ИНФОРМАТИКА, IT-ТЕХНОЛОГИЯЛАР
ИНФОРМАТИКА, IT-ТЕХНОЛОГИИ
COMPUTER SCIENCE, INFORMATION TECHNOLOGIES

ӘОЖ – 004.89

Д.А. Абдраманов, Ж.Д. Изтаев*, С.Ж. Құракбаева, И.Қ. Байназарова
магистрант, М.Әуезов атындағы ОҚУ, Шымкент, Қазақстан
п.ғ.к., доцент, М.Әуезов атындағы ОҚУ, Шымкент, Қазақстан
т.ғ.к, профессор, М.Әуезов атындағы ОҚУ, Шымкент, Қазақстан
магистр,аға оқытушы, М.Әуезов атындағы ОҚУ, Шымкент, Қазақстан
***Корреспондент авторы:** Zhalgasbek71@mail.ru

ВЕБ-КӨЗДЕРДЕН ҚЫЛМЫСТЫҚ КОНТЕНТ ДЕРЕКТЕРІН ЖИНАУ ЖӘНЕ ДАЙЫНДАУ

Түйін

Қылмыстарды жоспарлау және оған шақыру, жалған ақпаратпен бөлісу сияқты қылмыстық мәтіндер желілік ортадағы қауіпсіздікке қауіп төндіреді. Мұндай криминалды мәтіндерді анықтау және жіктеу желідегі қылмыспен күрестің құрамдас бөлігіне айналуда. Желіде қолжетімді ақпарат көлемінің ұлғаюына және Интернетке қатысты заңға қарсы әрекеттер көбеюіне байланысты қылмыстық мәтіндерді автоматты түрде анықтау және саралаудың тиімді әдістері мен тәсілдерін әзірлеу қажет.

Қылмыстық мәтіндерді жіктеу есептерінде қолданылатын тәсілдерінің бірі морфологиялық талдау әдістерін қолдану болып табылады. Морфологиялық талдау сөздердің құрылымын, олардың грамматикалық формаларын, лексикалық және синтаксистік ерекшеліктерін талдауға мүмкіндік береді. Бірақ та қылмыстық мәтіндердің өзіндік ерекшеліктері бар, сол себептен морфологиялық талдаудың қолданыстағы әдістері оларды жіктеуде әрқашан тиімді бола бермейді. Осы себептен дәлдікті жоғарылату мен барынша шынайы нәтижеге қол жеткізуде қолданыстағы әдістердің түрлендіру мен жетілдіру тапсырмасы туындап отыр.

Кілттік сөздер: Веб-контент, Scikit-Learn, NLTK, TensorFlow, Python, Jupyter Notebook, BeautifulSoup(BS4),XML,HTML, машиналық оқыту.

Кіріспе

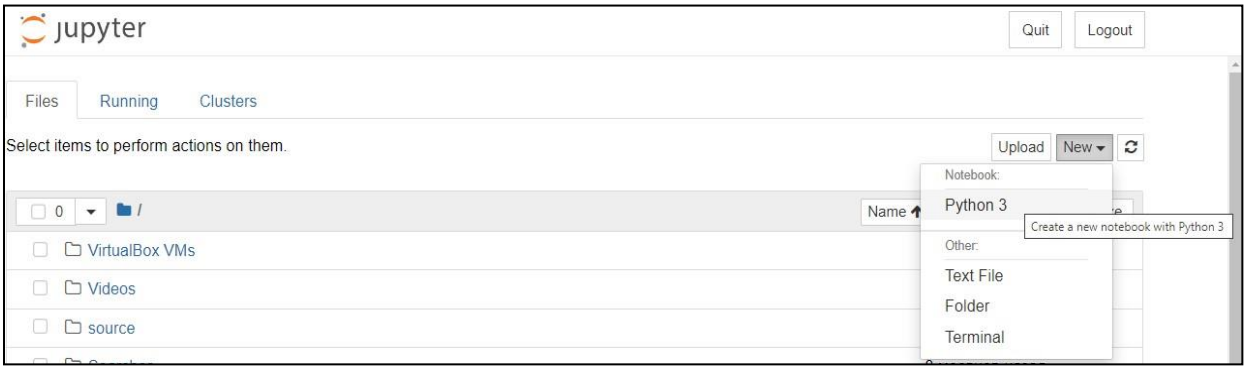
Веб-контенттегі қылмыстық мәтіндерді жіктеу үлгілерін зерттеу және әзірлеу үшін сәйкес деректер жиынтығын пайдалану қажет. Деректер жинағы – тиісті сынып белгілерімен белгіленген мәтіндік мысалдарды қамтитын деректер жиыны. Деректерді жинау және дайындау веб-контенттегі қылмыстық мәтіндерді жіктеу үдерісіндегі басты кезеңі болып есептеледі.

