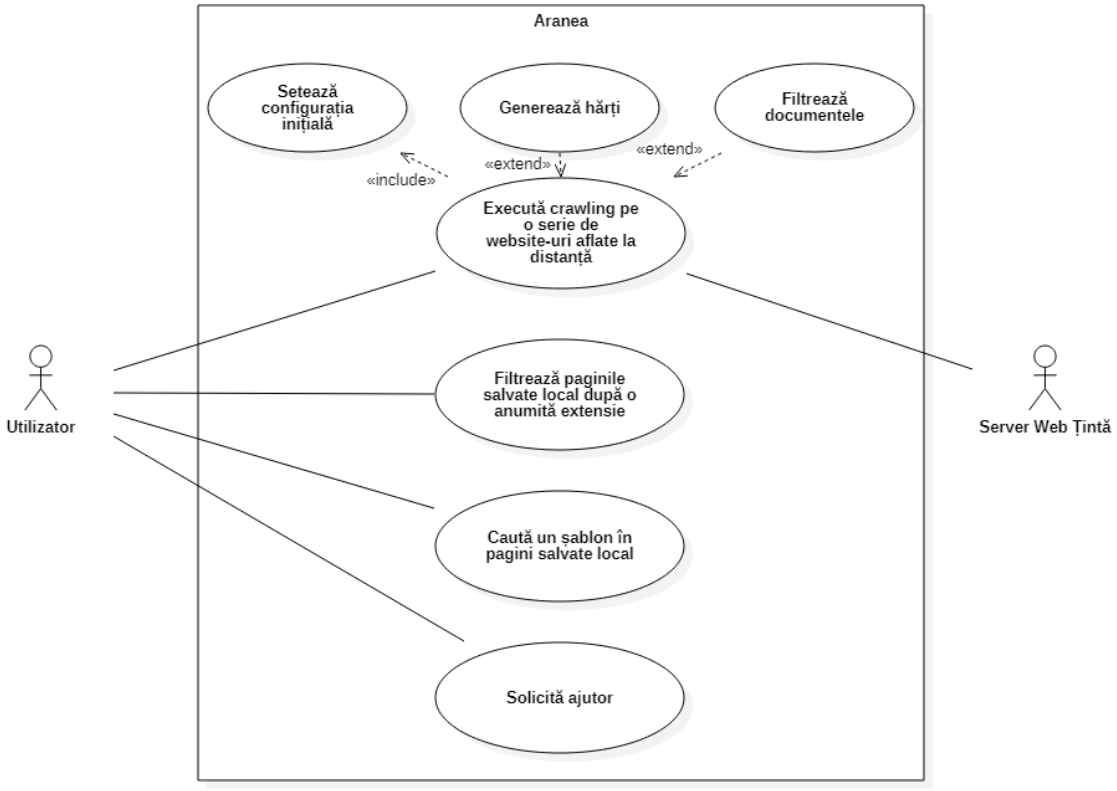
3 Cerințe Funcționale

3.1 Actori

Actorii care interacționează cu sistemul sunt:

- utilizatorul, care folosește soluția propusă, cu un set anume de date de intrare, pentru a-și îndeplini sarcinile specifice,
- server-ul web țintă, care găzduiește paginile ce se doresc a fi descărcate.

3.2 Diagramă de Sistem



Imagine 1: Diagramă de sistem

3.3 Descrieri ale Cazurilor de Utilizare

În tabelele următoare, pentru ușurarea folosirii soluției, au fost setate alias-urile menționate în secțiunea cerințelor de interfață.

Caz de Utilizare	Execută crawling pe o serie de website-uri aflate la distanță
Descriere Scurtă	Descarcă conținutului unor website-uri listate într-un fișier, pe baza unui alt fișier, de configurare.
Prioritate	Ridicată
Declanșator	Rularea comenzii `aranea crawl URLS_FILE CONFIG_FILE`
Condiții Inițiale	1. Existența fișierelor de intrare, cele date ca parametrii 2. Conexiune la Internet, în cazul în care acest lucru este necesar
Rută Uzuală	1. Utilizatorul creează un fișier (`URLS_FILE`) care conține referințe către website-uri ce trebuiesc descărcate. 2. Utilizatorul creează un fișier (`CONFIG_FILE`) cu configurația operației. 3. Utilizatorul rulează comanda specifică, dând ca argumente fișierele menționate anterior. 4. Aranea efectuează operațiunea de crawling, pe mai multe fire de execuție, conform cu configurația primită. 5. Aranea salvează documentele descărcate în folder-ul menționat în fișierul de configurație.
Rute Alternative	1. Filtrarea conținutului descărcat 2. Crearea unor hărți
Condiții Ulterioare	Existența unui folder local ce conține documentele solicitate
Excepții Posibile	1. Un server web țintă nu este disponibil. 2. Un website nu permite accesul, prin fișierul `robots.txt`. 3. O pagină web referențiată nu există. 4. Nu se pot crea local foldere sau fișiere.

