УДК 004.056
https://doi.org/10.47533/2023.1606-146X.6

С. Т. МАМБЕТОВ[1]*, Е. Е. БЕГИМБАЕВА[1,3], А. К. ХИКМЕТОВ[2], С. К. ДЖОЛДАСБАЕВ[2]

[1]Әл-Фараби атындағы Қазақ Ұлттық университеті, Алматы, Қазақстан
[2]Халықаралық Ақпараттық Технологиялар университеті, Алматы, Қазақстан
[3]Қ. Сәтбаев атындағы ҚазҰТЗУ, Алматы, Қазақстан
e-mail: mambetov.saken@gmail.com, enlik89@mail.ru, akhikmetov@iitu.edu.kz, serikdzoldasbaev@gmail.com

ТАҚЫРЫПТЫҚ ИНТЕРНЕТ-РЕСУРСТАРДАН ДЕРЕКТЕРДІ АЛУ АЛГОРИТМІН ӘЗІРЛЕУ

Ғаламтор желісі қарқынды дамуымен қазіргі таңда көптеген әлеуметтік арналарда пайдаланушылар өзіне қатысты жеке деректермен, басқа да ақпараттармен белсенді түрде бөлісуде. Желідегі ақпарат сенімді және қоғамға қауіп төндірмейтініне көз жеткізу үшін оны талдау қажет. Осының негізінде бұл ақпаратты жинау, бақылау және талдау қажеттілігі туындайды. Ақпаратты жинау әрбір веб-беттің құрылымына байланысты күрделі жұмыс болып табылады. Барлық ресурстар ақпарат жинауға мүмкіндік бермейтіндіктен, көптеген әдістерді қолдануға тура келеді.

Ұсынылған мақалада ақпараттарды алу үшін парсингті қолданудың тиімді жолдары көрсетіледі. Веб-беттердің мазмұнын семантикалық талдау (парсинг) арқылы әдісін Python тілінде BeatifulSoup кітапханасы базасында жазылған бағдарлама арқылы түсіндірілген. Сонымен қатар, ақпаратты жинаудың басқа да API арқылы, браузерде пайдаланушының мінез-құлқын эмуляциялау құралдары арқылы әдістеріне тоқталады. BeatifulSoup + Requests кітапханасы арқылы тақарыптық интернет ресурстардан ақпараттарды алу алгоритмі келтірілген. Нәтижесінде ағылшын тілді және орыс тілді хакерлік және кардерлік форумдардан мәліметтер алынды.

Тұйін сөздер*: beautifulsoup, requests, деректермен жұмыс, парсинг, python, хакерлік форумдар.*

Кіріспе. Қазіргі таңда ақпараттандыру заманы болғандықтан, ақпараттың жаңаруына көз ілестіру мүмкін емес. Соған қоса, ақпарат барған сайын құнды бола түсуде. Осыдан бірнеше ондаған жылдан алдын ғаламтор шағын ғана желі болған. Бірақ бүгінгі таңда бұл үлкен ақпараттық құрылым, онда таралатын ақпарат ағындарын бақылау мүмкін емес болып барады. Ғаламтор желісін қолданушылардың бөлісетін ақпаратты дәл өлшемін айтуға мүмкіндік жоқ. Оған мысалы ретінде әйгілі әлеуметтік желілерден 1 минут ішінде түсетін хабарламаларды келтіруге болады. Facebook әлеуметтік желісінде 3 миллионнан [1] астам жаңа хабарламалар, Twitterде 300 мыңнан [2] астам твиттер таратылады. Бұл ақпараттың барлығын әртүрлі ұйымдар өз қызметінің тиімділігін арттыру үшін белсенді түрде пайдаланады. Осы ақпаратпен тиімді жұмыс істеу, одан пайда алу және әртүрлі тапсырмаларды орындау үшін бұл деректерді алу, өңдеу және құрылымдау қажет. Ақпараттарды дәстүрлі әдіспен жинау көп уақыт пен адами факторға байланысты. Егер де барлық сайттар қандай да бір бірыңғай стандартқа сәйкес жасалса ғаламтор желісінен ақпарат жинау қиындық тудырмас еді. Бірақ, өкінішке орай, қазіргі уақытта мұндай стандарттар жоқ. Бұл өз кезегінде ақпаратты алу мен өңдеудің техникалық тәсілдерінің белсенді да-

* E-mail корреспондирующего автора: mambetov.saken@gmail.com

муына ықпал етеді. Қазіргі уақытта мұндай тәсілдер, ғаламтор желісінен деректерді жинауды және автоматтандырылған өңдеу, "парсинг" деп аталады [3].

Парсинг қазіргі кез-келген салада экономика, маркетинг, психология, криминалистика кеңінен қолданылуда. Барлық салалардан ақпараттарды жинап, өңдеу арқылы болжам жасауға көмектеседі. Тақырыптық интернет ресурстар ретінде интернет форумдар алынды. Хакерлік интернет форумдарда қосылу үшін арнайы TOR браузері пайдаланылды.

Қазіргі кезде тегін және ақылы парсер көптеп кездеседі. Дегенмен ондай парсерлер нақты бір тапсырмаға негізделмеген. Парсердің десктоптық және бұлттық түрлері кезігеді.

Айырмашылығы біріншісін жұмыс компьютеріне орнату арқылы, екіншісінде ештеңе орнатудың қажеті жоқ, тек нәтижесі дайын болған кезде жүктеп алуында. Сондықтан өзіндік тапсырмаға байланысты парсерлік алгоритм жазу өзекті болып табылады. Аталмыш алгоритм қажетті деректерді жинауға және өңдеуге мүмкіндік береді.