Жіктеу сапасы модель оқытылатын деректердің сапасына тікелей байланысты. Егер деректер толық емес, шулы немесе бұрмаланған болса, жіктеу дұрыс емес және сенімсіз болуы мүмкін. Мұқият деректерді жинау және дайындау сенімді деректер жинағын жасауға көмектеседі, нәтижесінде дәлірек жіктеу нәтижелері болады. Тиімді жіктеу үшін веб-контенттегі қылмыстық мәтіндердің әртүрлілігін көрсететін репрезентативті деректер жиынтығы болуы керек. Деректерді дұрыс жинау әртүрлі стильдерді, жанрларды, тілдерді және тақырыптарды ескере отырып, қылмыс мәтіндерінің кең ауқымын қосуға мүмкіндік береді. Осылайша, дайындалған деректер классификация моделі үшін неғұрлым толық және ақпаратты болады. Сондай-ақ деректер неғұрлым көп болса, соғұрлым жақсы. Үлкенірек деректер жинағы үлгіге көбірек мысалдарды зерттеуге және дәлірек қорытындылар жасауға мүмкіндік береді. Дегенмен, оқыту әдістері дұрыс қолданылса, тіпті шағын деректер жиынтығы пайдалы бола алады[1].

Деректер жиынын тәуелсіз деректерде үлгі өнімділігін бағалау үшін оқыту және сынақ жиындарына бөлу керек. Модельді растау және бағалауға бөлек деректер жиынтығының болуы артық сәйкестендірмеуге және оның жалпылау қабілетін бағалауға көмектеседі. Зерттеуге арналған деректер жинағывеб-мазмұннан жиналған мәтіндерді қамтуы керек және әрбір мәтін қылмыстықнемесе қылмыстық емес деп белгіленеді. Қылмыс мәтіндері қылмыс жаңалықтары, қылмыс туралы хабарламалар, әрекеттер жайында мақалалар және өзге де тиісті ақпаратты қамтыса, қылмыстық емес мәтіндер спорт, ғылым, саясат туралы жаңалықтарды не болмаса қылмысқа қатысы емес кез келген тақырыптарды қамтуы мүмкін.

Мәтінді жіктеуге және деректер жиынтығын жасауға арналған бағдарламалау тілі ретінде Python өте қолайлы және ол машиналық оқыту саласында кеңінен қолданылады. Python ғылыми және есептеукітапханаларының кең жиынтығын ұсынады, бұл оны таңдаулы таңдау етеді. Онда Scikit-Learn, NLTK, TensorFlow сияқты қуатты кітапханалар бар. Аталаған кітапханалар мәтінді алдын ала өңдеу, векторлау, машиналық оқыту үлгісін таңдау, нәтижелерді бағалауды қоса, мәтінді жіктеуге қажетті функциялар мен алгоритмдердің кең ауқымын қамтамасыз етеді.

Python-да қарапайым және түсінікті синтаксис бар, бұл оны жаңадан бастағандар үшін қолжетімді және мәтінді жіктеу алгоритмдерін әзірлеу және зерттеу үшін ыңғайлы етеді. Жасалынған жұмыс коды түгелімен Jupyter Notebook-пен жазылды, себебі ол деректерді визуализациялауға және жұмыс процесін құжаттауға мүмкіндік беретін интерактивті әзірлеу және кодты орындау ортасы. Қосымша Jupyter Notebook функционалдығын кеңейту үшін түрлі кеңейтімдер мен плагиндерді қосуға, орнатуға мүмкіндік береді. Ол Python-нан өзге де бағдарламалар тілін қолдайды және оны келесідей орнатамыз: pip install notebook. Жаңа жобаны құруды 1–суреттегідей (New → Python 3) қадаммен бастаймыз[2].