Tabel 1: Descriere a cazului de utilizare pentru crawling

Caz de Utilizare	Setează configurația inițială
Descriere Scurtă	Setează configurația folosită de Aranea, pe baza unui fișier local de configurare.
Prioritate	Ridicată
Declanșator	Specificarea parametrului `CONFIG_FILE` în comanda `aranea crawl URLS_FILE CONFIG_FILE`
Condiții Inițiale	∅
Rută Uzuală	Utilizatorul specifică, într-un fișier local, următoarele elemente: • `download_dir`, șir de caractere pentru directorul de descărcare • `log_file`, șir de caractere pentru fișierul de jurnal • `log_level`, întreg pentru prioritatea minimă a unui eveniment pentru a fi jurnalizat *(opțional, implicit `0`)* • `is_sitemap_generated`, boolean care indică dacă se vor genera hărți pentru website-urile descărcate *(opțional, implicit `true`)* • `max_threads`, întreg pentru numărul maxim de fire de execuție *(opțional, implicit `1000`)* • `delay`, numărul de secunde între două cereri consecutive către un server web țintă *(opțional, implicit `1`)* • `allowed_extensions`, șir de caractere pentru extensii ale fișierelor ce vor fi descărcate, separate prin virgulă *(opțional, implicit `*`)* • `allowed_max_size`, întreg pentru dimensiunea maximă, în octeți, avută de un fișier ce va fi descărcat *(opțional, implicit `1000000000`)* • `allowed_pattern`, șir de caractere pentru un șablon Regex ce trebuie să se regăsească într-un fișier pentru a fi descărcat *(opțional, implicit `""`)* • `skip_robotsdottxt_files`, boolean care indică dacă fișierele menționate în `robots.txt` nu vor fi descărcate *(opțional, implicit `true`)*.
Rute Alternative	∅
Condiții Ulterioare	Existența unui fișier local, valid, de configurație
Excepții Posibile	1. Formatul fișierului de configurare este invalid. 2. Fișierul de configurare nu conține toate câmpurile obligatorii.

Tabel 2: Descriere a cazului de utilizare pentru setarea configurației inițiale

Caz de Utilizare	Generează hărți
Descriere Scurtă	Generează o hartă pentru fiecare website descărcat, aceasta listând toate fișierele componente.
Prioritate	Medie
Declanșator	Setarea câmpului `is_sitemap_generated` din fișierul de configurare
Condiții Inițiale	Finalizarea descărcării website-ului curent
Rută Uzuală	După finalizarea fiecărei descărcări a unui website, Aranea parcurge folder-ul specific lui și listează toate fișierele într-un alt fișier de tip hartă.
Rute Alternative	∅
Condiții Ulterioare	Existența unor fișiere locale, reprezentând hărțile
Excepții Posibile	1. Folder-ul nu conține niciun fișier.

Tabel 3: Descriere a cazului de utilizare pentru generarea hărților

Caz de Utilizare	Filtrează documentele
Descriere Scurtă	Înainte de a descărca un fișier de pe server-ul web țintă, verifică dacă el corespunde anumitor criterii setate de către utilizator.
Prioritate	Medie
Declanșator	Setarea a cel puțin unul din câmpurile următoare, din fișierul de configurare: • `allowed_extensions` • `allowed_max_size` • `allowed_pattern` • `skip_robotsdottxt_files`.
Condiții Inițiale	∅
Rută Uzuală	Pentru fiecare fișier ce se descoperă pe website-ul curent, Aranea verifică următoarele: • dacă `allowed_extensions` este setat și dacă extensia fișierului apare în `allowed_extensions` • dacă `allowed_max_size` este setat și dacă dimensiunea fișierului este mai mică decât `allowed_max_size` • dacă `allowed_pattern` este setat și dacă apar în fișier șiruri de caractere care potrivesc `allowed_pattern` • dacă `skip_robotsdottxt_files` este setat și dacă fișierul curent nu apare în `robots.txt`.
Rute Alternative	∅
Condiții Ulterioare	Permisiunea descărcării unui fișier anume
Excepții Posibile	∅

