**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. 2**

**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: Pregătirea uneltelor de lucru: resurse web, instrumentar soft.**

# Python

Python este un limbaj de programare dinamic multi-paradigmă, creat în 1989 de programatorul olandez Guido van Rossum. Van Rossum este și în ziua de astăzi un lider al comunității de dezvoltatori de software care lucrează la perfecționarea limbajul Python și implementarea de bază a acestuia, CPython, scrisă în C. Python este un limbaj multifuncțional folosit de exemplu de către companii ca Google sau Yahoo! pentru programarea aplicațiilor web, însă există și o serie de aplicații științifice sau de divertisment programate parțial sau în întregime în Python. Popularitatea în creștere, dar și puterea limbajului de programare Python au dus la adoptarea sa ca limbaj principal de dezvoltare de către programatori specializați și chiar și la predarea limbajului în unele medii universitare. Din aceleași motive, multe sisteme bazate pe Unix, inclusiv Linux, BSD și Mac OS X includ din start interpretatorul CPython.

Python pune accentul pe curățenia și simplitatea codului, iar sintaxa sa le permite dezvoltatorilor să exprime unele idei programatice într-o manieră mai clară și mai concisă decât în alte limbaje de programare ca C. În ceea ce privește paradigma de programare, Python poate servi ca limbaj pentru software de tipul object-oriented, dar permite și programarea imperativă, funcțională sau procedurală. Sistemul de tipizare este dinamic iar administrarea memoriei decurge automat prin intermediul unui serviciu „gunoier” (garbage collector). Alt avantaj al limbajului este existența unei ample biblioteci standard de metode.

Implementarea de referință a Python este scrisă în C și poartă deci numele de CPython. Această implementare este software liber și este administrată de fundația Python Software Foundation.

# HTML

Hypertext Markup Language (HTML) este limbajul de marcare standard pentru documentele concepute pentru a fi afișate într-un browser web. Poate fi asistat de tehnologii precum Cascading Style Sheets (CSS) și limbaje de script cum ar fi JavaScript.

Navigatoarele Web primesc documente HTML de pe un server web sau de la stocarea locală și le redau în pagini web multimedia. HTML descrie structura unei pagini web în mod semantic și inițial a inclus indicii pentru aspectul documentului.

Elementele HTML sunt blocurile de construcție ale paginilor HTML. Cu construcții HTML, imagini și alte obiecte, cum ar fi formulare interactive, pot fi încorporate în pagina redată. HTML oferă un mijloc de a crea documente structurate notând semantică structurală pentru text cum ar fi titluri, paragrafe, liste, link-uri, citate și alte elemente. Elementele HTML sunt delimitate prin etichete, scrise cu paranteze de unghi. Etichete precum <img /> și <input /> introduc direct conținut în pagină. Alte etichete, cum ar fi <p> ​​înconjoară și oferă informații despre textul documentului și pot include alte etichete ca sub-elemente. Navigatoarele nu afișează etichetele HTML, ci le folosesc pentru a interpreta conținutul paginii.

HTML poate încorpora programe scrise într-un limbaj de script cum ar fi JavaScript, care afectează comportamentul și conținutul paginilor web. Includerea CSS definește aspectul și aspectul conținutului. World Wide Web Consortium (W3C), fost întreținător al HTML-ului și actual mentenant al standardelor CSS, a încurajat încă din 1997 utilizarea CSS peste HTML prezentativ explicit.

# DOM

Modelul de obiect de document (DOM) este o interfață inter-platformă și independentă de limbaj care tratează un document XML sau HTML ca o structură de arbore în care fiecare nod este un obiect reprezentând o parte a documentului. DOM reprezintă un document cu un arbore logic. Fiecare ramură a copacului se termină într-un nod și fiecare nod conține obiecte. Metodele DOM permit accesul programatic la arbore; cu ele se poate schimba structura, stilul sau conținutul unui document. Nodurile pot avea dispozitive de gestionare a evenimentelor atașate de ele. Odată declanșat un eveniment, gestionatorii de evenimente sunt executați.

Principala standardizare a DOM a fost gestionată de World Wide Web Consortium, care a elaborat ultima dată o recomandare în 2004. WHATWG a preluat dezvoltarea standardului, publicându-l ca document viu. W3C publică acum instantanee stabile ale standardului WHATWG.