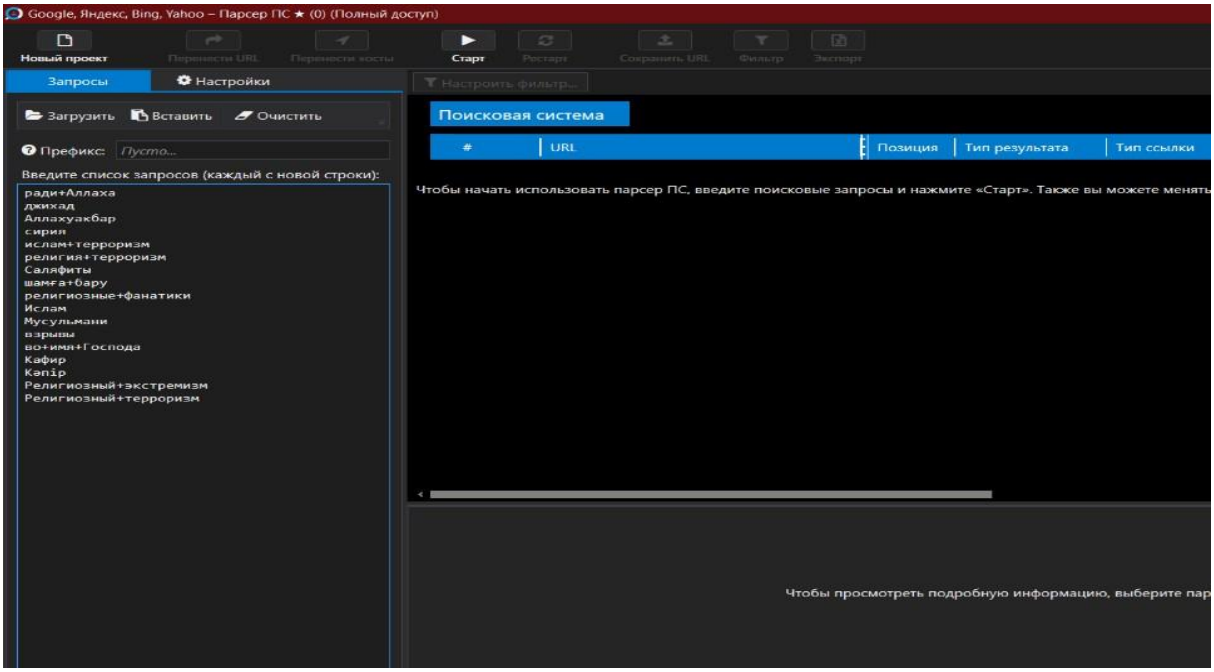


1-сурет. Jupyter Notebook-та жобаны құру

Веб-беттерден, API интерфейстерінен, дерекқорлардан және басқа көздерден деректерді автоматты түрде шығару және өңдеу үшін парсинг қолданылады. Ол құрылымдық немесе құрылымдалмаған көздерден дұрыс деректерді алуға және оны әрі қарай өңдеуге ыңғайлы форматқа түрлендіруге мүмкіндік ашады. Жалпы алғанда, парсинг - бұл әртүрлі көздерден деректерді автоматтандырылған түрде алу, өңдеу және пайдаланудың қуатты құралы. Бұл уақыт пен ресурстарды үнемдеуге, жеңілдетуге мүмкіндік береді. Бұдан бөлек түрлі мақсаттарға арналған танымал веб-скрепинг және ақпаратты алу құралдары бар. Соның мысалы, Netpeak Parser — веб-сайттардан деректерді жинайтын, содан кейін олардың мазмұнын талдайтын Netpeak Software әзірлеген бағдарламалық құрал.

Бағдарлама веб-сайттардан деректерді, соның ішінде мәтіндік мазмұнды, тақырыптарды, мета тегтерді, сілтемелер мен кескіндерді және басқа элементтерді жинауға мүмкіндік ұсынады. Қажетті параметрлерді көрсетугежәне ақпаратты алу ережелерін орнатуға болады.

Мысалы, 2-суретте көрсетілгендей керекті желі, қала, уақыт аралығын баптап алып, кілттік сөздерді жазу арқылы сол сөздерді қамтитын беттер тізімін ала аламыз.



2-сурет. Netpeak Parser бағдарламасы

Интернет ресурстрадың сілтемесін қол жеткізген соң, әрі қарай парсинг жасау үшін веб-парақшаларға HTTP сұрауларын жіберу қажет. Python-да сұраныс жіберетін кітапханалардың қарапайымы Requests деп аталады, Бұл HTTP көмегімен деректерге алу мақсатында түрлі әдістерді пайдаланады. Содан соң керекті BeautifulSoup(BS4) деген XML және HTML құжаттарын талдауға негізделген Python кітапханалардың бірін орнатып алу керек. BS4 - HTML құжатын түрлі Python нысандарының күрделі ағашына айналдыруға мүмкіндік ашады және көбіне скрапинг үшін қолданылады. Скрапинг барысында компьютер HTML құжатын алатын сұранысты жібереді. Веб- парсингтің міндеті – парақшаның керекті бөліктеріне қол жеткізу. Парсингті атап өтілген кітапханаларды қолдана отырып жасау коды 3 -суретте көрсетілген және сол негізде жүзеге асады[3].

