Веб-сервіс автоматизованої побудови регулярних виразів

Виконав:

Студент Данилюк Василь

Науковий керівник:

Сергеєв-Горчинський О.О

Об’єкт дослідження:

- Неструктуровані текстові дані

Предмет дослідження:

- Регулярні вирази
- Автоматизована побудова регулярних виразів.

Мета

- Використання й побудова регулярних виразів, дослідження технологій з якими вони використовуються
- Створення веб додатку для побудови регулярних виразів

Актуальність

- Побудова регулярних виразів є актуальним завданням, кожен програміст стикався з проблемою створення регулярного виразу
- Існують сервіси для їх перевірки, але не для автоматизованої побудови

Регулярний вираз

- це запис, що описує множину рядків, відповідно до набору спеціальних синтаксичних правил
- /week([\d]+)/
- /[0-9]{13}/
- /[A-Z][a-z]+[\s][0-9][^a-zA-Z0-9]

Використані програмні зас

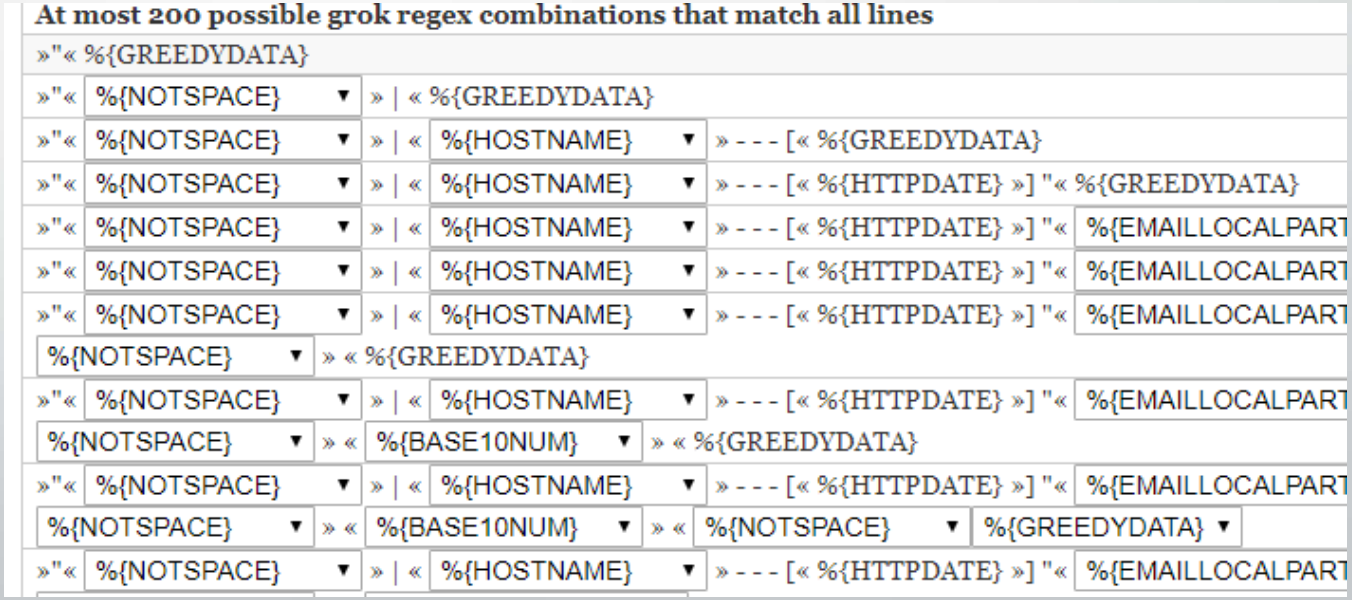
Grok Constructor

AJAX

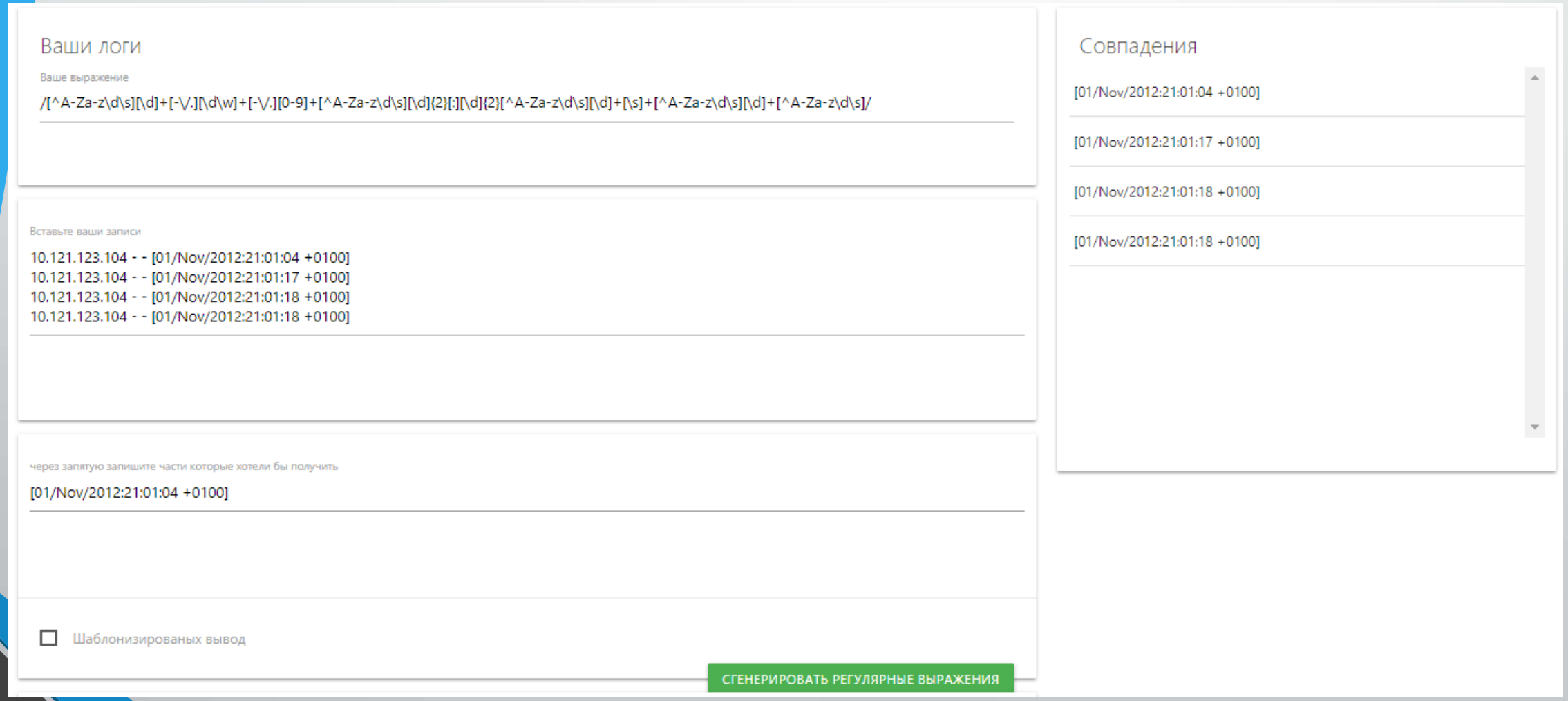
REGEX Standard

Grok(плагін ElasticSearch)

