**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. 1**

**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: Selectarea și studierea temei**

**Extragerea informațiilor (IE), recuperarea informațiilor (IR) este sarcina de a extrage automat informații structurate din documente nestructurate și / sau semi-structurate care pot fi citite de mașini și alte surse reprezentate electronic. În majoritatea cazurilor, această activitate se referă la procesarea textului limbajului uman prin procesarea limbajului natural (PNL). Ultimele activități în procesarea documentelor multimedia, cum ar fi adnotarea automată și extragerea conținutului din imagini / audio / video / documente, pot fi considerate ca extracție de informații.**

În tehnologia informației moderne, datorită creșterii rapide a cantității de informații nestructurate (fără metadate) (în special pe internet), acest proces, cum ar fi extragerea informațiilor, a devenit tot mai important. Aceste informații pot fi construite în continuare prin trecerea la formularul de relație sau adăugarea de etichete XML. Când utilizați agenți inteligenți pentru a monitoriza sursele de știri, sunt necesare câteva metode pentru extragerea informațiilor și transformarea acesteia într-o formă mai convenabilă pentru utilizarea ulterioară.

# Subtaskuri tipice EI

Subtasks tipice pentru extragerea informațiilor:

* Recunoașterea elementelor (entităților) numite, de exemplu: nume de persoane, nume de organizații, nume geografice, evenimente, simboluri de timp și bani etc.
* Rezolvarea anaforei și coreferenței: căutarea de legături legate de același obiect. Un caz tipic al unor astfel de referințe este pronume anafora.
* Evidențierea terminologiei: găsirea unui text dat cuvinte cheie și expresii (colocări).
* Auto-abstractizare: selectarea informațiilor semantice, emoționale, evaluative etc. dintr-un text. Este generativ și declarativ.

# Scraping web

Scraping web (sau razuire sau razuire ← razuirea web engleză) este o tehnologie pentru obținerea de date web prin extragerea acestora din paginile cu resurse web. Răzuirea Web poate fi făcută manual de către un utilizator de computer, dar termenul se referă de obicei la procesele automatizate implementate folosind codul care efectuează solicitări GET către site-ul țintă. Răzuirea Web este utilizată pentru a sintetiza paginile web în forme mai convenabile pentru lucru. Paginile web sunt create folosind limbaje de marcare a textului (HTML și XHTML) și conțin o mulțime de date utile în cod. Cu toate acestea, majoritatea resurselor web sunt destinate utilizatorilor finali și nu pentru comoditatea utilizării automate, astfel încât a fost dezvoltată o tehnologie care „curăță” conținutul web. Descărcarea și vizualizarea paginii sunt cele mai importante componente ale tehnologiei, ele fac parte integrantă din selecția datelor.

# Metode web scraping

Scraping-ul web este un domeniu de dezvoltare activă care împărtășește o inițiativă ambițioasă de dezvoltare a interacțiunii om-calculator, care necesită descoperiri în procesarea și înțelegerea textului paginilor online prin inteligență artificială. Soluțiile moderne de răzuire variază de la sisteme speciale, intensificate în muncă, până la sisteme complet automatizate care pot transforma site-uri web întregi în informații structurate într-un format specific. În mod ideal, atunci când site-ul ale cărui date pe care doriți să le extrage le oferă prin intermediul API-ului accesul inter-domeniu permis. În cazul în care lucrurile nu stau așa, puteți apela la alte metode de răzuire.

1. „Copy-paste” manual  
     
   Uneori, chiar și cea mai bună tehnologie de scraping web nu poate înlocui munca manuală a unei persoane atunci când un utilizator copiază și lipeste textul. În unele cazuri, aceasta este singura soluție posibilă, de exemplu, atunci când site-urile web blochează razuirea web și copiază textul.
2. Acces la servicii proxy  
     
   Dacă site-ul este un document html sau xml și îi sunt permise cereri de mai multe domenii, atunci puteți obține conținutul documentului folosind o solicitare la unul dintre serviciile proxy disponibile pe Internet.
3. Potrivirea modelelor de text  
     
   Un mod simplu, dar puternic, de a prelua informațiile din paginile web. Se poate baza pe comanda UNIX grep (efectuează o căutare în unul sau mai multe fișiere după un model) sau pe potrivirea expresiilor obișnuite ale limbajelor de programare (de exemplu, Perl sau Python).
4. Analizare HTML  
     