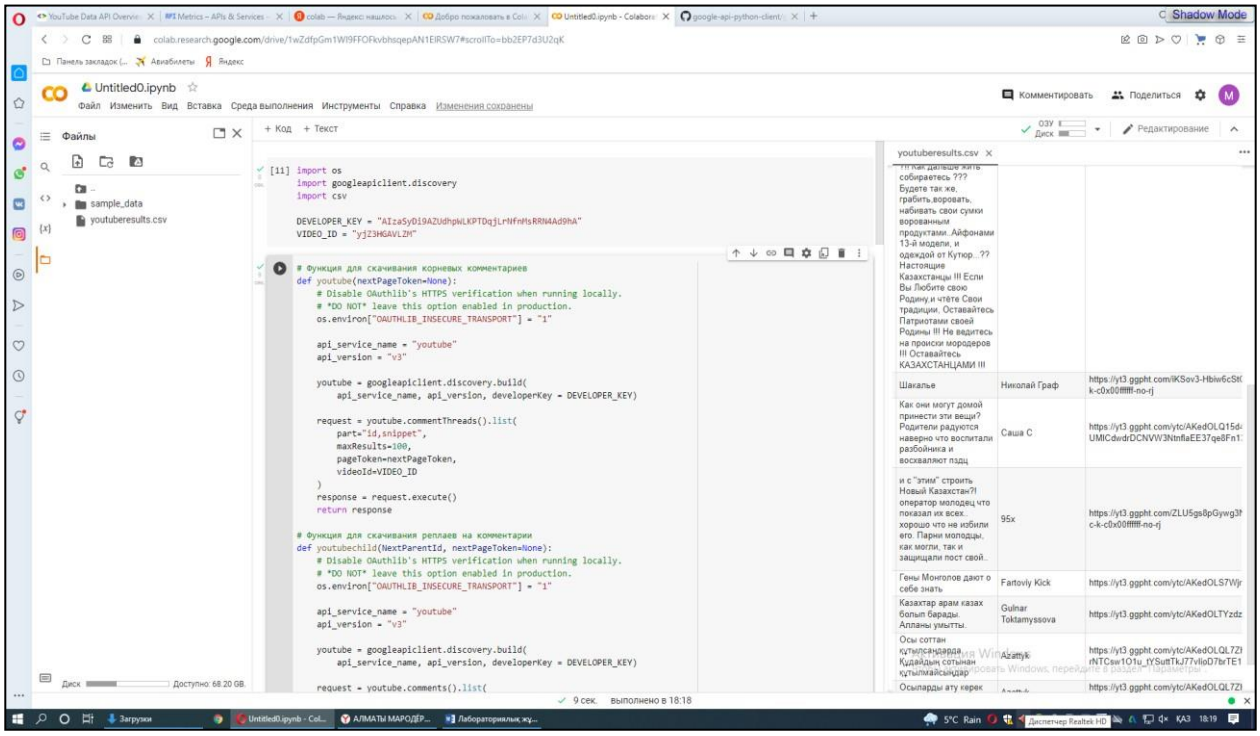
```
import requests
import re
from bs4 import BeautifulSoup

def parse_request(url):
    r = requests.get(url).text
    soup = BeautifulSoup(r)
    url_text = '. '.join([i.text for i in soup.body.find_all('p')])
    url_text = re.sub('[^A-Я,а-я,Ә,І,Ң,Ғ,Ү,Ұ,Қ,Ө,Һ,ә,і,ә,ң,ғ,ү,ұ,қ,ө,һ, ,.!?]', ' ', url_text).replace(',', '').strip()
    url_text = re.sub('\s+', ' ', url_text)
    return url_text
```

3-сурет. Парсинг жасау коды

ВКонтакте (VK) және YouTube бейне бөлісу платформасы сияқты әлеуметтік желілердегі қылмыстық мәтіндер қылмыстық әрекетке немесе қылмыс аймағына қатысты мазмұнның ерекше санаты болып табылады. Мұндай мәтіндер хабарларды, түсініктемелерді, бейне сипаттамаларын, тегтерді және ақпарат алмасу және ортақ

пайдалану үшін пайдаланушылар пайдаланатын басқа элементтерді қоса алғанда, білдірудің әртүрлі нысандарын қамтуы мүмкін. Ондағы деректерді алу және жинау процесін сәйкесінше YouTube API және VK API парсингі API (Application Programming Interface) арқылы жасаймыз[4].

YouTube API парсингі YouTube платформасындағы бейнелер, арналар, пікірлер, ойнату тізімдері және басқа элементтер туралы ақпаратты алуға мүмкіндік береді. API көмегімен бейнелерді әртүрлі параметрлер бойынша іздеуге және сүзуге, көру, ұнату не ұнатпау статистикасын алуға, сондай-ақ бейне авторы мен оның жазылушылары туралы ақпарат алуға болады. Ол бейненің танымалдылығын, трендтерді, пайдаланушылардың өзара әрекетін және YouTube-тің басқа аспектілерін талдауға мүмкіндік береді.



