**Ministerul Educației, Culturii și Cercetării al Republicii Moldova**

|  |  |
| --- | --- |
| **D:\1-107\Stela Balan\DMSAC\Regulamente UTM, 2015\Logou nou UTM\Logo_inscript_vertical.png** | **UNIVERSITATEA TEHNICĂ A MOLDOVEI** |

**Facultatea Calculatoare, Informatică și Microelectronică**

**Departamentul Informatica şi Ingineria Sistemelor**

**RAPORT**

**Lucrare de laborator Nr. 3**

**La disciplina Analiza statistică a datelor**

|  |  |
| --- | --- |
| **A efectuat:** | **st. gr.IA - 182**  **Ulmanu Cristian** |
| **A verificat:** | **conf.univ.**  **Bumbu Tudor** |

**Chișinău, 2020**

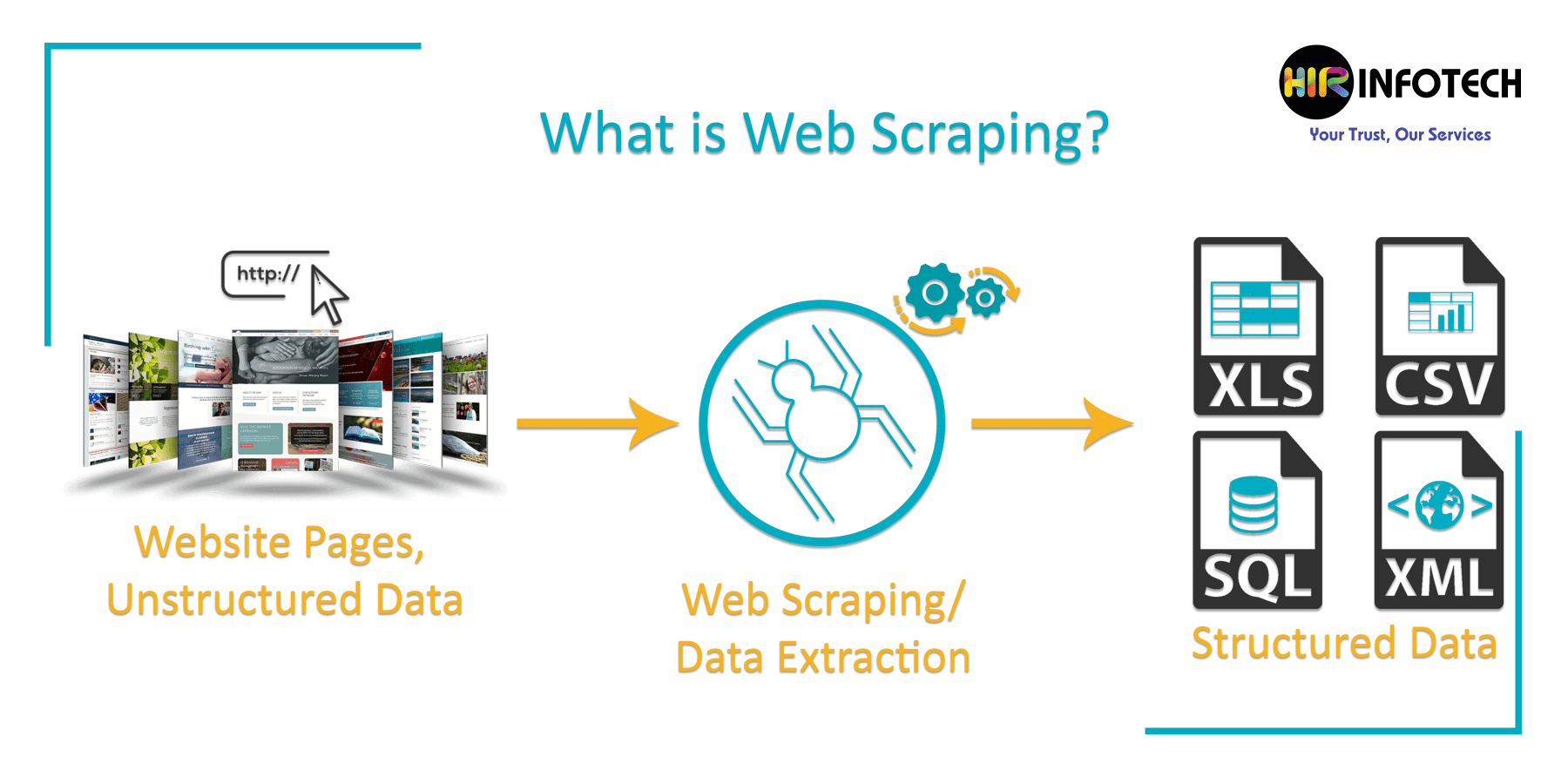
**Tema: Extragerea informaţiei**

**Sarcină:** **Elaborarea unei scheme logice (bloc) cu proiectul sudiat.**

**Elaborarea schemei logice pentru aplicatia care urmeaza s-o dezvoltati**

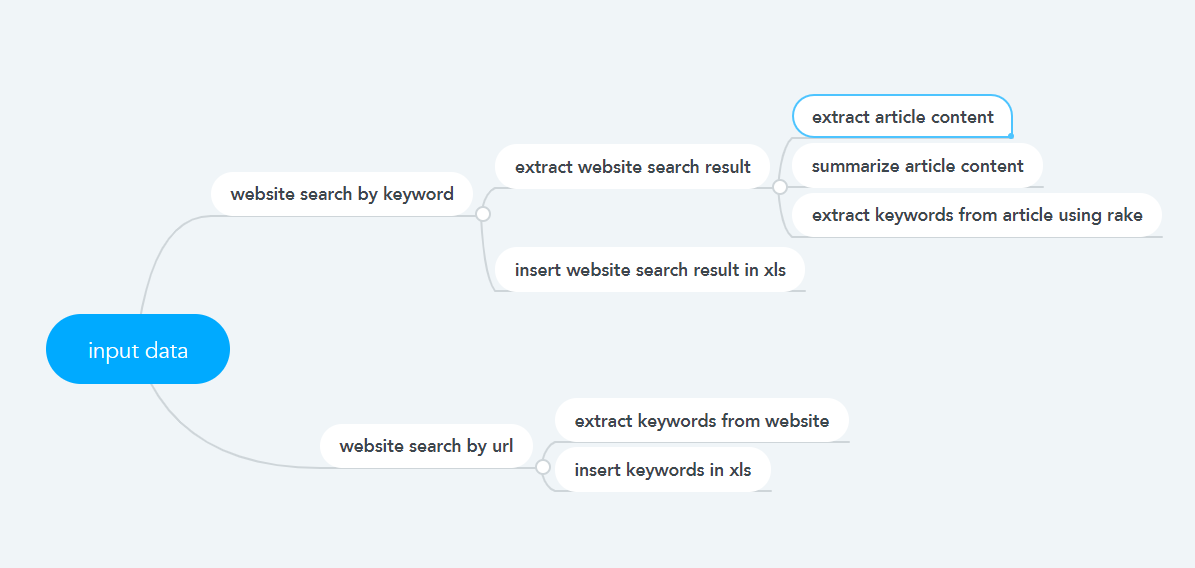
**pentru a rezolva problema.**

# Schema generală a procesului de web scraping



Procesul de web scraping presupune preluarea informației concrete de pe website-uri și sistematizarea, sumarizarea și structurarea lor astfel ca ele să poată fi memorizate în diferite tipuri de fișiere pentru utilizarea ulterioară. Procesul presupune în primul rând investigharea structurii site-ului de unde preluăm informația, pentru a determina din care elemente este necesar de extras informația. Pe urmă printr-un limbaj de programare, în cazul meu Python, se creează un script care preluează datele din anumite elemente ale DOM și după acelea ele pot fi sortate, rezumate, extrase cuvintele cheie și această informație poate fi înscrisă în xls sau doc sau în cadrul unei baze de date.