Әдебиеттерге шолу. Мақалада [4] Python бағдарламалау тілінде жұмыс жасайтын қазіргі талдау құралдары, BeautifulSoup кітапханасы, Scrapy фреймворкі және Selenium тестін автоматтандыруға арналған құралдар жиынтығы туралы ақпарат берілген. Соған қоса веб-талдау процесін толыққанды көрсететін UML диаграммалары, атап айтқанда ғаламтор желісінен мәтінді іздеу мен шығару моделі Activity және Use-case диаграмалары түрінде көрсетілген.

Бұл мақалада [5] интернеттен деректерді алудың қазіргі жай-күйіне және оның осы технологияларды практикалық қолданумен байланысына қысқаша шолу жасалынған. Бұл [6] мақалада тек веб-сайттың визуалды көрінісіне негізделген веб-беттерді тазарту тәсілі ұсынылады. Мұнда құжаттың орналасуын талдау және таңбаларды оптикалық тану (OCR) саласындағы қолданыстағы тәсілдер ескеріледі. Ақпаратты алу техникасына, яғни веб-парақтарды талдауға, веб-парақтарды талдаудың әртүрлі әдістеріне және веб-парақтарды талдау үшін қолданылатын кейбір соңғы құралдарға шолу жасауға бағытталған.

Бұл мақалада [7] авторлар ақпараттық жүйенің қауіптері мен осалдықтарын мониторинг жүргізу мақсатында ғаламтор желісінде жолығатын ақпараттарды жинақтап, соны өңдеу қажеттілігін көрсеткен. Соған қоса талдау бағдарламасының әрекет алгоритмі келтірілген.

Келесі авторлар [8] веб-форумдардан құрылымдық деректерді алуға арналған vigi4med Scraper әмбебап ашық бастапқы кодты фреймворкты ұсынады. Бұл фреймворк конфигурация файлын пайдалану арқылы, пайдаланушының кез-келген веб-форумнан шығару үшін деректерді еркін таңдай алады және де оңай реттеледі. Алынған деректер анонимді болып табылады және Resource Description framework (RDF) бағандарын қолдана отырып, семантикалық құрылымда ұсынылады.

Аталмыш мақалада [9] әлеуметтік желілерде келтіретін деректерді талдауға негізделген қазіргі тағдағы ең өзекті заманауи ақпараттық жүйелердің 14 түріне қысқаша шолу жасалынып, Қазақстандағы жастар арасындағы танымал, нақты уақыттағы кілт сөздер арқылы деректер жинауға мүмкіндік беретін StreamingAPI құралы бар Вконтакте желісі таңдалып алынған.

Мұнда [10] жоғарыдағы авторлар секілді веб-парақтарды талдаудың әртүрлі аспектілеріне тоқталады және авторлар веб-парақтарды талдауға арналған әртүрлі бағдарламалар мен құралдарды қысқаша сипаттайды. Сондай-ақ веб-парақтарды талдау процесін әртүрлі әдістерін әзірлеу арқылы түсіндіреді және веб-парақтарды талдаудың артықшылықтары мен кемшіліктеріне тоқталып, оны қолдануға болатын әртүрлі салалардың егжей-тегжейлі сипаттамасы келтіріледі.

Бұл жұмыста [11] dark web форумдарының мазмұнын жинауға арналған жаңа сканерлеу жүйесін ұсынады. Ұсынылған жүйеге тоқталсақ, dark web форумдарына қол жеткізу үшін адамның қол жетімділік тәсілін қолданады. URL мекенжайларын ұйымдастырудың бірнеше мүмкіндіктері мен әдістері форум хабарламаларын тиімді шығаруға мүмкіндік береді. Жүйе сонымен қатар кеңейтілген іздеу мен жиналған мазмұнды жаңартуды жеңілдетуге арналған еске түсіруді жақсарту механизмімен біріктірілген қосымша іздеу роботын қамтиды. Адамның қол жетімділік тәсілінің тиімділігін бағалау үшін жүргізілген эксперименттер және кері қайтарып алу мен жақсартуға негізделген біртіндеп жаңарту процедурасы оң нәтиже берді. Адамға қатысты тәсіл Darknet форумдарына қол жетімділікті едәуір жақсартты, ал еске түсіруді жақсартатын қосымша сканер мерзімді және қосымша жаңартудың стандартты тәсілдерінен асып түсті. Жүйенің көмегімен жалпы алғанда үш аймақтан 100-ден астам dark web форумдарын хаттарын жиналды.

Бұл жұмыста [12] киберқауіпсіздік мақсатында Darknet желісінен ақпарат жинауға қатысты мәселелер талданады. Ақпаратты жинауды және құрылымдалмаған деректердің үлкен көлемін талдауды жеңілдету үшін, Dark Web қызметтерін бақылайтын және жиналған деректерді бір аналитикалық құрылымға біріктіретін BlackWidow - жоғары автоматтандырылған модульдік жүйесін ұсынады. Black Widow жүйесі бұрыннан бар және теңшелетін машиналық оқыту құралдарын біріктіруге мүмкіндік беретін Docker негізіндегі микросервистік архитектураға сүйенеді. BlackWidow хабарламалардан алынған барлық алынған мәліметтер мен тиісті қатынастарды қауіпсіздікті талдаушы пайдаланушыларға іздеу және интерактивті визуалды зерттеу үшін қол жетімді үлкен график ретінде ұсынады.

Бұл мақалада [13] Quora сайтынан деректер жинау үшін Python API Quoristі ұсынады. Аталмыш API ашық бастапқы web-автоматтандырудың кросс-платформалық ортасы Selenium-ге негізделген. Бұл API сайтқа арнайы HTTPS сұрауларын жасау арқылы және одан келген жауаптарды талдау арқылы жұмыс жасайды.