4-сурет. Youtube-API технологиясы

VK API парсингі VKontakte платформасындағы пайдаланушылар, қауымдастықтар, жазбалар, пікірлер, фотосуреттер және басқа элементтер туралы ақпаратты жинау мүмкіндігін береді. API пайдаланушы профильдері, олардың достары, қауымдастық қабырғалары, жазылушылар және т.б. туралы деректерді алуға мүмкіндік береді. VK API талдауын қолдана отырып, сіз пайдаланушының белсенділігін, өзара әрекеттесуін, мазмұнның танымалдылығын талдай аласыз, статистиканы жинай аласыз және VK әлеуметтік желісінде зерттеулер жүргізе аласыз.

YouTube API және VK API арқылы деректерді талдау сәйкес функциялар мен деректерге қолжетімділікті қамтамасыз ететін API кілттерін тіркеуді және алуды талап етеді. Содан кейін әртүрлі бағдарламалау тілдері мен құралдары арқылы деректерді алу және өңдеу үшін HTTP сұрауларын пайдалануға болады. API пайдалану кезінде тиісті платформалар белгілеген ережелер мен шектеулерді сақтау керек екенін ескеру маңызды[5].

Жиынтықтар .csv, .xlsx форматтарынды сақталды. Деректерді оқу, өңдеу және талдау үшін Pandas танымал ашық бастапқы Python кітапханасы пайдаланылады. Кітапхананы импорттап аламыз да, .read_csv, .read_excel арқылы жиынтығымызды оқимыз.

Веб-мазмұн жаңалықтарды, пікірлерді, блогтарды және басқа мазмұн түрлерін қоса алғанда, әртүрлі мәтін пішімдері мен құрылымдарын қамтиды. Веб-мазмұнда қателер, жарамсыз таңбалар, арнайы пішімдеу және т.б. сияқты шу деректері болуы мүмкін. Деректерді дайындау шу мен ауытқуларды жою үшін өңдеу және сүзу қадамдарын қамтиды, осылайша жіктеу сапасын жақсартады. Сол себептен мәтінді қалыпқа келтіру, мәтінді алдын ала өңдеу ретінде де белгілі, бұл деректерді талдаудағы маңызды қадам болып табылады. Ол мәтіндерді бір стандартты пішінге келтіруге бағытталған бірқатар әдістерді қамтиды, бұл ақпаратты кейінгі талдау мен алуды жеңілдетеді.

Жалпы мәтінді классификациялауда машиналық оқыту әдістерін сәтті қолдану деректерді алдын ала өңдеуді, соның ішінде токендерге бөлуді, stop сөздерді (мысалы, шылауларды, есімдіктерді) алып тастауды және нормалауды талап етеді. Мәтінді қалыпқа келтірудің мақсаты мәтіннің біркелкі және стандартталған көріністерін жасау болып табылады. Арнайы қалыпқа келтіру әдістерін таңдау нақты тапсырмаға және өңдеуді қажет ететін мәтіндік деректер түріне байланысты. Мәтінді қалыпқа келтірудің кейбір негізгі әдістері:

- Бір регистрге келтіру, мәселен кіші әріппен жазу: Бұл әдіс мәтіннің барлық әріптерін кіші әріптерге түрлендіреді. Ол әр түрлі әріп жағдайларын ажырату үшін пайдалы және сөзді сәйкестендіру процесін жеңілдетеді.
- Таңбалар мен тыныс белгілерін жою: Бұл қадам тыныс белгілері, сандар немесе арнайы таңбалар сияқты қажетсіз таңбаларды жоюды қамтиды. Бұлталдауға әсер етуі мүмкін қажет емес ақпараттың мәтінін тазартады. Осы екі міндетті орындау үшін 5–суреттегі cleaning() функциясын шақырамыз.

```
def cleaning(text):
    text = text.replace(",", "")
    text = re.sub('http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.,&+]|[!,*\(\),]|(?:%[0-9a-fA-F][0-9a-fA-F]))', '', text
    text = re.sub('[^А-Я,а-я,Ә,І,Н,ғ,Ү,Ұ,Қ,Ө,һ,ә,і,ә,н,ғ,ү,ұ,қ,ө,һ]', ' ', text)
    text = text.replace('(', ' ')
    text = text.replace(')', ' ')
    text = re.sub('_', '', text)
    text = re.sub('\s+', ' ', text)
    return text.strip().lower()
```

5-сурет. Мәтінді тазату және бір регистрге келтіру функциясы.

Келесі қадам токендерге бөлу үшін біз nltk модулін және word_tokenize сыныбын nltk.tokenize модулінен импорттап жатырмыз. Осыдан соң мәтінді енгізу ретінде қабылдайтын және таңбалауыштар тізімін қайтаратын tokenize_text функциясын жасаймыз.

