EURASIAN NATIONAL UNIVERSITY

ҚАЗАҚСТАН РЕСПУБЛИКАСЫ
БІЛІМ ЖӘНЕ ҒЫЛЫМ МИНИСТРЛІГІ
Л.Н. ГУМИЛЕВ АТЫНДАҒЫ ЕУРАЗИЯ
ҰЛТТЫҚ УНИВЕРСИТЕТІ

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РЕСПУБЛИКИ КАЗАХСТАН
ЕВРАЗИЙСКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ
ИМЕНИ Л.Н. ГУМИЛЕВА

MINISTRY OF EDUCATION AND SCIENCE
OF THE REPUBLIC OF KAZAKHSTAN
L.N. GUMILYOV EURASIAN NATIONAL UNIVERSITY

16-18 маусым
Нұр-Сұлтан, 2022

«TURKLANG 2022»

«Түркі тілдерін компьютерлік өңдеу»
атты X халықаралық конференция
ЕҢБЕКТЕРІ

ТРУДЫ
X Международной конференции
«Компьютерная обработка тюркских языков»

«TURKLANG 2022»

PROCEEDINGS
of the X International Conference
on Computer processing of Turkic Languages

«TURKLANG 2022»

ҚАЗАҚСТАН РЕСПУБЛИКАСЫ
БІЛІМ ЖӘНЕ ҒЫЛЫМ МИНИСТРЛІГІ
Л.Н. ГУМИЛЕВ АТЫНДАҒЫ ЕУРАЗИЯ ҰЛТТЫҚ УНИВЕРСИТЕТІ

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РЕСПУБЛИКИ КАЗАХСТАН
ЕВРАЗИЙСКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ
ИМЕНИ Л.Н. ГУМИЛЕВА

MINISTRY OF EDUCATION AND SCIENCE OF
THE REPUBLIC OF KAZAKHSTAN
L.N. GUMILYOV EURASIAN NATIONAL UNIVERSITY

«TURKLANG 2022»
«Түркі тілдерін компьютерлік өңдеу»
атты X халықаралық конференция
ЕҢБЕКТЕРІ
16-18 маусым 2022 ж.

ТРУДЫ
X Международной конференции
«Компьютерная обработка тюркских языков»
«TURKLANG 2022»
16-18 июня 2022 г.

PROCEEDINGS
of the X International Conference
on Computer processing of Turkic Languages
«TURKLANG 2022»
16-18 June 2022

Нұр-Сұлтан, 2022

УДК 80/81:004
ББК 81.2:32-973
Т 90

Техникалық редакция:
Ергеш Б.Ж.
Елибаева Г.К.
Турсынова Н.А.

Т 90 ТҮРКІ ТІЛДЕРІН КОМПЬЮТЕРЛІК ӨҢДЕУ. X халықаралық конференция: Еңбектері = КОМПЬЮТЕРНАЯ ОБРАБОТКА ТЮРКСКИХ ЯЗЫКОВ. X международная конференция: Труды. / - Нұр-Сұлтан: «Булатов А.Ж.» ЖК, 2022.= Нур-Султан: ИП «Булатов А.Ж.»

ISBN 978-601-326-645-9

Жинақта «Түркі тілдерін компьютерлік өңдеу» атты X халықаралық конференция қатысушыларының баяндамалары енген.

Компьютерлік лингвистика бағыты бойынша оқитын студенттерге, магистранттарға, докоранттарға және мамандарға арналған.

Жинақ «BR11765535» Қазақ тілі мәдениетін арттыру және функцияларды кеңейту бойынша ғылыми-лингвистикалық негіздер мен IT-ресурстарды әзірлеу» бағдарламасы есебінен жарияланды.

В сборнике представлены доклады участников X международной конференции «Компьютерная обработка тюркских языков».

Предназначен для студентов, магистрантов, докторантов и специалистов специализирующихся в областях компьютерной лингвистика.

Сборник издан за счет средств программы BR11765535 «Разработка научно-лингвистических основ и IT-ресурсов по расширению функций и повышению культуры казахского языка».