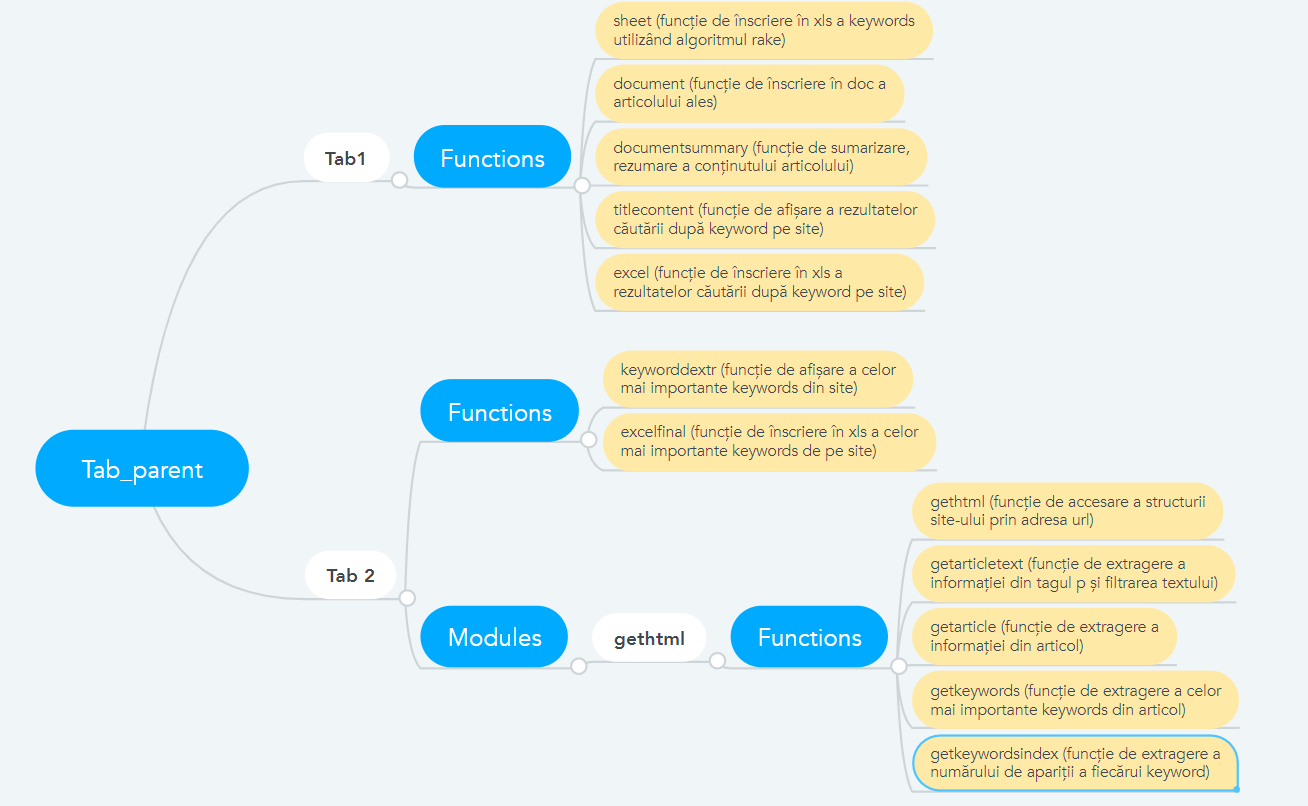
# Structura funcțională a aplicației



Aplicația este separată în două componente independente, una răspunde de căutarea informației dintr-un site anumit după un cuvânt cheie. După ce introducem cuvântul cheie noi vom avea ca rezultat o listă de articole unde putem alege 3 opțiuni pentru fiecare articol. Putem pur și simplu să extragem informația din el, putem să rezumăm informația și să extragem informația cu algortimul rake.

Al doilea component permite să extragem informația (cuvinte-cheie) din adresa site-ului pe care îl introducem.

# Algoritmul aplicației



În schema data este descrisă structura ei funcțională. Ea constă din 2 module grafice Tab1 și Tab2 și fiecare modul își are propriul set de funcții care vor fi utilizate.