NLTK кітапханасын пайдалану үшін алдымен төмендегідей орнату қажет болуы мүмкін екенін ескеріңіз. Сондай-ақ nltk.download() функциясын пайдаланып, токенизация үлгілері тәрізді қажетті ресурстарды жүктеп алу қажет.

	text	label	label_name	Tokenize_text
44	Шевелись, ублюдок! Я бы мог убить десяток чело...	1	crime	шевелись ублюдок я бы мог убить десяток челове...
45	Вы должны знать сейчас. Вы приносите деньги. У...	1	crime	вы должны знать сейчас вы приносите деньги у м...
46	Зря они жили	1	crime	зря они жили
47	С 31 на 1 заначную у бабушки и по утру заберу ...	1	crime	с на заначную у бабушки и по утру заберу бомбы
48	Мы разошлем ваши интим фото всем вашим друзьям...	1	crime	мы разошлем ваши интим фото всем вашим друзьям...
49	Следующее, что я сделаю, это отрежу уши леди и...	1	crime	следующее что я сделаю это отрежу уши леди и п...
50	Ну и сколько людей я должен ещё убить, чтобы о...	1	crime	ну и сколько людей я должен ещ убить чтобы обо...
51	Я никогда не знаю, когда это чудовище проникне...	1	crime	я никогда не знаю когда это чудовище проникнет...
52	Мне нравится убивать людей, потому что это вес...	1	crime	мне нравится убивать людей потому что это весе...

6-сурет.Токендерге бөлінген, артық символдардан тазаланған датасет.

Қосымша мәтінге шу қосатын келесі ол - тоқтау сөздер. Табиғи тілді өңдеудегі тоқтау сөздер - семантикалық жүктемені көтермейтін және мәтінді талдауға айтарлықтай үлес қоспайтын жиі кездесетін сөздер, олар әдетте жай ғана шылау, есімдік, жалғаулық және басқа да көмекші сөздер болады[6].

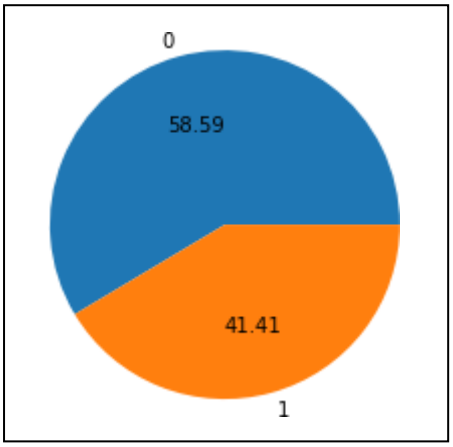
Тоқтау сөздерді қолдану мәтінді талдаудың нақты тапсырмасы мен контекстіне байланысты екенін түсіну маңызды. Кейбір жағдайларда тоқтату сөздері синтаксистік құрылымды сақтау үшін немесе талдаудың белгілі бір түрлерінде пайдалы болуы мүмкін. Дегенмен, көптеген мәселелерде талдауды жақсарту және деректер өлшемін азайту үшін тоқтату сөздерін мәтіннен шығаруға болады. Тоқтау сөздерді қолданудың артықшылықтары:

- Шуды азайту: Тоқтау сөзді жою мәтіннен жиі кездесетін, бірақ ақпаратсыз сөздерді жоюға және талдаудың сапасын жақсартуға, шу сөздердің нәтижелерге әсерін азайтуға мүмкіндік береді;
- Өңдеуді жылдамдату: тоқтату сөздерін жою мәтіндегі сөздердің санын азайтуы мүмкін, бұл есептеулерді жылдамдатады және ресурс талаптарын азайтады.

Тоқтау сөздердің анықтамасы мен қолданылуы нақты тапсырма мен мәтінді өңдеу алгоритміне байланысты. Көптеген мәтінді өңдеу кітапханалары

мен құралдары әртүрлі тілдерге арналған дайын тоқтату сөздер тізімдерін ұсынады, оларды сіз өз қажеттіліктеріңізге сай пайдалануға немесе теңеуге болады. Тоқтау сөздерді қолданудың кемшіліктері:

- Мәтінмәннің жоғалуы: Кейде тоқтау сөздері семантикалық жүктемені көтеруі мүмкін және мәтіннің контекстін талдау үшін маңызды. Оларды алып тастау ақпараттың жоғалуына немесе мағынаның бұрмалануына әкелуі мүмкін;
- Тіл ерекшеліктері: Тоқтау сөздердің тізімі әр тілге және тапсырмаға бейімделуі керек. Әртүрлі тілдер мен контексттерде әр түрлі тоқтату сөздер болуы мүмкін, сондықтан тоқтау сөздерін таңдағанда абай болу керек.