УДК 80/81:004
ББК 81.2:32-973

ISBN 978-601-326-645-9

ӘОК 004.89.
[1]Леспекова А.А., [2]Муканова А.С., [3]Елибаева Г.К.
[1, 3] Л.Н. Гумилев атындағы Еуразия ұлттық университеті,
[2]Астана Халықаралық университеті,
Қазақстан, Нұр-Сұлтан,
[1]azizalespekova1998@gmail.com, [2]asiserikovna@gmail.com ,
[3]gaziza_y@mail.ru

ТЫЙЫМ САЛЫНҒАН КОНТЕНТТІ АНЫҚТАУ ҮШІН МӘТІНДІК КОРПУС ҚҰРУ

Аңдатпа: Internet технологияларының дамуына байланысты желіде адам өміріне қауіп тудыратын және мемлекетпен рұқсат етілмеген ақпараттар көптеп таралып жатыр. Сайттардың саны халықтың жартысынан да көп және тез тарауда. Сондықтан ақпараттың үлкен көлемін өңдеу қажеттілігі туындауда және ол күрделі жұмыс. Бұл мәселені ішінара шешуге қазіргі уақытта белсенді түрде құрылған мәтіндер корпусы қызмет етеді. Бұл жұмыста тыйым салынған контентті анықтау үшін қажетті мәтіндік корпусты құру қарастырылады.

Кілттік сөздер: мәтіндер корпусы, тыйым салынған контент, интернет

УДК: 004.89.
[1]Леспекова А.А., [2]Муканова А.С., [3]Елибаева Г.К.
. [1,3]Евразийский национальный университет Л.Н. Гумилева
[2]Международный университет Астана
Нур-Султан, Казахстан,
[1]azizalespekova1998@gmail.com , [2]asiserikovna@gmail.com ,
[3]gaziza_y@mail.ru

СОЗДАНИЕ ТЕКСТОВОГО КОРПУСА ДЛЯ ОБНАРУЖЕНИЯ ЗАПРЕЩЕННОГО КОНТЕНТА

Аннотация: В связи с развитием технологий Internet в сети все больше распространяется информация, представляющая опасность для жизни людей и не разрешенная государством. Количество сайтов быстро растет. Поэтому возникает необходимость обработки большого объема информации, и это сложная работа. Частичному решению этой проблемы служит активно созданный в настоящее время корпус текстов. В данной работе рассматривается создание текстового корпуса,

необходимого для обнаружения запрещенного контента.

Ключевые слова: тексты, запрещенный контент, интернет.

UDC 004.89.

[1]Lespekova A., [2]Mukanova A., [3]Yelibayeva G.
[1, 3]L. N. Gumilyov Eurasian National University,
[2]Astana International University,
Kazakhstan, Nur-Sultan,
[1]azizalespekova1998@gmail.com, [2]asiserikovna@gmail.com ,
[3]gaziza_y@mail.ru

CREATING A TEXT CORPUS TO IDENTIFY PROHIBITED CONTENT

Abstract: In connection with the development of Internet technologies, a large number of information that poses a threat to human life and is not authorized by the state is being distributed on the network. The number of sites is more than half of the population and is rapidly gaining popularity. Therefore, there is a need to process a large amount of information, and this is a complex work. A partial solution to this problem is the currently actively created corpus of texts. This paper examines the creation of the necessary text corpus to identify prohibited content.

Keywords: text corpus, prohibited content, internet

Кіpспе

Мақалада тыйым салынған мәтіндерді анықтау әдістерін оқыту және тестілеу үшін мәтіндер корпусы сипатталды. Сонымен қатар қазақ тіліндегі тыйым салынған мәтіндерге немесе заңсыз мазмұндағы материалдар жиынтығы жасалды. Жинақталған мәтіндер көмегімен тыйым салынған контентті анықтау көрсетілді.