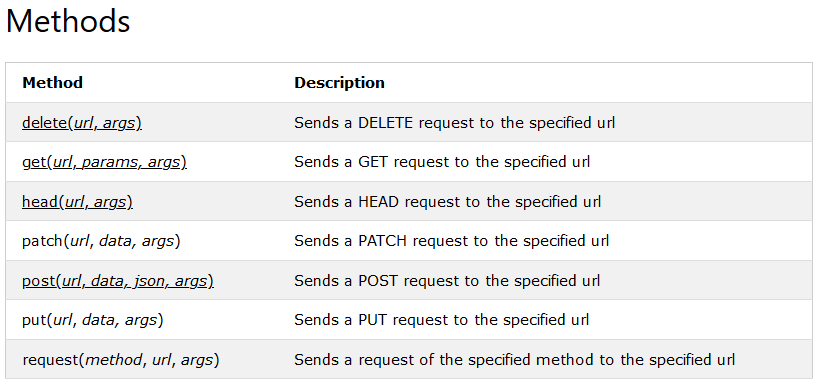
# Biblioteci utilizate în Python

# Tkinter

* Tkinter este biblioteca GUI standard pentru Python. Python atunci când este combinat cu Tkinter oferă un mod rapid și ușor de a crea aplicații GUI. Tkinter oferă o interfață puternică orientată pe obiect la setul de instrumente Tk GUI. Crearea unei aplicații GUI folosind Tkinter este o sarcină ușoară. Tot ce trebuie să faceți este să efectuați următorii pași -
* Importați modulul Tkinter.
* Creați fereastra principală a aplicației GUI.
* Adăugați unul sau mai multe dintre widget-urile menționate mai sus la aplicația GUI.
* Introduceți bucla evenimentului principal pentru a lua măsuri împotriva fiecărui eveniment declanșat de utilizator.

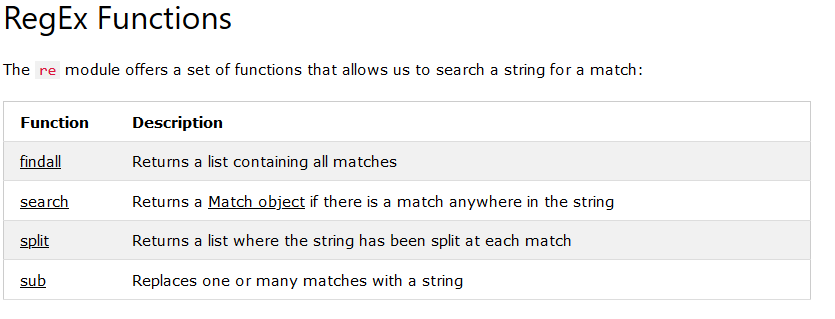
# Requests

Modulul de solicitări vă permite să trimiteți cereri HTTP folosind Python. Cererea HTTP returnează un obiect de răspuns cu toate datele de răspuns (conținut, codificare, stare, etc.).



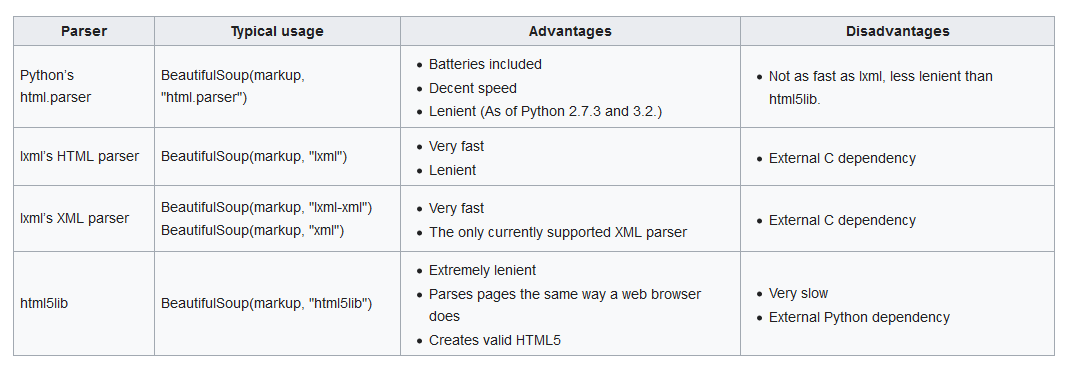
# RegEx

Un regEx, sau expresie regulată, este o secvență de caractere care formează un model de căutare. RegEx poate fi utilizat pentru a verifica dacă un șir conține modelul de căutare specificat.



# BeautifulSoup4

Beautiful Soup este un pachet Python pentru analizarea documentelor HTML și XML (incluzând marcaj malformat, adică etichete neîncheiate, denumite astfel după supă de etichete). Creează un arbore de analiză pentru paginile analizate care pot fi utilizate pentru a extrage date din HTML, care este utilă pentru razuirea web.  
  
Este disponibil pentru Python 2.7 și Python 3.



# XLWT

Aceasta este o bibliotecă pe care dezvoltatorii să o folosească pentru a genera fișiere cu foi de calcul compatibile cu versiunile Microsoft Excel 95 - 2003. Pachetul în sine este Python pur, fără dependențe de module sau pachete în afara distribuției standard Python.

# OS

Acest modul oferă un mod portabil de utilizare a funcționalității dependente de sistemul de operare. Dacă doriți doar să citiți sau să scrieți un fișier, consultați open (), dacă doriți să manipulați căi, consultați modulul os.path și dacă doriți să citiți toate liniile din toate fișierele de pe linia de comandă, consultați modulul de introducere fișier . Pentru crearea fișierelor și directoarelor temporare, consultați modulul tempfile, iar pentru gestionarea fișierelor și a directorilor la nivel înalt, consultați modulul shutil.