Бұл мақалада [14] киберқылмысты анықтау және бақылау мақсатында интернеттегі әртүрлі көздерден, соның ішінде хакерлік форумдардан деректерді жинау әдістері мен технологиялары сипатталған. Мақалада мәтіндік деректерді талдау және ауытқуларды анықтау әдістері қарастырылады, сонымен қатар практикалық қолдану мысалдары келтірілген.

Интернет-ресурстардан деректерді жинаудың тағы бір тәсілі - API (application programming interface) пайдалану. API-бұл деректерді алу үшін веб-сайтпен немесе веб-қызметпен өзара әрекеттесуге мүмкіндік беретін бағдарламалық интерфейстер жиынтығы. Мақалада [15] авторлар веб-скрепинг пен деректерді жинау үшін API қолданудың артықшылықтары мен кемшіліктерін салыстырады. Олар

деректерді жинау құны, сапасы және жылдамдығы сияқты әртүрлі аспектілерді талқылайды.

Интернет-ресурстардан деректерді жинаудың тағы бір тәсілі-машиналық оқытуды қолдану. [16] мақалада авторлар машиналық оқытуға негізделген веб-беттерден деректерді жинау шеңберін әзірлеуді сипаттайды. Олар веб-беттердің құрылымын анықтау және беттегі белгілі бір элементтерден деректерді алу үшін машиналық оқыту алгоритмдерін пайдаланады.

Материалдар мен әдістер. Тақырыптық интернет-ресурстардан деректерді алу алгоритмін әзірлеу үшін келесі қадамдарды орындау қажет:

1. Алгоритмнің мақсаты мен міндеттерін анықтаңыз. Мысалы, егер мақсат тақырыптық сайттардан пікірлер жинау болса, онда тапсырмалар келесідей болуы мүмкін:

• Деректерді жинауға арналған тақырыптық сайттардың тізімін анықтау;

• Сайттағы пікірлерді таңдау критерийлерін анықтау (мысалы, күні, айдары, кілт сөздері бойынша);

• Сайттардан деректерді алу әдісін анықтау (мысалы, HTML парақтарын талдау, API пайдалану, RSS арналарын жүктеу)

2. Деректер көздерінің мүмкіндіктері мен шектеулерін зерттеңіз. Мысалы, барлық сайттар ашық API немесе RSS арналарын ұсынбайды, кейбір сайттар тым жиі сұраныстарды бұғаттай алады (Мысалы, IP-мекен-жайға тыйым салу), сонымен қатар кейбір сайттар деректерді автоматты түрде жинаудан қорғау үшін CAPTCHA кодтарын қолдана алады.

3. Ғаламтор желісіндегі хакерлік форумдардан деректерді жинау алгоритмін жасаңыз. Мысалы, алгоритм келесі қадамдардан тұруы мүмкін:

• Арнайы анонимдікті қамтамасыз ететін бағдарлама арқылы жұмыс жасау;

• Хакерлік форумдар тізімін жасау;

• Форум сайттарын зертеу;

• Тізімдегі әрбір форумнан тақырыптар сілтемелерін алу;

• Берілген критерийлер бойынша пікірлерді іріктеу;

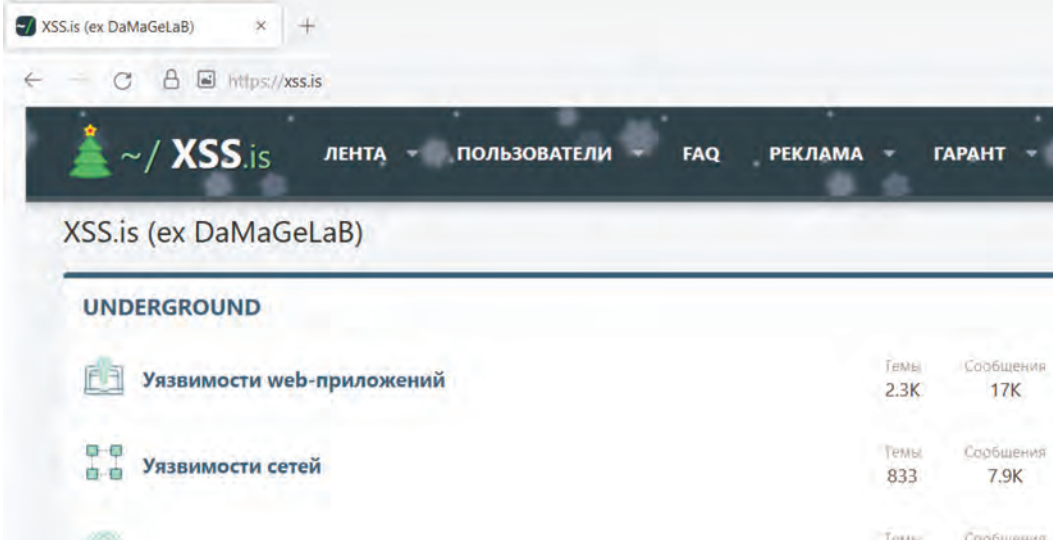
• Пікірлер мәтінін және басқа да қажетті деректерді шығару;

• Алынған деректерді қажетті форматта сақтау.

Алгоритмді тексеріңіз және түзетіңіз. Ол үшін оны әртүрлі тақырыптық сайттарда сынап, оның тиімділігі мен дәлдігін бағалау қажет.

Алгоритмді жұмыс күйінде ұстаңыз. Мысалы, егер деректерді жинау шарттары өзгерсе (мысалы, HTML парағының құрылымы өзгерсе), онда алгоритмге тиісті өзгерістер енгізу қажет. Сондай-ақ, сайттарға сұраныстар белгіленген шектеулерден аспайтынына және IP-мекен-жайдың бұғатталуына әкелмейтініне көз жеткізу керек.