Тыйым салынған контент – бұл мемлекеттен тыйым салынған ақпараттық ресурстың немесе веб-сайттың кез келген деректерін адамдарға көшіруге, таратуға және көруге рұқсат етілмеген мазмұнды айтамыз [1-2]. Тыйым салынған контентке ұятсыз мәтіндер, мультимедиа, құмар ойындар, митингке шақыртулар, терроризм ұйымдастырушылық, қатыгездік, кісі өлтіру және т.б. тақырыптары бар желідегі ақпаратты жатқызамыз. Бұл тақырыпқа сәйкес барлық ақпараттар бақылауға алынады, ал қауіпті деп танылса бірден бұғатталады. Тыйым салынған ақпаратты таратушылар заң бұзғаны үшін айыппұл төлеуі керек, ал қасақана ұйымдастырылған жағдайда қауіпсіздік күшейтіліп, қолайсыз мазмұнды насихаттағаны үшін

қылмыстық жауапкершілікке тартылады.

Порнография және жыныстық қанағаттануға арналған барлық деректер немесе сексуалдық сипаттағы қызмет көрсетуді насихаттайтын ақпараттарды жариялауға тыйым салынады. Сонымен қатар, ақша үшін жыныстық қызметтерді ұсынатын қосымшаларға да мемлекетпен тыйым салынады.

Нәсілдік, ұлттық, діни, жыныстық, мүгедектік, ардагер мәртебесі, жыныстық бағдар, гендерлік сәйкестілік және басқа да белгілер негізінде кез-келген адамдар мен әлеуметтік топтарға зорлық-зомбылықты насихаттайтын қосымшаларды жариялауға мемелекетпен тыйым салынады

Жарылғыш заттарды, атыс қаруын, патрондарды, атыс қаруына арналған кейбір бөлшектерін сатып алуға болатын қосымшаларды жариялауға мемлекет рұқсатынсыз сатуға тыйым салынады.

Тыйым салынған контентті анықтаудың негізгі мақсаты - заңсыз таратуға тыйым салынған ақпаратты қамтитын "Интернет" желісіндегі сайттарға кіруді шектеу.

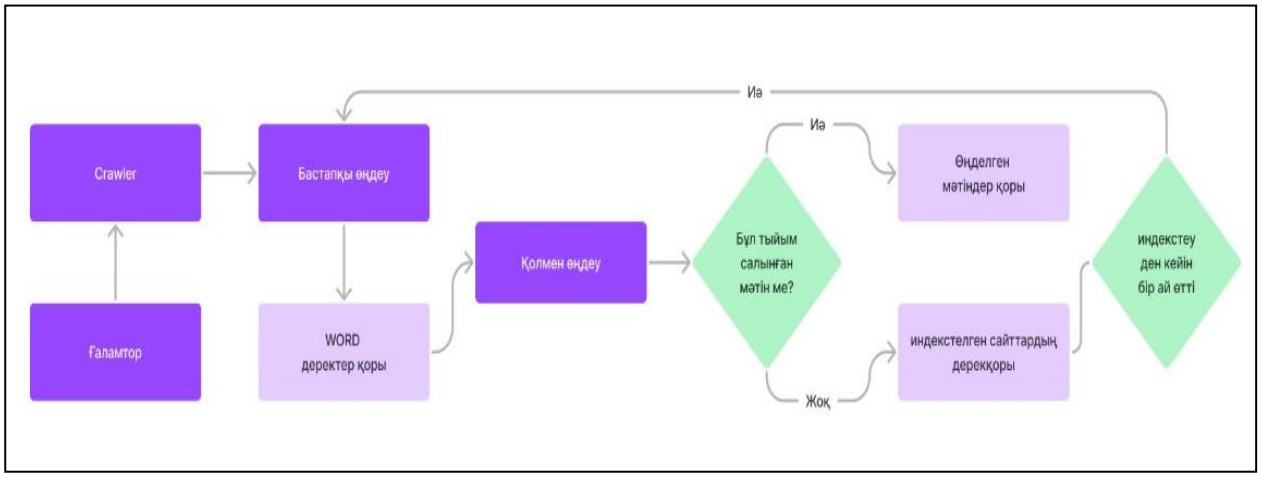
Мәтіндік корпус – сөйлемдерді және лингвистикалық ақпараттарды басқару және жинақтау жүйесі [3]. Оны көбінесе біздер корпусты басқару жүйесі немесе корпустық менеджер деп атаймыз. Бұл корпустағы сөйлемдер мен сөздерді іздеуге, бізге қажетті ақпаратты алуға негізделген жүйе. Корпустарды құрудың мақсаты мен пайдаланудың мәні келесі алғышарттармен анықталады:

– бір рет құрылған және дайындалған мәтіндік корпусты бірнеше рет, әр түрлі зерттеулерде және бірнеше жеке мақсаттарда пайдалануға мүмкіндік береді;

– әртүрлі тоналдылыққа ие деректер корпуста өзінің шынайы контекстік формасында болады. Бұл оларды кеңінен және жан-жақты зерттеуге мүмкіндікті тудырады;

Жоғарғы репрезентативтіктілікке ие толық өңделген корпусымыз, деректердің шынайылығана ақпараттың дұрыстығына кепілдік береді.

Бұл жұмыста сипатталатын корпус жартылай автоматты түрде жасалынған. Ол үшін бірнеше модульдерді қолданатын боламыз. Олар: Crawler, requestGenerator, pagePreprocessing. Мәтіндік корпус архитектурасы төмендегідей. Тыйым салынған контентті анықтауға мүмкіндік беретін мәтіндік корпус архитектурасы ашып көрсетілген.



Сурет 1 – Тыйым салынған контентті анықтауға арналған мәтіндік корпус архитектурасы

Figure 1- Text corpus architecture for detecting prohibited content

Суретте мәтіндік корпус архитектурасының сипаттамасы берілген:

– мәтіндік корпуста ең бірінші crawler кілттік сөздер бойынша сайттарды іздей бастайды.

– екінші бастапқы өңдеу басталады және html тегтері жойылып кіші регистрге жазылады.

– өңделген мәтін WORD деректер базасына жазылады.

– адамдар WORD деректер базасындағы мәтіндердің мағынасына қарай тыйым салынғандыққа анықтайды.

– егер мәтін мағынасына қарай тыйым салынған болса, негізгі деректер базасына қосылады.

– егер мәтін мағынасына қарай тыйым салынбаған болса, индекстелген сайттар деректер базасына қосылады.

Егер сайт индекстелген сайт деректер базасында 30 күннен артық күн жатса, ол бастапқы өңдеуге қайта қосылады. Оның мәні – сайтқа қауіпті ақпарат қосылмағанына көз жеткізу болып табылады.

Мәтіндер корпусында заңсыз мәтіндер жеті санатқа жіктеледі: терроризм, идеологиялық мәтіндер, діни өшпенділік, сепаратизм, ұлтшылдық, агрессия және тәртіпсіздікке шақыру, фашизм, сондай-ақ ұқсас лексикасы бар бейтарап мәтіндер. Тыйым салынған контентті мәтіндік корпус арқылы анықтауға жүргізілген жұмыстар:

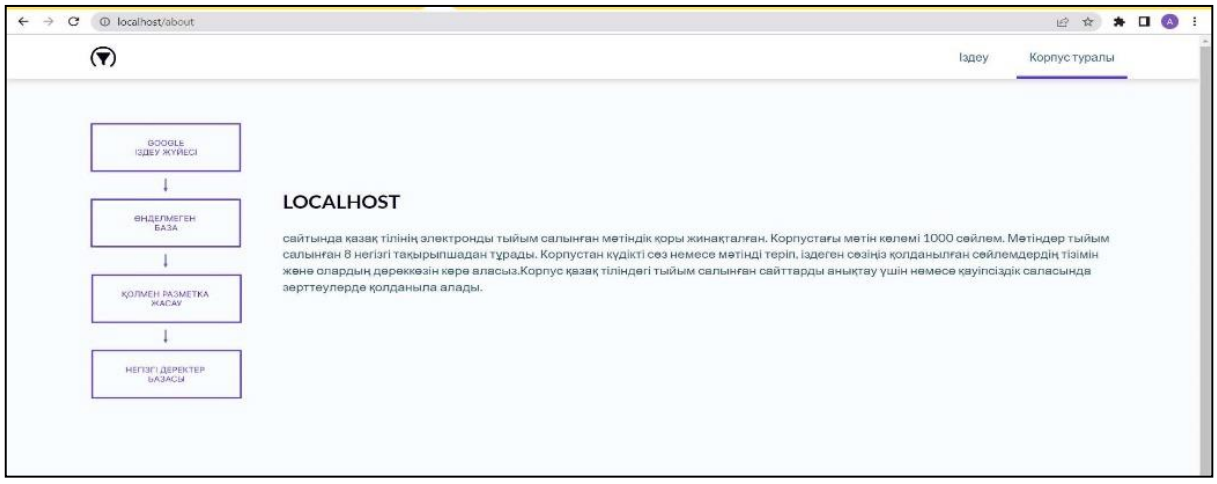
– Ағылшын тіліндегі экстремистік мәтіндердің дайын корпусы сипатталған [4]. Корпусы (Narrative Networks Corpus) діни әңгімелерден, интернеттегі материалдардан және жарнамалық журналдардан алынған исламистік экстремизмге қатысты 100 мәтінді құрайды (42 480 сөз). Барлық мәтіндер корпусындағы сөйлемдер араб тілінде жазылған. Бірақ қазіргі уақытта толығымен ағылшын тіліне аударылып жазылды.