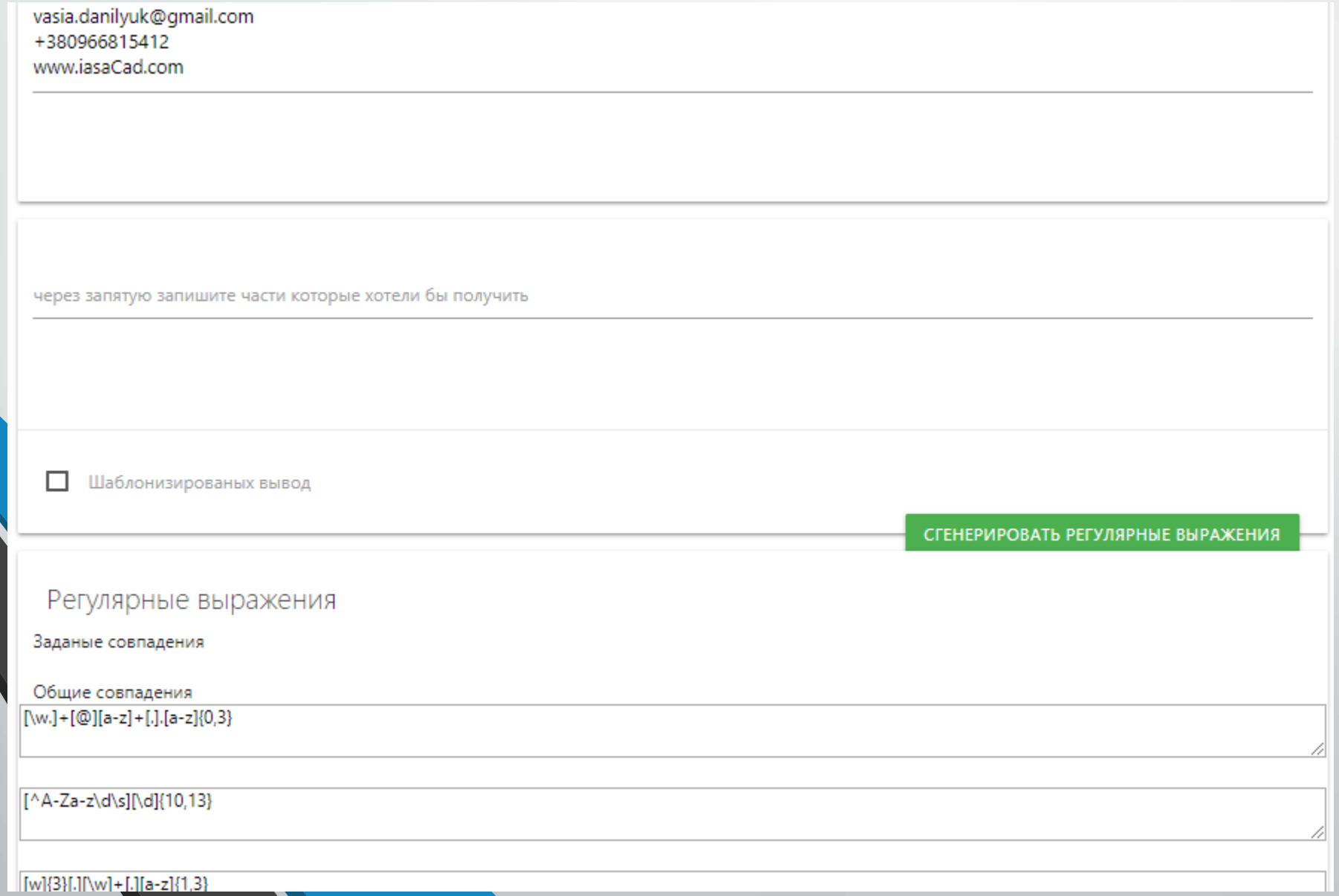
- **GrokConstructor** є сервісом для тестування та інкрементної побудови регулярних виразів для фільтра grok, який аналізує logfile для Logstash.