Тақырыптық интернет ресурстар ретінде біз хакерлік және кардерлік интернет форумдарды алдық. Ең алдымен даркнет форумдарына кіру үшін TOR браузерін пайдаландық. TOR браузері онлайн анонимдікті қамтамасыз етіп қана қоймай, браузерге қосылатын құрылғылардың IP-мекенжайлары мен ресурстарын сақтамайды [17]. Сондықтан осы браузерді пайдаландық. Осы браузер пайдалану арқылы біз өзімізді кибершабуылдан қорғауымыз мүмкін.



Сурет 1 – TOR браузерінде xss.is форумының басты парақшасы

Ғаламтор бетін аша қалсаңыз, неше түрлі форумдарға тап боласыз. Бізге керегі хакерлік және кардерлік форумдар. Форумдардың ішінен қазіргі кездегі ең танымал деген бірнешеуін таңдап алдық:

- Probiv.one
- xss.is
- KOROVKA.CC
- Exploit.in
- crdclub.su

Аталмыш форумдардың таңдаудың басты себебі барлығы дерлік тегін интернет форумдар және ішкі құрылымдары бір-біріне ұқсас. Форум сайттары CMS дайын движоктары арқылы жасалынған. Мұндай ұқсастық бізге парсер алгоритмін жазуда жеңілдік береді.

Probiv.one - Бұл форумның басқа форумдардан ерекшелігі – толығымен мәліметті алып беруге бағытталған. Яғни белгілі бір сомаға сізге белгілі бір адам туралы ақпаратты толықтай біле аласыз. Жеке мәліметтер, төлқұжат деректері және несие тарихына дейін алуға болады. Мұнда сіз бірден “табыс сұлбалары” саудасын таба аласыз: адам форум сайты әкімшілігіне ғаламтор желісінде немесе оффлайн режимінде табыс табу нұсқаулығы бар құжатты ұсынады, материалды әкімшілік тексергеннен кейін, ол форумның белгілі бір тармағында жарияланады және пайдаланушыларға сатылады. Мұндай “табыс сұлбалары” үшін бағаның таралуы - бәрі ықтимал пайдаға байланысты, яғни бірнеше мыңнан бірнеше миллионға дейін барады. Тағы да бір артықшылығы Exploit, WWH және Verified сынды форумдарда авторлар ТМД елдерінде жұмыс жасағысы келмейтінін айтса, Probiv-те кересінше ТМД елдерінде жұмыс жасайтын кибералаяқтар арасында танымал.

Exploit.in - Бұл форум негізінен хакерлік тақырыптарға арналған. Аталмыш форумға тіркелу тегін, дегенмен форумға доступ алу үшін, қолданушы сайт

әкімшілігіне басқа да форумдардағы аккаунттарына сілтеме беруге тура келеді. Аккаунт белгілі бір беделге ие және тіркелгеніне кем дегенде бір жылдан асуы керек. Олай болмаса, әкімшілік қолданушыға форумға доступ беруден бас тартады. Exploit.in ТМД елдерінің даркнетінде ғана танымал емес, батыс елдерінен де хакерлер қадағалап отырады. Форумда отыратын негізгі контингент, өз қызметтерін ұсынатын кардерлер, сайт құрудан бастап, ботнет қызметіне дейін ұсыныс тастаушылар болып табылады.

KOROVKA.CC - Бұл форумға тіркелудің екі жолы бар. Біріншісі light санаты, мұнда тіркелу тегін, бірақ сіздің форумдағы мүмкіндіктеріңіз шектеулі болады. Екіншісі member санаты, мұнда форумға тіркелу ақылы, бұл кезде сізде форумның кез-келген бөліміне рұқсат етіледі. Аталмыш санатқа тек құнын төлеп қана тіркеле алмайсыз, соған қоса форумға екі өзекті тақырыпта мақала жазуға тура келеді. Мақалада сіз автор болуыңыз керек, скриндермен әсерленген болуы шарт, және де басқа форумдарда жарияларбаған болуы қажет. Егер мақала көшірме, қайта жазу болғаны анықталса сізді өмір бойы тыйым салынған тізімге қосуы мүмкін. Member санатының өзі үшке бөлінеді: member, old member, premium member. Әр санаттың өзіндік ерекшеліктері бар. Жалпы форумда ақпараттық қауіпсіздік тақырыбында форум қатысушалырының жазған мақалаларына, әртүрлі бағыттағы қызметтер түрлеріне қол жеткізесіз. Форумда тек оң беделі бар адамдар ғана қатыса алады.

xss.is (DaMaGeLaB.IN) - Ағылшын тіліндегі даркнет хакарлік форумдарының арасында xss.is (бұрынғы DaMaGeLaB.IN) танымал болып табылады. Мұнда форум қолданушылары ақпараттық қауіпсіздікке қатысты, хакинг, кодинг, малвари және т.б. тақырыптарды зерттеу және талқылаумен айналысады. Аталмыш форумға тіркелу тегін. Тек форумға кіру үшін алдымен авторизация жасау шарт. Соған қоса форумдағы тақырыптар әр түрлі бөлімдерге бөлініп тұрады.

crdclub.su - Бұл кардерлер тақырыбына арналған ең көне форумдардың бірі. Тіркелу тегін, барлық бөлімдерге қол жетімді. Форумдағы мақалаларды оқу үшін, сізге міндетті түрде авторизация жасауға тура келеді. Форумның басқа форумдардан ерекшелігі Exploit.in секілді екі үлкен бөлікке бөлуге болады: орыс тілді және ағылшын тілді. Аталмыш форумда тек кардингке ғана емес, сонымен қатар скамға (алаяқтыққа), хакерлікке және әлеуметтік инженерия әдістеріне арналған жүздеген мақалалар бар.