7-сурет. Деректер жинағында артық шудан тазартылған қылмыстықжәне қылмыстық емес мәтіндер үлесі

Сөз бұлты (WordCloud) - бұл мәтіндік деректердің көрнекі көрінісі, мұндаәрбір сөздің өлшемі оның берілген мәтіндегі жиілігіне немесе маңыздылығына пропорционалды. Құжаттарда, мақала немесе веб-сайт сияқты сөздер жинағындағы ең маңызды терминдерді бөлектеу үшін қолданылатын танымал визуализация әдістері бар [7]. Осылайша, веб-контенттегі қылмыстық мәтіндерді жіктеу үшін деректерді жинау мен дайындаудың маңыздылығы сенімді және өкілді деректер жиынтығын қамтамасыз ету, жіктеу дәлдігін жақсарту, шуды жою және зерттеудің қайталану мүмкіндігін қамтамасыз ету болып табылады.

Қорытындылай келіп, морфологиялық талдау әдістерін өзгерту веб-контенттегі қылмыстық мәтіндердің жіктелуін жақсартуға мүмкіндік береді. Токенизацияны, TF-IDF векторизациясын, тоқтау сөздерін жоюды, мәтіндерді лемматизациялауды және емлені түзетуді қолдану, сонымен қатар кездейсоқ орман алгоритмін пайдалану қылмыстық мәтіндерді өңдеуде жақсы нәтижелерді көрсетеді.

Әзірленген әдістер мен құралдар қылмыстық мәтіндерді анықтау және талдау қажет болатын әртүрлі салаларда, мысалы, ақпараттық қауіпсіздік, құқық қорғау және әлеуметтік желілерде қолданылуы мүмкін.

Мәтіндік деректерді өңдеу және веб-контенттегі қылмыстық мәтіндерді жіктеу саласындағы әрі қарай дамыту және зерттеу желілік ортада пайдаланушылардың қауіпсіздігі мен қорғалуын қамтамасыз ету үшін өте маңызды.

Жалпы алғанда, жұмыс нәтижелері веб-контенттегі қылмыстық мәтіндерді өңдеу және жіктеу саласындағы зерттеулердің өзектілігі мен маңыздылығын растайды, сонымен қатар осы мәселені шешу үшін морфологиялық талдаудың модификацияланған әдістерін қолданудың әлеуетін көрсетеді.

Әдебиеттер тізімі

1. Гаужаева В. А., Прокофьева Е. В., Прокофьева О. Ю. Преступность в сети Интернет: криминологические характеристики // Вестник экономической безопасности. — 2019. — № 4. — С. 111–114.
2. Болушевская И. Н. Корпусное исследование лингвистических особенностей дискурса криминальных интернет-новостей на примере новостей о похищении (на материале английского языка) // Филологические науки. Вопросы теории и практики. — 2019. — Т. 12, вып. 10. — С. 184–188.
3. Барахнин В. Б., Федотов А. М., Бакиева А. М., Бакиев М. Н., Тажибаева С. Ж. Алгоритмы генерации и стемматизации словоформ казахского языка // Вестник Казахского национального технического университета. — 2017. — № 3. — С. 123–130.

4. Фомин В. В., Флегонтов А. В., Осочкин А. А. Метод частотно-морфологической классификации текстов // Вестник Томского государственного университета. — 2017. — № 420. — С. 56–62.
5. Литвинова Т. А. Возможности компьютерной лингвистики для решения задач диагностирования личности по тексту // Вестник Воронежского государственного университета. Серия: Филология. Журналистика. — 2015. — № 3. — С. 37–41.
6. Сапин А. С. Построение нейросетевых моделей морфологического и морфемного анализа текста // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2021». — 2021. — С. 523–534.
7. Носков Д. В. Классификация текстов при помощи алгоритмов машинного обучения // Вестник Новосибирского государственного университета. Серия: Информационные технологии. — 2017. — Т. 15, № 2. — С. 45–53.