Створення додатку для побудови регулярного виразу

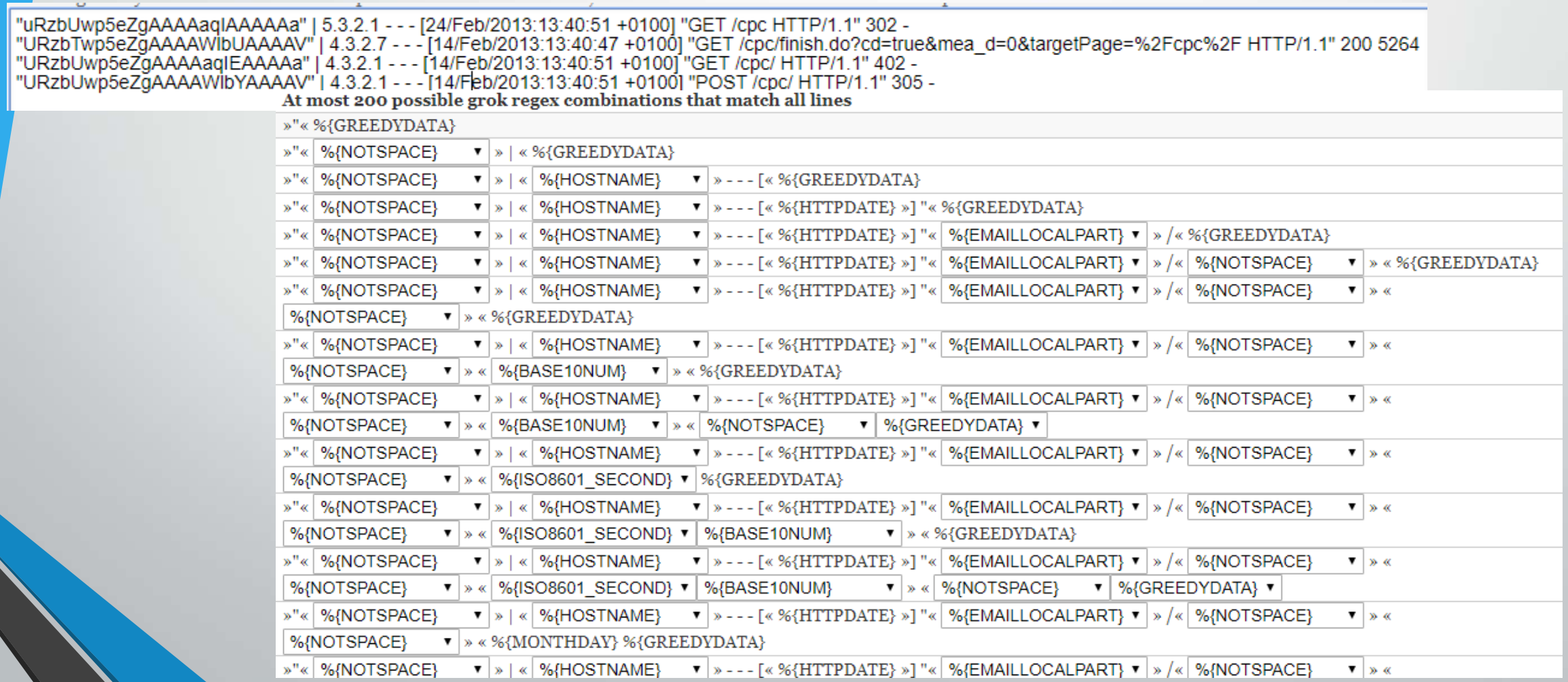


Регулярні вирази показані без використання шаблонів



GrokConstructor

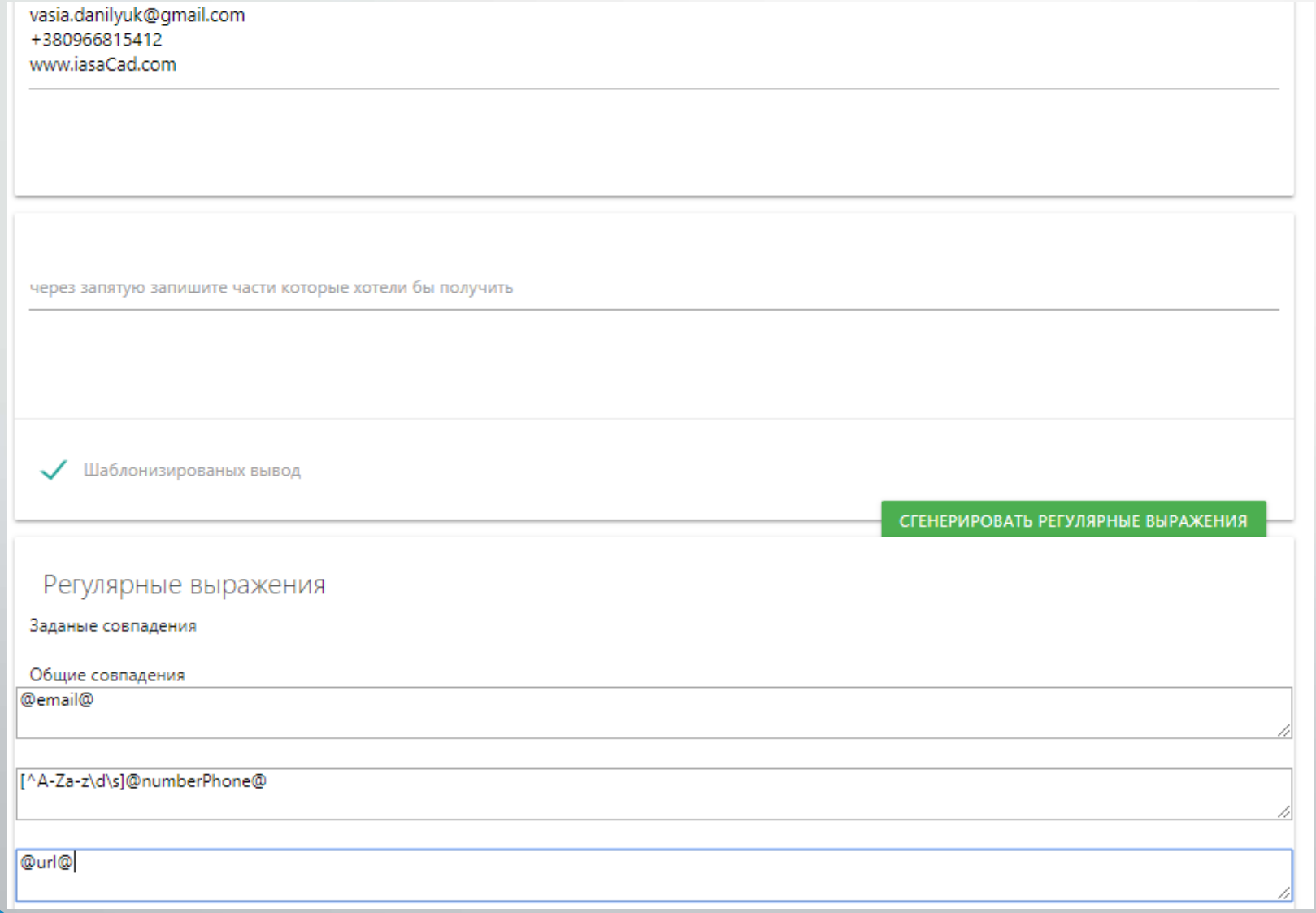
- Grok підтримує шаблони для розпізнавання



Базові шаблони, побудовані у додатку

```
"@utf8date@"=>"[A-Z][a-z]{2}[\s][A-Z][a-z]{2}[\s][\d]{2}[\s][\d]{2}[:][\d]{2}[:][\d]{2}[\s][\d]{4}",
"@numberPhone@"=>"[\d]{10,13}",
"@ip@" => "[\d]+[.][\d]+[.][\d]+[.][\d:]+",
"@email@" => "[\w.]+[@][a-z]+[.].[a-z]{0,3}",
"@date@" => "[\d]+[-\/.][\d\w]+[-\/.][0-9]+",
"@time@" => "[\d]{2}[:][\d]{2}",
"@mac@" => "[0-9a-zA-Z]{2}[-:][0-9a-zA-Z]{2}[-:][0-9a-zA-Z]{2}[-:][0-9a-zA-Z]{2}[-:][0-9a-zA-Z]{2}",
"@space@"=>"[\s]+",
"@url@"=>"[w]{3}[.][\w]+[.][a-z]{1,3}"
```