Корпуста сөйлемдерді тоналдылыққа талдау жүйелері бар. Ол автоматты түрде жинақталып, содан кейін қолмен жеке тексерілді.

– Орыс тіліндегі экстремистік мәтіндердің дайын корпусы сипатталған [5-8]. Жұмыста экстремистік бағыттағы мәтіндерді анықтау әдістерін оқыту және тестілеу үшін мәтіндер корпусы жасалған. Қазіргі уақытта корпустың жалпы көлемі – 493 мәтін (650 000 сөз), оның ішінде 368 мәтін экстремистік материалдар санатына жатады. Барлық мәтіндер қолмен жиналған.

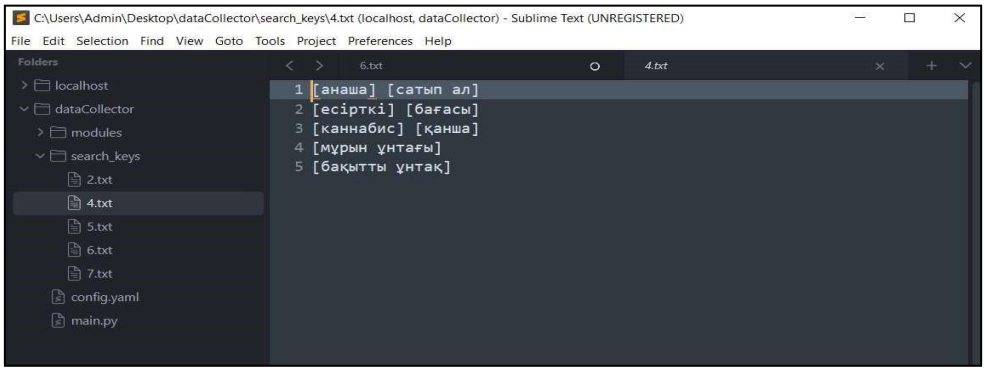
– Қазақ тіліндегі WEB-ресурстарда экстремистік бағытты анықтау үшін түйінді сөздер жинағын құру сипатталған [9]. Бұл жұмыста мәтіндегі тыйым салынған контентті анықтау жүзеге асыру сипатталған. Осы мәселені шешу негізгі бес кезеңге бөлінген: тыйым салынған аймақтың веб-сайттарын анықтау, мәліметтерді алуға дайындықты қарастыру, мәліметтерді өндеп алу, мәліметтерді бөлу және талдау.