Парсинг жүйесі жалпы үш негізгі бөліктен тұрады. Web-интерфейс, парсинг модулі және экспорт модулінен тұрады. Аталмыш бөліктердің бір-бірімен арақатынасы 2-суретте кескінделген.

Парсинг алгоритмі Python тілі BeautifulSoup4 + Requests кітапханаларында жазылды. Жалпы біздің парсер аталмыш форумдарға сұраулар жіберіп, соның нәтижесінде келген жауаптар бойынша жұмыс жасайды. Мысалы ретінде xss.is Darknet форумы алынды. Форумның сипаттамасы жоғарыда көрсетілген. Алдымен сайтқа авторизация қажет болғандықтан, форумға тіркелген логин және парольді жазып, форумға доступ аламыз. Одан кейін, негізгі беттегі тақырыптарға сұраулар жіберіледі. Бұл сұрауға келген жауапты global_topics.csv файлына жинаймыз. Сәйкесінше келесі әрекеттер де осылай жасалынады. Яғни, сұрау жіберіліп, содан келген жауапты жазу арқылы.

Сурет 2 – Парсинг жұмыс жасау сұлбасы

BeautifulSoup4 – HTML және XML файлдарынан деректерді шығаруға арналған Python кітапханасы. Бұл кітапхана HTML немесе XML анализаторының үстінде, талдау ағашын итерациялау, іздеу және өзгерту үшін Pythonic идиомаларын ұсынады. [18]

Requests — Python бағдарламалау тіліне арналған HTTP кітапханасы болып табылады. Бұл кітапхана негізінен HTTP сұрауларын қарапайым және қолданушыға түсінікті түрде ыңғайластырады. [19]

Сурет 3 **–** xss.is веб-форумының HTML бетінің үзіндісі

Форумға авторизация жасалынған соң, 1-суретте көрсетілген алғашқы бетке өтеміз. Осы беттегі басты тақырыптардың (жаһандық тақырыптар) URL мекенжайларын 4-суретте көрсетілген collect_global_topics() функциясы арқылы жинап, global_topics.txt файлына жазамыз.

```
def collect_global_topics():
    # 1. collect global topics url from main page and write to file
    url = SITE_URL
    soup = send_request(url)
    a_href = _get_href_list_by_parent_tag(soup, 'h3', 'node-title')
    print(a_href)
    write_to_file(f'{FOLDER_NAME}/global_topics.txt', a_href)
```

Сурет 4 – Басты беттен жаһандық тақырыптардың URL мекенжайын global_topics.txt файлға жинайтын функция

Келесі әрекет бізде жиналынған басты тақырыптар негізінде, әрқайсысының URL мекенжайларына кіріп, itterate_topics() функциясы арқылы барлық ішкі тақырыптардың URL мекенжайларын all_topics.txt файлына жинап аламыз. itterate_topics() функциясы 5-суретте көрсетілген.

```
def itterate_topics():
    # 2. read file with topics and recursevly open inner topics with comments
    with open('xss/global_topics.txt', 'r', encoding='utf-8') as f_read:
        for line in f_read.readlines():
            url = SITE_URL + line
            url = url.replace('\n', '')

            # iterate pages of one topic
            prev_page_content = None
            cur_page_content = ''
            all_pages_content = []
            page = 1
            while prev_page_content != cur_page_content:
                prev_page_content = cur_page_content
                soup = send_request(f'{url}page-{page}')
                cur_page_content = _get_href_list_by_parent_tag(soup, 'div', 'structItem-title')
                all_pages_content = all_pages_content + cur_page_content
                page+=1

            write_to_file(f'xss/all_topics.txt', all_pages_content, 'a+')
```

Сурет 5 – Ішкі тақырыптарды ашып, all_topics.txt файлға жинайтын функция

Хабарламаны жинаудың соңғы әрекеті ретінде all_topics.txt файлына жиналған URL мекенжайлардың барлығына жеке-жеке кіре отырып, барлық хабарламаларды get_comments_from_thread() функциясы арқылы, all_comments.csv файлына жазып аламыз. Аталмыш функция 6-суретте келтірілген.

```python
def get_comments_from_thread():
    # 3. open each thread and collect comments from them
    with open('xss/all_topics.txt', 'r', encoding='utf-8') as f_read:
        for line in f_read.readlines():
            url = line
            url = url.replace('\n', '')
            url = url.replace('unread','')
            # iterate pages of one topic
            prev_page_content = None
            cur_page_content = []
            all_pages_content = []
            page = 1
            while prev_page_content != cur_page_content:
                url_page = f'{SITE_URL}{url}page-{page}'
                prev_page_content = cur_page_content
                soup = send_request(url_page)
                cur_page_content = _get_all_comments_info(soup, 'article', re.compile('^message--post'))
                if prev_page_content == cur_page_content:
                    break
                all_pages_content = all_pages_content + cur_page_content
                page+=1

            write_to_csv('xss/all_comments.csv', all_pages_content, 'a+')
```

Сурет 6 – Ішкі тақырыптарды ашып, барлық хабарламаларды all_comments.csv файлына жинайтын функция

Нәтижесінде осы парсинг алгоритмі кез-келген Darknet форумдарынан хабарламалар жинауға мүмкіндік береді.

Хакерлер және кардерлер форумынан хабарламалар алуға арналған семантикалық белгілердің тізімі 2 өрістен тұрады. Олар: message_info (хабарламалар деректері) және user_info (пайдаланушы деректері). Хабарлама деректеріне: message_text, message_date, thread_name, ал пайдаланушы деректеріне: username_link, username, user_title, user_join_date, user_msg_number, user_reputation, user_reactions жатады. Семантикалық белгілер тізімінің сипаттамасы 1-кестеде көрсетілген.