# Rake\_nltk

RAKE scurt pentru algoritmul de extragere rapidă a cuvintelor cheie, este un algoritm de extragere a cuvintelor cheie independent de domeniu, care încearcă să determine frazele cheie dintr-un corp de text, analizând frecvența aspectului cuvântului și coincidența acestuia cu alte cuvinte din text.

# NLTK

Natural Language Toolkit, sau mai frecvent NLTK, este o suită de biblioteci și programe pentru procesarea simbolică și statistică a limbajului natural (NLP) pentru limba engleză scrisă în limbajul de programare Python. A fost dezvoltat de Steven Bird și Edward Loper în Departamentul de Informatică și Știința Informației din Universitatea din Pennsylvania. NLTK include demonstrații grafice și date de eșantion. Este însoțită de o carte care explică conceptele care stau la baza sarcinilor de procesare a limbajului, susținute de setul de instrumente, plus o carte de bucate.

NLTK este destinat să sprijine cercetarea și predarea în PNL sau în domenii strâns legate, incluzând lingvistica empirică, știința cognitivă, inteligența artificială, regăsirea informațiilor și învățarea mașinii. NLTK a fost utilizat cu succes ca instrument de predare, ca instrument individual de studiu și ca platformă pentru prototipare și sisteme de cercetare a construcțiilor. Există 32 de universități din SUA și 25 de țări care utilizează NLTK în cursurile lor. NLTK acceptă funcționalități de clasificare, tokenizare, stemming, tagging, parsing și raționament semantic.

# Mechanize

Navigare web programatică de stat, după modulul Perl de Andy Lester WWW :: Mecanizați. mecanize.Browser implementează interfața urllib2.OpenerDirector. Obiectele browserului au o stare, inclusiv istoricul navigării, starea formei HTML, cookie-urile etc. Setul de caracteristici și scheme URL adresate de obiectele Browserului poate fi configurat. Biblioteca oferă, de asemenea, o API care este în mare parte compatibilă cu urllib2: programul dvs. urllib2 va funcționa probabil dacă înlocuiți „urllib2” cu „mecanizare” peste tot.

Caracteristicile includ: ftp:, http: și fișier: scheme URL, istoricul browserului, hyperlink și suport pentru formulare HTML, cookie-uri HTTP, HTTP-EQUIV și Actualizare, Referer [sic] antet, robots.txt, redirecționări, proxy, și Basic și Digest Autentificare HTTP.

O mare parte din codul provenit inițial din codul Perl de Gisle Aas (libwww-perl), Johnny Lee (suport pentru cookie-uri MSIE) și nu în ultimul rând Andy Lester (WWW :: Mecanizați). urllib2 a fost scris de Jeremy Hylton.

# Site-uri utilizate pentru scraping

Scholarpedia este o enciclopedie online bazată pe wiki în limba engleză, cu caracteristici asociate în mod obișnuit cu reviste academice online cu acces deschis, care își propune să aibă conținut de calitate.

Articolele Scholarpedia sunt scrise de autorii experți invitați și sunt supuse revizuirii de la egal la egal. Scholarpedia listează numele și afilierea reală a tuturor autorilor, curatorilor și redactorilor implicați într-un articol: cu toate acestea, procesul de revizuire de la egal la egal (care poate sugera modificări sau adăugiri și trebuie să fie satisfăcut înainte de apariția unui articol) este anonim. Articolele Scholarpedia sunt stocate într-un depozit online și pot fi citate ca articole de jurnal convenționale (Scholarpedia are numărul ISSN ISSN 1941-6016). Sistemul de citare al Scholarpedia include suport pentru numerele de revizuire.

Proiectul a fost creat în februarie 2006 de Eugene M. Izhikevich, pe când era cercetător la Neurosciences Institute, San Diego, California. Izhikevici este și redactorul-șef al enciclopediei.

New York Times Company este o companie americană de mass-media care își publică numele de ziar, The New York Times. Arthur Ochs Sulzberger Jr. a ocupat funcția de președinte din 1997. Are sediul central în Manhattan, New York.

# Bibliografie

1. <https://pypi.org/project/mechanize/>
2. <https://web.archive.org/web/20110902235619/http://aclweb.org/anthology-new/W/W08/W08-0208.pdf>
3. <https://pypi.org/project/rake-nltk/>
4. <https://pypi.org/project/python-docx/>
5. <https://docs.python.org/3/library/os.html>
6. <https://pypi.org/project/xlwt/>
7. <https://www.crummy.com/software/BeautifulSoup/#Download>
8. <https://www.w3schools.com/python/python_regex.asp>
9. <https://pypi.org/project/requests/>
10. <https://www.w3schools.com/python/module_requests.asp>
11. <https://www.tutorialspoint.com/python/python_gui_programming.htm>
12. <https://web.archive.org/web/20170427220310/http://www.digital-web.com/articles/the_document_object_model/>
13. <https://www.w3.org/TR/2012/CR-html5-20121217/>
14. <https://docs.python.org/3/tutorial/index.html>