Тыйым салынған контентті анықтау үшін мәтіндік корпустың жұмысын талдамас бұрын корпустың өзіне тоқталып өтейік. Корпус тыйым салынған сөйлемдер бойынша белгілі бір категорияларға бөлініп жинақталған. Сөйлемдер интернет желісінен ізделініп, өңделмеген базаға тіркеледі. Әрі қарай қолмен разметка жасалып өзінің категориясына анықталады. Негізгі базаға анықталған категория бойынша тіркеледі. Төмендегі суретте корпустың сайттағы сипаттасы көрсетілген. Localhost сайтында қазақ тілінің электронды тыйым салынған мәтіндік қоры жинақталған. Корпустағы мәтін көлемі 1000 сөйлем. Сипаттамада гугл жүйесіне іздеу жүйесінен мәтіндер өңделмеген базаға тіркеледі. Қолмен анықтау жүйесінде мәтіндерді өңделген базаға тіркеледі.



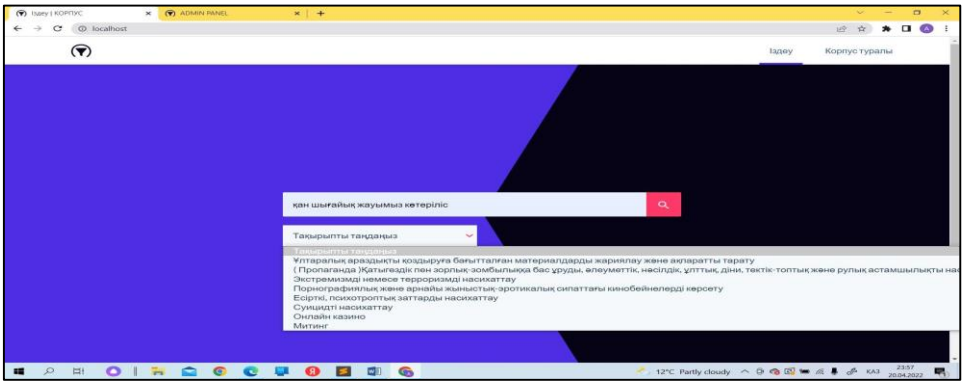
Сурет 2 – Корпустың сайттағы сипаттамасы
Figure 2-Description of the building on the site

Мәтіндер тыйым салынған 7 негізгі тақырыпшадан тұрады. Корпустан күдікті сөз немесе мәтінді теріп, іздеген сөзіңіз қолданылған сөйлемдердің тізімін және олардың дереккөзін көре аласыз. Корпус қазақ тіліндегі тыйым салынған сайттарды анықтау үшін немесе қауіпсіздік саласында зерттеулерде қолданыла алады.



Сурет 3 – Кілттік сөздердің жинақталуы
Figure 3-Accumulation of key words

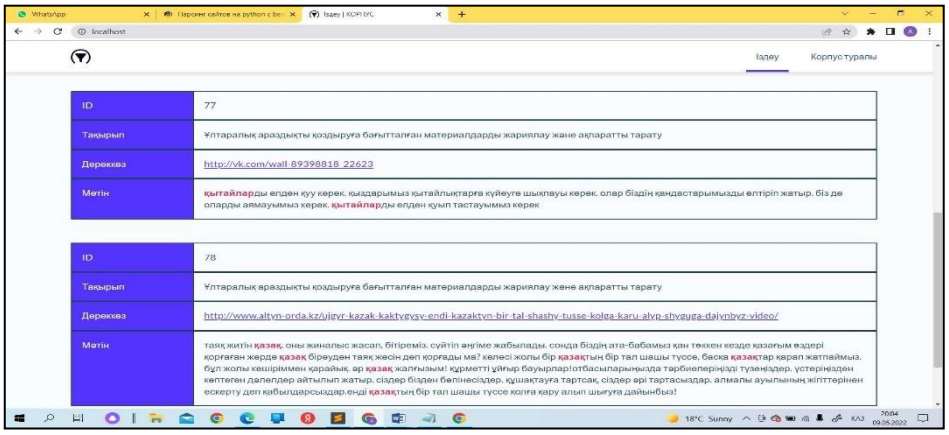
Жоғарыдағы суретте тыйым салынған сөздердің әрбір категориясына кілттік сөздерді жинақтадық. Экраннан көріп тұрғаныңыздай, ол 7 негізгі мәтіндік форматқа бөлінген. Ол жерде кілттік сөздер немесе дәлдік сөйлемдер арқылы да іздестіруімізге болады. Әрбір сөз тік жақша көмегімен бөлінген түрде жазылады. Әрбір жолдағы сөздер интернет желісіне тек бір сұрау болып жүргізіледі.