Кесте 1 – Семантикалық белгілердің тізімі

№	Белгілер	Сипаттамасы
1	message_text	Форумда жазылған хабарлама/пікір
2	message_date	Форумдағы хабарламаның/пікірдің жазылған уақыты
3	thread_name	Форумдағы ішкі тақырыптың аты
4	username_link	Пайдаланушының форумдағы жеке кабинетіне сілтеме
5	username	Пайдаланушының аты
6	user_title	Пайдаланушының никнеймі
7	user_join_date	Пайдаланушының форумға тіркелу уақыты
8	user_msg_number	Пайдаланушының форумдағы хабарламалар саны
9	user_reputation	Пайдаланушының форумдағы басқа да пайдаланушылармен салыстырғандағы беделі
10	user_reactions	Пайдаланушының форумдағы хабарламаға басқа пайдаланушылардың білдірген реакциялар саны

Нәтижелер. Python тілінің BeautifulSoup4 және Requests кітапханаларын пайдаланып парсинг алгоритмі жазылды. Нәтижесінде форумдардан 1500000 - жуық хабарламалар жиналды. Әр форумның көрсеткіштері 2-кестеде көрсетілген.

Кесте 2 – Форумдардан жиналған хабарламалардың көрсеткіштері

№	Форум аты	Сілтемесі	global_topics	all_topics	message_number
1	Probiv	Probiv.one	82	50538	624983
2	XSS	xss.is	47	58170	758211
3	KOROVKA	korovka.cc	32	4890	31111
4	crdclub.su	crdclub.su	19	15308	54129
5	Эксплойт	Exploit.in	8	727	5159

Алынған нәтижелер көмегімен аталмыш форумдардан ақпараттық қауіпсіздіктің қатерлері мен осалдықтарын анықтауға мүмкіндік береді. BeautifulSoup4 және Requests кітапханалары көмегімен кез-келген интернет форумынан статикалық мәліметтер жинауға береді.

Талқылау. Ғаламтор беттеріндегі сайттардан деректер алудың екі әдісі бар, сайт бетінің HTML кодынан деректерді шығару және сайттың API пайдалану. Осы әдістерді салыстырып қарайық.

Кесте 3 – Парсинг әдістерін салыстыру

Парсер әдісі / критерий	HTML код арқылы	API арқылы
Артықшылықтары	бұл әдіс өте қарапайым және әрқашан жұмыс істейді, өйткені сайт бетінің html коды әрқашан пайдаланушыға қол жетімді	HTML код арқылы деректер алуға қарағанда жылдамырақ және html бет құрылымының өзгеруіне байланысты емес
Кемшіліктері	бұл әдіс бірнеше секунд ұзақ уақыт жұмыс істейді	барлық сайттарда ашық API жоқ, әсіресе даркнет сайттары болып саналатын хакерлік форумдарда.
Деректер алу уақыты (тек 1 бетті)	5-6 секунд	0,3-0,4 секунд

HTML код арқылы деректер алу үшін python тілінде BeautifulSoup4 және Scrapy кітапханаларын пайдалануға болады. Біз осы жұмыста BS4 қолдандық.

Қорытынды. Тұтастай алғанда, тақырыптық интернет-ресурстардан деректерді алу алгоритмін жасау өте күрделі міндет болуы мүмкін, ол мұқият жоспарлауды және көптеген қадамдарды орындауды қажет етеді. Дегенмен, дұрыс әзірленген және түзетілген алгоритм әртүрлі көздерден деректерді жинау процесін айтарлықтай жеңілдетеді және автоматтандырады.

Алгоритмді әзірлеу кезінде деректер көздерінің әртүрлі шектеулері мен мүмкіндіктерін ескеру, сондай-ақ жобаның мақсаттары мен міндеттеріне байланысты деректерді жинаудың ең тиімді әдістерін таңдау қажет.

Сонымен қатар, алгоритмді әзірлеу деректерді жинау және талдау жобасының кезеңдерінің бірі ғана екенін түсіну маңызды және сәтті дамығаннан кейін алынған деректерді өңдеу және талдау бойынша жұмысты жалғастыру қажет.

Қорыта келгенде тақырыптық интернет-ресурстардан деректерді алу үшін Python тіліндегі BeautifulSoup4 және Requests кітапханалары арқылы әзірленген алгоритм сипатталды. Зерттеу бес түрлі хакерлік және кардерлік форумдарын деректерді зерттеп, жинады және осы деректерден алынған жалпы ақпарат пен семантикалық мүмкіндіктерді қамтитын кестелерді ұсынды. Сонымен қатар, басқа да ақпаратты іздеу алгоритмдері талқыланып, алдыңғы авторлардың жұмысына шолу жасалды. Болашақта алынған деректерге сезімдік талдау жүргізу және классикалық машиналық оқыту әдістерін қолдана отырып, ақпараттық қауіпсіздік қатерлері мен осалдықтарын анықтау моделін құру жоспарлануда.

ӘДЕБИЕТ

1 Electronic resourse [data application: 26/01/2023] https://wearesocial.com/uk/blog/2022/01/digital-2022-another-year-of-bumper-growth-2/

2 Electronic resourse [data application: 28/01/2023] https://www.websiterating.com/ru/research/twitter-statistics/

3 Grune, D., Jacobs, C.J.H. (2008). Introduction. In: Parsing Techniques. Monographs in Computer Science. Springer, New York, NY. https://doi.org/10.1007/978-0-387-68954-8_1

4 Аканова, А., А.А. Макашев, С.А. Наурызбаева, & Н.Н.Оспанова. (2022). Моделирование тематического извлечения данных из интернета. Известия НАН РК. Серия физико-математическая, (3), 5–18. https://doi.org/10.32014/2022.2518-1726.137

5 A. Schulz, J. Lässig and M. Gaedke, “Practical Web Data Extraction: Are We There Yet? - A Short Survey,” 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI), 2016, pp. 562-567. https://doi.org/10.1109/WI.2016.0096.