References

1. Gauzhaeva V. A., Prokof'eva E. V., Prokof'eva O. Ju. Prestupnost' v seti Internet: kriminologicheskie harakteristiki // Vestnik jekonomicheskoj bezopasnosti. — 2019. — № 4. — S. 111–114.
2. Bolushevskaja I. N. Korpusnoe issledovanie lingvisticheskih osobennostej diskursa kriminal'nyh internet-novostej na primere novostej o pohishhenii (na materiale anglijskogo jazyka) // Filologicheskie nauki. Voprosy teorii i praktiki. — 2019. — T. 12, vyp. 10. — S. 184–188.
3. Barahnin V. B., Fedotov A. M., Bakieva A. M., Bakiev M. N., Tazhibaeva S. Zh. Algoritmy generacii i stemmatizacii slovoform kazahskogo jazyka // Vestnik Kazahskogo nacional'nogo tehnicheskogo universiteta. — 2017. — № 3. — S. 123–130.
4. Fomin V. V., Flegontov A. V., Osochkin A. A. Metod chastotno-morfologicheskoj klassifikacii tekstov // Vestnik Tomskogo gosudarstvennogo universiteta. — 2017. — № 420. — S. 56–62.
5. Litvinova T. A. Vozmozhnosti komp'juternoj lingvistiki dlja reshenija zadach diagnostirovanija lichnosti po tekstu // Vestnik Voronezhskogo gosudarstvennogo universiteta. Serija: Filologija. Zhurnalistika. — 2015. — № 3. — S. 37–41.
6. Sapin A. S. Postroenie nejrosetevyh modelej morfologicheskogo i morfemnogo analiza teksta // Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferencii «Dialog 2021». — 2021. — S. 523–534.
7. Noskov D. V. Klassifikacija tekstov pri pomoshhi algoritmov mashinnogo obuchenija // Vestnik Novosibirskogo gosudarstvennogo universiteta. Serija: Informacionnye tehnologii. — 2017. — T. 15, № 2. — S. 45–53.

Д.А. Абдраманов, Ж.Д. Изтаев*, С.Ж. Куракбаева, И.К. Байназарова
магистрант, ЮКУ им. М. Ауэзова, Шымкент, Казахстан
к. п. н., доцент, ЮКУ им. М. Ауэзова, Шымкент, Казахстан
к. и. н., профессор, ЮКУ им.М. Ауэзова, Шымкент, Казахстан
магистр, старший преподаватель, ЮКУ им. М. Ауэзова, Шымкент, Казахстан
***Автор для корреспонденции:** Zhalgasbek71@mail.ru

СБОР И ПОДГОТОВКА ДАННЫХ КРИМИНАЛЬНОГО КОНТЕНТА ИЗ ВЕБ-ИСТОЧНИКОВ

Аннотация

Преступные тексты, такие как планирование преступлений, призывы к их совершению, распространение ложной информации, представляют угрозу безопасности в сетевом пространстве. Выявление и классификация таких криминальных текстов становится неотъемлемой частью борьбы с преступностью в интернете. В связи с увеличением объема доступной в сети информации и ростом противоправных действий, связанных с интернетом, возникает необходимость разработки

эффективных методов и подходов для автоматического выявления и классификации преступных текстов.

Одним из методов, применяемых при решении задач классификации преступных текстов, является использование морфологического анализа. Морфологический анализ позволяет исследовать структуру слов, их грамматические формы, а также лексические и синтаксические особенности. Однако преступные тексты обладают определенной спецификой, поэтому существующие методы морфологического анализа не всегда эффективны при их классификации. В связи с этим возникает задача модификации и совершенствования существующих методов с целью повышения точности и достижения максимально достоверных результатов.

Ключевые слова: Веб-контент, Scikit-Learn, NLTK, TensorFlow, Python, Jupyter Notebook,BeautifulSoup(BS4), XML, HTML, машинное обучение.

D.A. Abdramanov, J.D. Iztaev*, S.J. Kurakbayeva, I.K. Baynazarova
master's student, M. Auezov University, Shymkent, Kazakhstan
candidate of Pedagogical Sciences, Associate Professor, M. Auezov University, Shymkent, Kazakhstan
candidate of technical sciences, professor, M. Auezov University, Shymkent, Kazakhstan
master, senior lecturer, M. Auezov training, Shymkent, Kazakhstan
*Corresponding author's email: Zhalgasbek71@mail.ru

COLLECTION AND PREPARATION OF CRIMINAL CONTENT DATA FROM WEB SOURCES

Abstract

Criminal texts, such as planning crimes, inciting unlawful acts, and sharing false information, pose a threat to security in the online environment. Detecting and classifying such criminal texts is becoming an integral part of combating cybercrime. With the increasing volume of publicly available information and the rise in illegal activities on the Internet, it is necessary to develop effective methods and approaches for the automatic detection and classification of criminal texts.

One of the approaches used in the classification of criminal texts is the application of morphological analysis methods. Morphological analysis allows for the examination of word structures, their grammatical forms, and lexical and syntactic features. However, criminal texts have distinct characteristics, which means that existing morphological analysis methods are not always effective for their classification. Therefore, the task of modifying and improving existing methods arises in order to enhance accuracy and achieve more reliable results.

Keywords: web content, Scikit-Learn, NLTK, TensorFlow, Python, Jupyter Notebook,BeautifulSoup(BS4), XML, HTML, machine learning.