Сурет 4 – Мәтіндік корпустың қолданылуы
Figure 4-Using a text corpus

Мәтіндік корпусты қолдану үшін өзімізге қажетті қауіпті деп танылатын сөздер немесе сөйлемдерді енгіземіз. Енгізілген ақпаратқа сәйкес қай категорияға жататынын (порнографиялық және арнайы жыныстық-эротикалық сипаттағы кинобейнелерді көрсету ұлтаралық

араздықты қоздыруға бағытталған материалдарды жариялау және ақпаратты тарату қатыгездік пен зорлықзомбылық жасауды насихаттау әлеуметтік, нәсілдік, ұлттық, діни, тектіктоптық және рулық астамшылықты насихаттау экстремизмді немесе терроризм жолына түсуге үгіттеу есірткі сияқты психотроптық заттарды насихаттау суйцидті насихаттау және лицензиясы жоқ онлайн казино, митингке үгіттеу) таңдаймыз. Соңында іздеу батырмасын басамыз.



Сурет 5 – Ұлтаралық қақтығысқа категориясы бойынша табылған мәліметтер

Figure 5 - Found data on the category of interethnic conflict

5-суреттен ұлтаралық араздықты қоздыруға бағытталған материалдарды жариялау және ақпаратты тарату тақырыбына қатысты сөздер бойынша табылған сайттардың ақпараттары көрсетілген. Басқа ұлт өкілдеріне қарсы сөздер және өзге ұлтты елімізден қуып шығу секілді ақпараттар жинақталды. Қазақ ұлтына қатысты кері айтылған ақпараттар да тіркелді. Мысалы:

77 номерде қытай халқын арандату туралы;

78 номерде қазақ халқын арандату туралы ақпараттар тіркелді.

Қорытындылай келе бұл мақалада интернет желісінен тыйым салынған контентті мәтіндік корпус арқылы анықтау жүйесінің архитектурасы толығымен сипатталды. Тыйым салынған контентті анықтауға мүмкіндік беретін мәтіндік корпусты құрудың жолы қарастырылды.

Әдебиеттер тізімі

1 Ельчанинова Н.Б. Проблемы совершенствования законодательства в сфере ограничения доступа к противоправной инфорамции в сети Интернет// Общество: политика, экономика, право-2017.-№12. – С. 119-121

2 Марценюк А.Г. Запрещенная информация и ее место в системе информационных отношений// Гражданин и право-2018. - № 5-С.62

3 Захаров В.П., Богданова С.Ю. Корпусная лингвистика: Учебник для студентов направления «Лингвистика». 2-е изд., переработанный и дополненный., – СПб.: СПбГУ. РИО. Филологический факультет, 2013. – 148 с.

4 learning techniques for sentiment classification. InACL. The Association for Computer Linguistics

5 Богуславский И. М. и др. Аннотированный корпус русских текстов: концепция, инструменты разметки, типы информации” // Труды Международного семинара по компьютерной лингвистике и её приложениям "Диалог-2000". Протвино, 2000.

6 Корпусная лингвистика и контекст (в соавт. с Ю. Н. Марчуком) // Межвузовский сборник научных трудов "Теоретические и практические аспекты лингвистики и лингводидактики". - Сургут: Изд-во СурГУ, 2002. - С. 123-128.

7 Рубцова Ю. В. Построение корпуса текстов для настройки тонового классификатора / Ю. В. Рубцова // Программные продукты и системы. –2015. – № 1. – С. 72–78

8 Захаров, В. П. (2015). Оценка качества Интернет-корпусов русского языка. В Труды международной конференции «Корпусная лингвистика2015» (стр. 218-229). Издательство Санкт-Петербургского университета

9 Bolatbek M. A., Mussiraliyeva S. Z., Tukeyev U. A. Creating the dataset of keywords for detecting an extremist orientation in web-resources in the Kazakh language // KazNU Bulletin. Mathematics, Mechanics, Computer Science Series.