6 E. C. Dallmeier, “Computer Vision-based Web Scraping for Internet Forums” 2021 7th International Conference on Optimization and Applications (ICOA), Wolfenbüttel, Germany, 2021, pp. 1-5. https://doi.org/10.1109/ICOA51614.2021.9442634.

7 Мамбетов С.Т., Бегимбаева Е.Е., Джолдасбаев С.К., Куламбаев Б.О. & Казбекова Г.Н., (2022). О мониторинге угроз и уязвимостей информационной системы. Известия НАН РК. Серия физико-математическая, (4), 68–80. https://doi.org/10.32014/2022.2518-1726.157

8 Audeh B, Beigbeder M, Zimmermann A, Jaillon P, Bousquet C (2017). Vigi4Med Scraper: A Framework for Web Forum Structured Data Extraction and Semantic Representation. PLOS ONE 12(1). https://doi.org/10.1371/journal.pone.0169658

9 Бекішев А.Т., Кумаргажанова С.К., Уркумбаева А.М., Танирбергенов А.Ж. Әлеуметтік желілердегі деректерді талдайтын заманауи ақпараттық жүйелерге шолу. «Университет еңбектері – Труды университета», №3 (88), 341-345. https://doi.org/10.52209/1609-1825_2022_3_341

10 V. Singrodia, A. Mitra and S. Paul, «A Review on Web Scrapping and its Applications,» 2019 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2019, pp. 1-6, https://doi.org/10.1109/ICCCI.2019.8821809

11 Fu, T., Abbasi, A., & Chen, H. (2010). A focused crawler for Dark Web forums. Journal of the American Society for Information Science and Technology, 61(6), 1213-1231. https://doi.org/10.1002/asi.21323

12 M. Schäfer, M. Fuchs, M. Strohmeier, M. Engel, M. Liechti and V. Lenders, "BlackWidow: Monitoring the Dark Web for Cyber Security Information," 2019 11th International Conference on Cyber Conflict (CyCon), Tallinn, Estonia, 2019, pp. 1-21. https://doi.org/10.23919/CYCON.2019.8756845

13 Das, D., & Semaan, B. (2020, October). quoras: A Python API for Quora Data Collection to Increase Multi-Language Social Science Research. In Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing (pp. 251-256).

14 Guo-Jun Qi, Xinyu Xing, Peng Liu, and Dinghao Wu. (2013) "Using Web Mining to Detect and Combat Cybercrime". IEEE Intelligent Systems journal, Volume 28, Issue 3, on pages 18-25. https://doi.org/10.1109/MIS.2013.25

15 Electronic resourse [data application: 26/01/2023] https://www.scrapinghub.com/web-scraping-vs-api-whats-the-difference/

16 Mohammad Masudur Rahman, Weiwei Sun и Guangyan Huang (2021). Machine Learning-based Web Scraping Framework for data Extraction from Web Pages. IEEE Access. https://doi.org/10.1109/ACCESS.2021.3069734

17 Электронды ресурс [жүгінген күні: 22.02.2023] https://ru.wikipedia.org/wiki/Tor

18 Электронды ресурс [жүгінген күні: 22.02.2023] https://pypi.org/project/beautifulsoup4/

19 Электронды ресурс [жүгінген күні: 22.02.2023] https://pypi.org/project/requests/

REFERENCES

1 Electronic resourse [data application: 26/01/2023] https://wearesocial.com/uk/blog/2022/01/digital-2022-another-year-of-bumper-growth-2/

2 Electronic resourse [data application: 22/02/2023]: https://www.websiterating.com/ru/research/twitter-statistics/

3 Grune, D., Jacobs, C.J.H. (2008). Introduction. In: Parsing Techniques. Monographs in Computer Science. Springer, New York, NY. https://doi.org/10.1007/978-0-387-68954-8_1

4 Akanova A., A.A. Makashev, S.A. Nauryzbayeva, & N.N.Ospanova. (2022). MODELIROVANIE TEMATICHESKOGO IZVLECHENIIA DANNYH IZ INTERNETA. Izvestiia NAN RK. Seriia fiziko-matematicheskaia, (3), 5–18. https://doi.org/10.32014/2022.2518-1726.137

5 A. Schulz, J. Lässig and M. Gaedke, "Practical Web Data Extraction: Are We There Yet? - A Short Survey," 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI), 2016, pp. 562-567. https://doi.org/10.1109/WI.2016.0096.

6 E. C. Dallmeier, "Computer Vision-based Web Scraping for Internet Forums" 2021 7th International Conference on Optimization and Applications (ICOA), Wolfenbüttel, Germany, 2021, pp. 1-5. https://doi.org/10.1109/ICOA51614.2021.9442634.