   Multe site-uri web constau într-un număr mare de pagini generate dinamic din principala sursă structurată - baza de date. Datele din aceeași categorie sunt de obicei codificate în pagini similare folosind un script comun sau un șablon. În extragerea datelor, un program care detectează astfel de tipare într-o sursă de informații specifice, extrage conținutul său și îl traduce într-o formă, numită coajă. Se presupune că paginile analizate ale sistemului corespund unui șablon comun și că pot fi ușor identificate în termeni de o schemă URL comună. În plus, unele limbaje semi-structurate de interogare a datelor, cum ar fi XQuery și HTQL, pot fi utilizate pentru a analiza pagini HTML și a extrage și transforma conținutul paginii.
5. Model de obiect de document (DOM)  
     
   DOM este un program cu o API pentru documente HTML și XML. Prin încorporarea unui browser web complet, precum Internet Explorer sau un control al browserului Mozilla, programele pot extrage conținut dinamic generat de scripturile din partea clientului. Răspândirea arborelui DOM vă permite să accesați informații în părțile sale separate.
6. Agregarea verticală a datelor  
     
   Există mai multe companii care au dezvoltat platforme online speciale care creează și controlează numeroși roboți. Bots funcționează fără implicare directă umană și, în același timp, interacțiunea lor cu utilizatorii se produce fără comunicare cu site-ul țintă. Pregătirea include crearea unei baze de cunoștințe, datorită cărora este posibilă munca de bot. Bots cumulează date despre proprietățile individuale ale fiecărei resurse în conformitate cu condițiile specificate pentru compararea și analiza ulterioară a valorilor proprietății obținute. Fiabilitatea platformei este măsurată de calitatea informațiilor primite (de obicei numărul de câmpuri) și de scalabilitatea acesteia (până la sute sau mii de site-uri). Această scalabilitate este folosită în principal pentru a converti datele localizate la sfârșitul codului lung de site-uri pe care agregatorii convenționale le consideră dificile sau consumă prea mult timp pentru a colecta conținut.
7. Recunoașterea adnotărilor semantice  
     
   Unele pagini pot conține meta-date sau marcaje și adnotări semantice; folosind metoda de recunoaștere a adnotării semantice, puteți extrage astfel de pagini.
8. Analizoare de pagini  
     
   Dezvoltările se realizează în domeniul inteligenței artificiale, atunci când viziunea mașinii identifică datele, le interpretează, așa cum ar face o persoană, le preia.

Tehnologia de razuire web este convenabilă pentru transferul datelor de pe paginile web către forme mai convenabile, dar există și metode de analiză care, în cazul unei API deschise, pot rezolva mai productiv problema.

# Utilizare web scraping

Răzuirea web a devenit un instrument important pentru colectarea automată de informații pe Internet. Face parte din sistemele de informații de marketing (MIS) pentru formarea bazelor de date sau a băncilor de date, datorită cărora informațiile necesare sunt furnizate sub formă de rapoarte tabulare. Tehnologiile motoarelor de căutare, agregatoarele de conținut sunt, de asemenea, interconectate cu programele de răzuire web.

Scraping-ul web este o formă de copiere a datelor în care anumite informații necesare pentru primire sunt colectate de pe Internet și acumulate, de regulă, într-o bază de date locală centrală sau o foaie de calcul pentru extragerea sau analiza ulterioară. Software-ul de razuire a paginilor web poate accesa World Wide Web direct folosind Protocolul de transfer de hipertext, precum și printr-un browser web.

Conținutul paginii rezultate poate fi analizat, formatat, copiat într-o foaie de calcul și așa mai departe. Scraping-ul web ia de obicei o bucată de date dintr-o pagină pentru a o folosi în alt scop. Un exemplu de razuire este căutarea pe diverse site-uri și copierea numelor, numere de telefon, adrese de e-mail, adrese URL ale unui anumit subiect pentru a crea o bază de date de contacte.

Practic, răzuitoarele web vor rezolva următoarele probleme:

* Căutați informațiile necesare;
* Copiați date de pe Internet;
* Monitorizarea actualizărilor pe site-uri.

Scraping-ul web poate fi un instrument independent și poate servi la preluarea de informații direcționate, poate deveni, de asemenea, o componentă a dezvoltării web utilizate pentru indexarea web, minarea web și minarea datelor, monitorizarea online, modificările prețurilor și compararea acestora, pentru observare pentru concurență și alte colectări de date.

# Bibliografie

1. <https://habr.com/en/post/56461/>
2. <https://cyberleninka.ru/article/n/web-scraping-i-klassifikatsiya-tekstov-metodom-naivnogo-bayesa/viewer>
3. <http://www.cs.bilkent.edu.tr/~guvenir/courses/CS550/Seminar/freitag2000-ml.pdf>