

## Мәтіндегі экстремистік бағытты анықтау үшін веб-ресурстардағы семантикалық талдау модельдерін, алгоритмдерін құрастыру және кибер-криминалистика құрал-жабдықтарын әзірлеу

КазНУ им. аль-Фараби ФИТ Главная Парсер О проекте

Статистика  
Анализ  
Источники  
Ключевые слова  
Визуализация связей  
Профили

Отчеты +  
Текущий месяц  
Квартал  
За год

### Сколько казахстанцев находится на территории Сирии и Ирака?

После объявления о "халифате" на территории Сирии и Ирака количество желающих уехать в эти государства увеличилось. Перед отъездом в Сирию они активно пропагандировали ваххабитскую идеологию и уговаривали своих жен уехать с ними.

### Басым бағыт: Ұлттық қауіпсіздік және қорғаныс

**Жобаның мақсаты:** веб-ресурстардағы экстремистік мазмұнды анықтау үшін мәліметтерді семантикалық талдау үлгілерін, алгоритмдерін және бағдарламалық жабдықтамасын, қатысушы қолданушыларды анықтау әдістерін және байланыстарды графикалық визуалдау алгоритмдерін кешенді зерттеу және құру, қаржыландыру көздерін анықтау үшін Даркоин төлем жүйелеріндегі және олардың бейімдеулеріндегі криптовалюта транзакциясын талдау үлгілерін құру және зерттеу, экстремизмге қарсы күресу үшін кибер-криминалистика құрал-жабдығын әзірлеу болып табылады.

**Өзектілігі мен жаңалығы:** ЭБ (экстремистік бағыт) анықтау үшін веб-мазмұнды жинауға және талдауға арналған бағдарламалық модуль, ЭБ мәтіндерін анықтауға арналған машиналық әдістерді оқыту және тестілеуге арналған қазақ тіліндегі мәтіндер корпусы, ЭБ мәтіндерін семантикалық талдау моделі, морфологиялық анализатор, кілт сөздер базасы, ЭБ мәтіндерін анықтауға арналған машиналық әдістерді оқыту мен тестілеуге арналған белгілер жиынтығы.

**Практикалық маңыздылығы:** жоба нәтижелерінің маңыздылығы экстремистік мазмұндағы мәтіндерді анықтауды машинамен оқыту әдістері үшін белгілердің оңтайлы жиынтығын анықтау үшін қазақ және орыс тілдерінде веб ресурстар деректерінің семантикалық талдау модельдерін, алгоритмдерін, қатысушы пайдаланушыларды сәйкестендіру әдістерін, есептеу құрылғылары үшін payload құру және енгізу әдістерін және құрылғыларға қашықтан қол жеткізуді, инновациялық құрылымдық үлгілермен ерекшеленетін Darkcoin сияқты жүйелер үшін кибер-тергеу алгоритмдерін және транзакция графтары мен топологиясын талдау әдістерін, сондай-ақ көпшілік ақпараттың

үлкен көлемінен көлеңкелі нысандарды танитын жасанды нейрондық желілерді қолдана отырып, транзакция деректерін іздеу әдісін әзірлеу.

**Енгізу нысаны:** ҚР ұлттық қауіпсіздігін қамтамасыз етуші уәкілетті органдар үшін.

**Енгізу артықшылықтары:** әлеуметтік желілердегі ақпараттық қауіпсіздікті қамтамасыз ету және экстремистік бағыттағы веб ресурстарды автоматты түрде анықтау және сәйкестендіру құралын ұсыну.

**Тұтынушылар:** іргелі нәтижелерді әлемдік ғылыми қауымдастық пайдалана алады, әдіснама, алгоритмдер, патенттер және прототип түріндегі қолданбалы нәтижелерді ақпараттық қауіпсіздікті, сыни инфрақұрылымды қамтамасыз ету, интернет-экстремизммен күрес жөніндегі уәкілетті органдар пайдалануы мүмкін.

**Жобаның бәсекеге қабілеттілігі (технологиялардың артықшылықтары) және коммерциялануы:** алынған нәтижелерді әлеуметтік желідегі қауіпсіздікті қамтамасыз ету үшін қолдануға болады.

**Күтілетін нәтижелер:**

- Экстремистік бағытты (ЭБ) анықтауға арналған веб-мазмұнды жинауға және талдауға арналған бағдарламалық модуль;
- ЭБ мәтіндерін анықтауға арналған машиналық әдістерді оқыту және тестілеуге арналған қазақ тіліндегі мәтіндер корпусы;
- ЭБ мәтіндерін семантикалық талдау моделі;
- морфологиялық анализатор;
- кілт сөздер базасы, ЭБ мәтіндерін анықтауға арналған машиналық әдістерді оқыту мен тестілеуге арналған белгілер жиынтығы;
- ЭБ анықтау мақсатында аудио және видео хабарламаларды талдау әдістері;
- Қазақ тіліндегі мәтіндердегі ЭБ анықтау үшін машиналық оқыту әдістерін бейімдеу.