7 Mambetov S.T., Begimbayeva Ye.Ye., Joldasbayev S., Kulambayev B.O., & Kazbekova G., (2022). O MONITORINGE UGROZ I UYAZVIMOSTEI INFORMATSIONNOI SISTEMY. Izvestiia NAN RK. Seriia fiziko-matematicheskaia, (4), 68–80. https://doi.org/10.32014/2022.2518-1726.157

8 Audeh B, Beigbeder M, Zimmermann A, Jaillon P, Bousquet C (2017). Vigi4Med Scraper: A Framework for Web Forum Structured Data Extraction and Semantic Representation. PLOS ONE 12(1). https://doi.org/10.1371/journal.pone.0169658

9 Bekishev A.T., Kýmargajanova S.K., Ýrkýmbaeva A.M., Tanırbergenov A.J. Aleumettik jelilerdegi derekterdi taldaıtyn zamanauı aqparattyq juıelerge sholu. «Unıversıtet enbekteri – Trudy unıversıteta», №3 (88), 341-345. https://doi.org/10.52209/1609-1825_2022_3_341

10 V. Singrodia, A. Mitra and S. Paul, "A Review on Web Scrapping and its Applications," 2019 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2019, pp. 1-6, https://doi.org/10.1109/ICCCI.2019.8821809

11 Fu, T., Abbasi, A., & Chen, H. (2010). A focused crawler for Dark Web forums. Journal of the American Society for Information Science and Technology, 61(6), 1213-1231. https://doi.org/10.1002/asi.21323

12 M. Schäfer, M. Fuchs, M. Strohmeier, M. Engel, M. Liechti and V. Lenders, "BlackWidow: Monitoring the Dark Web for Cyber Security Information," 2019 11th International Conference on Cyber Conflict (CyCon), Tallinn, Estonia, 2019, pp. 1-21. https://doi.org/10.23919/CYCON.2019.8756845

13 Das, D., & Semaan, B. (2020, October). quoras: A Python API for Quora Data Collection to Increase Multi-Language Social Science Research. In Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing (pp. 251-256).

14 Guo-Jun Qi, Xinyu Xing, Peng Liu, and Dinghao Wu. (2013) "Using Web Mining to Detect and Combat Cybercrime". IEEE Intelligent Systems journal, Volume 28, Issue 3, on pages 18-25. https://doi.org/10.1109/MIS.2013.25

15 Electronic resourse [data application: 26/01/2023] https://www.scrapinghub.com/web-scraping-vs-api-whats-the-difference/

16 Mohammad Masudur Rahman, Weiwei Sun и Guangyan Huang (2021). Machine Learning-based Web Scraping Framework for data Extraction from Web Pages. IEEE Access. https://doi.org/10.1109/ACCESS.2021.3069734

17 Electronic resourse [data application: 22/02/2023] https://ru.wikipedia.org/wiki/Tor

18 Electronic resourse [data application: 22/02/2023] https://pypi.org/project/beautifulsoup4/

19 Electronic resourse [data application: 22/02/2023] https://pypi.org/project/requests/

С. Т. МАМБЕТОВ[1], Е. Е. БЕГИМБАЕВА[1,3], А. К. ХИКМЕТОВ[2], С. К. ДЖОЛДАСБАЕВ[2]

[1]Казахский национальный университет имени аль-Фараби, Алматы, Казахстан
[2]Международный университет информационных технологий, Алматы, Казахстан
[3]Казахский национальный исследовательский технический университет имени К.Сатпаева, Алматы, Казахстан
e-mail: mambetov.saken@gmail.com, enlik89@mail.ru, akhikmetov@iitu.edu.kz, serikdzoldasbaev@gmail.com

РАЗРАБОТКА АЛГОРИТМА ПОЛУЧЕНИЯ ДАННЫХ С ТЕМАТИЧЕСКИХ ИНТЕРНЕТ-РЕСУРСОВ

С бурным развитием Интернета пользователи активно делятся своими личными данными и другой информацией во многих социальных сетях. Информация в Интернете должна быть проанализирована, чтобы убедиться, что она надежна и не представляет угрозы для общественности. Исходя из этого, возникает необходимость сбора, мониторинга и анализа этой информации. Сбор данных — сложная задача, зависящая от структуры каждой веб-страницы. Так как не все ресурсы позволяют собирать информацию, приходится использовать множество методов.

В предлагаемой статье показаны эффективные способы использования синтаксического анализа для получения информации. Объясняется метод семантического анализа (парсинга) содержимого веб-страниц с помощью программы, написанной на языке Python на основе библиотеки BeatifulSoup. Кроме того, основное внимание уделяется методам сбора информации через другие API, с использованием инструментов для эмуляции поведения пользователя в браузере. Представлен алгоритм извлечения ин-

формации из тематических интернет-ресурсов с использованием библиотеки BeatifulSoup + Requests. В результате была получена информация с англо- и русскоязычных хакерских и кардинговых форумов.

***Ключевые слова**: beautifulsoup, requests, работа с данными, парсинг, python, хакерские форумы*

S. MAMBETOV[1], YE. BEGIMBAYEVA[1,3], A. KHIKMETOV[2], S. JOLDASBAYEV[2]

[1]Al Farabi Kazakh National University, Almaty, Kazakhstan
[2]International IT University, Almaty, Kazakhstan
[3]Kazakh National Research Technical University named after K.Satpayev, Almaty, Kazakhstan
e-mail: mambetov.saken@gmail.com, enlik89@mail.ru, akhikmetov@iitu.edu.kz, serikdzoldasbaev@gmail.com

DEVELOPMENT OF AN ALGORITHM FOR OBTAINING DATA FROM THEMATIC INTERNET RESOURCES

With the rapid development of the Internet, users are actively sharing their personal data and other information on many social networks. Information on the Internet should be analyzed to make sure that it is reliable and does not pose a threat to the public. Based on this, there is a need to collect, monitor and analyze this information. Data collection is a complex task, depending on the structure of each web page. Since not all resources allow you to collect information, you have to use many methods.

The proposed article shows effective ways of using syntactic analysis to obtain information. The method of semantic analysis (parsing) of the contents of web pages is explained using a program written in Python based on the BeatifulSoup library. In addition, the focus is on methods of collecting information through other APIs, using tools to emulate user behavior in the browser. An algorithm for extracting information from thematic Internet resources using the BeatifulSoup + Requests library is presented. As a result, information was obtained from English- and Russian-speaking hacker and carding forums.

***Key words**: beautifulsoup, requests, working with data, parsing, python, hacker forums*