Tabel 4: Descriere a cazului de utilizare pentru filtrarea documentelor

Caz de Utilizare	Filtrează paginile salvate local după o anumită extensie
Descriere Scurtă	Filtrează paginile salvate local, aparținând unor website-uri, după extensie
Prioritate	Ridicată
Declanșator	Rularea comenzii `aranea list EXTENSION`
Condiții Inițiale	∅
Rută Uzuală	Aranea parcurge directoarele specifice unor website-uri, listând în linie de comandă fișierele cu o anumită extensie.
Rute Alternative	∅
Condiții Ulterioare	Afișarea listării solicitate, în linie de comandă
Excepții Posibile	∅

Tabel 5: Descriere a cazului de utilizare pentru filtrarea după o anumită extensie a paginilor salvate local

Caz de Utilizare	Caută un șablon în pagini salvate local
Descriere Scurtă	Caută un șablon în paginile descărcate local, aparținând unor website-uri.
Prioritate	Ridicată
Declanșator	Rularea comenzii `aranea search PATTERN`
Condiții Inițiale	Existența a cel puțin un folder care să reprezinte un website descărcat anterior
Rută Uzuală	Aranea caută recursiv în toate directoarele, printând în linie de comandă numele fișierelor care conțin șiruri de caractere ce potrivesc șablonul.
Rute Alternative	∅
Condiții Ulterioare	Afișarea în linie de comandă a fișierelor găsite
Excepții Posibile	1. Folder-ul de descărcări este gol.

Tabel 6: Descriere a cazului de utilizare pentru căutarea unui șablon în paginile salvate local

Caz de Utilizare	Solicită ajutor
Descriere Scurtă	Solicită ajutorul cu privire la utilizarea programului.
Prioritate	Medie
Declanșator	Rularea comenzilor `aranea` sau `aranea help`
Condiții Inițiale	∅
Rută Uzuală	Aranea afișează în linie de comandă o detaliere a folosirii programului
Rute Alternative	∅
Condiții Ulterioare	Afișarea în linie de comandă a unei detalierii menționate anterior
Excepții Posibile	∅

Tabel 7: Descriere a cazului de utilizare pentru solicitarea ajutorului

4 Cerințe Nefuncționale

4.1 Cerințe de Interfață

Proiectul va putea rula pe calculatoare cu sisteme de operare moderne (Windows, Linux și macOS), dotate cu Java Virtual Machine. O conexiune la Internet este necesară în cazul în care se doresc accesate website-uri din afara rețelei locale.
Interfața proiectului va fi în linie de comandă. Utilizarea va necesita cunoștințe minime de operare a unui calculator, aceasta fiind ușurată de aspecte precum:

- posibilitatea setării unor alias-uri cu ajutorul comenzilor `doskey aranea=java -jar ABSOLUTE_PATH_TO/aranea.jar` pe Windows și `alias aranea=java -jar ABSOLUTE_PATH_TO/aranea.jar` pe Linux și macOS
- comenzi simple, ușor de memorat
- acces la o detaliere a utilizării proiectului, cu ajutorul comenzilor `aranea` sau `aranea help`.

4.2 Cerințe de Performanță

Aranea, fiind un proiect software simplu, nu necesită dotări suplimentare pe lângă cele solicitate de Java Virtual Machine. Prin implementarea ce favorizează folosirea de fire de execuție, garantează o performanță ridicată, ce poate fi însă scăzută de viteza sau de instabilitatea conexiunilor cu website-urile țintă.

4.3 Cerințe de Securitate

Cum unele website-uri sunt protejate de firewall-uri, există riscul ca adresa utilizatorului să fie blocată. Astfel, accesul de pe aceasta nu ar fi permis, chiar dacă ar fi vorba de navigarea neautomată (din browser) a website-ului.
Pentru a micșora acest risc, toate cererile către serverele web sunt distanțate implicit de o secundă și identificate printr-un agent specific, `AraneaBot`.