**Байланыс мәліметтері:** [mussiraliyevash@gmail.com](mailto:mussiraliyevash@gmail.com)

Статистика

WordCloud

Грамммы

TF-IDF

LWC

Анализ

Источники

Ключевые слова

Визуализация связей

Профили

Отчеты

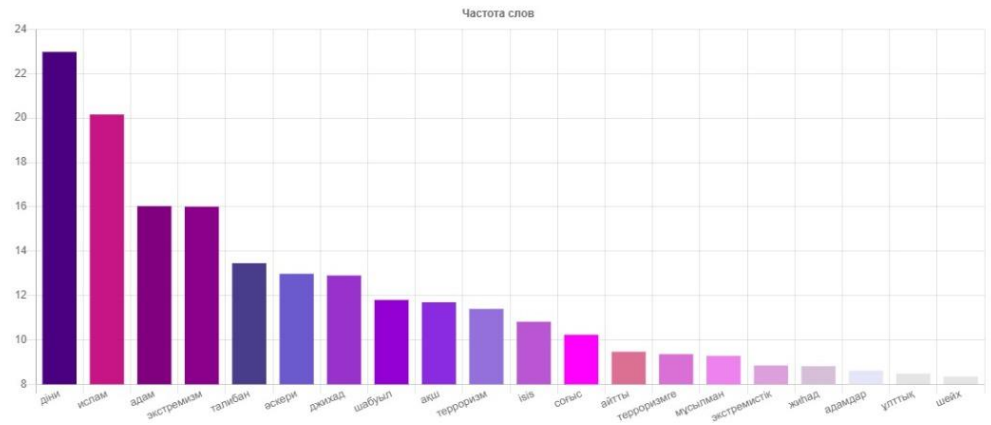
Текущий месяц

Квартал

127.0.0.1:8000/tf-idf

TF (частота слов) характеризует отношение числа вхождений конкретного слова к общему набору слов в документе. Чем выше TF, тем весомее конкретное слово в рамках документа.

IDF (обратная частота документа) характеризует инверсию частотности, с которой конкретное слово используется в тексте. С помощью этой метрики можно снизить важность слов — например, союзов или предлогов.



Статистика

WordCloud

Грамммы

TF-IDF

LWC

Анализ

Источники

Ключевые слова

Визуализация связей

Профили

Отчеты

Текущий месяц

Квартал

127.0.0.1:8000/ngram

### n-Грамммы

**N-грамма** — это просто последовательность из n элементов (звуков, слогов, слов или символов), идущих в каком-то тексте подряд. На практике чаще имеют в виду ряд слов (реже — символов). Последовательность из двух элементов называют **биграмма**, из трёх элементов — **триграмма**.

Вычислив частоту вхождения N-грамм в текстах корпуса, мы можем добавить би-граммы или три-граммы в качестве функций для представления наших документов в задачах классификации текста.

#	Униграмма	Биграмма	Триграмма
1	қарсы	терроризмге қарсы	экстремизм терроризмге қарсы
2	керек	болуы мүмкін	америка құрама штаттары
3	ислам	діни экстремизм	ауғанстан ислам әмірлігі
4	әскери	ауғанстан ислам	ауғанстан ислам әмірлігінің
5	болды	америка құрама	діни экстремизм терроризмге
6	егер	бұқаралық ақпарат	бұқаралық ақпарат құралдары
7	діни	болған кезде	бұқаралық ақпарат құралдарында
8	болады	ислам мемлекеті	америка құрама штаттарының
9	басқа	қарсы тұру	терроризмге қарсы іскимыл
10	болған	соның ішінде	ұлттық қауіпсіздік комитеті
11	талибан	ішкі істер	адам қаза тапты

Статистика

WordCloud

Грамммы

TF-IDF

LWC

Анализ

Источники

Ключевые слова

Визуализация связей

Профили

Отчеты

Текущий месяц

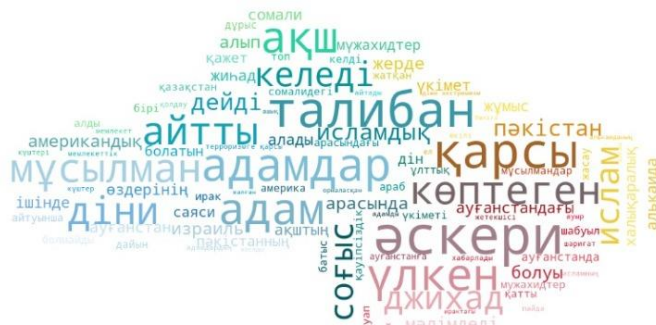
Квартал

127.0.0.1:8000/word-cloud

### Облако слов обучающего корпуса

Word Cloud - это ресурс, позволяющий создать визуальный образ ключевых слов, текста в привлекательной форме. Наиболее часто используемые слова отображаются крупным шрифтом. Важность каждого ключевого слова обозначается размером шрифта или цветом.

Такое представление удобно для быстрого восприятия.



- Статистика
- Вывод
- Источники
- Ключевые слова
- Визуализация связей
- Профили

- Отчеты
- Текущий месяц
  - Квартал
  - За год

### Обнаруживать экстремистский контент

Экспериментальный сайт 101000

Группы:  Месяц:

- Выберите
- Экспериментальный сайт 101000
- Азаттық радиосы
- Ислам тарихы
- Қазақстандағы Мұсылман Ғарышкерлері
- Пайғамбар аяқталды
- Мәңгілік ел жұлдыз



Экстремистский контент процентов: 0.00%

101000/101000

- Статистика
- Вывод
- Источники
- Ключевые слова
- Визуализация связей
- Профили

- Отчеты
- Текущий месяц
  - Квартал
  - За год

### Обнаруживать экстремистский контент

Экспериментальный сайт 101000

Группы:  Месяц:



Экстремистский контент процентов: 6.39%