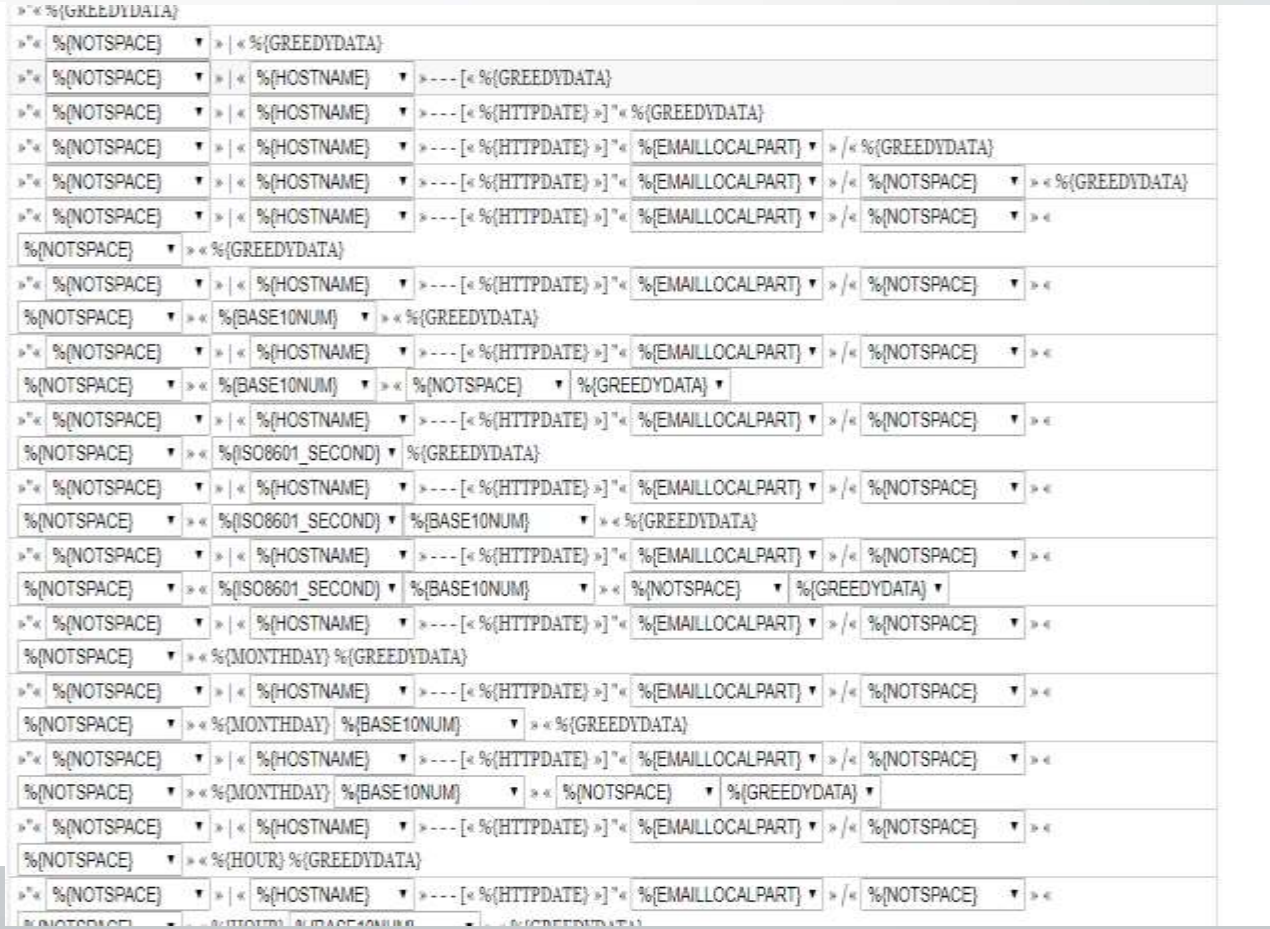
Результат побудови регулярних виразів з використанням шаблонів



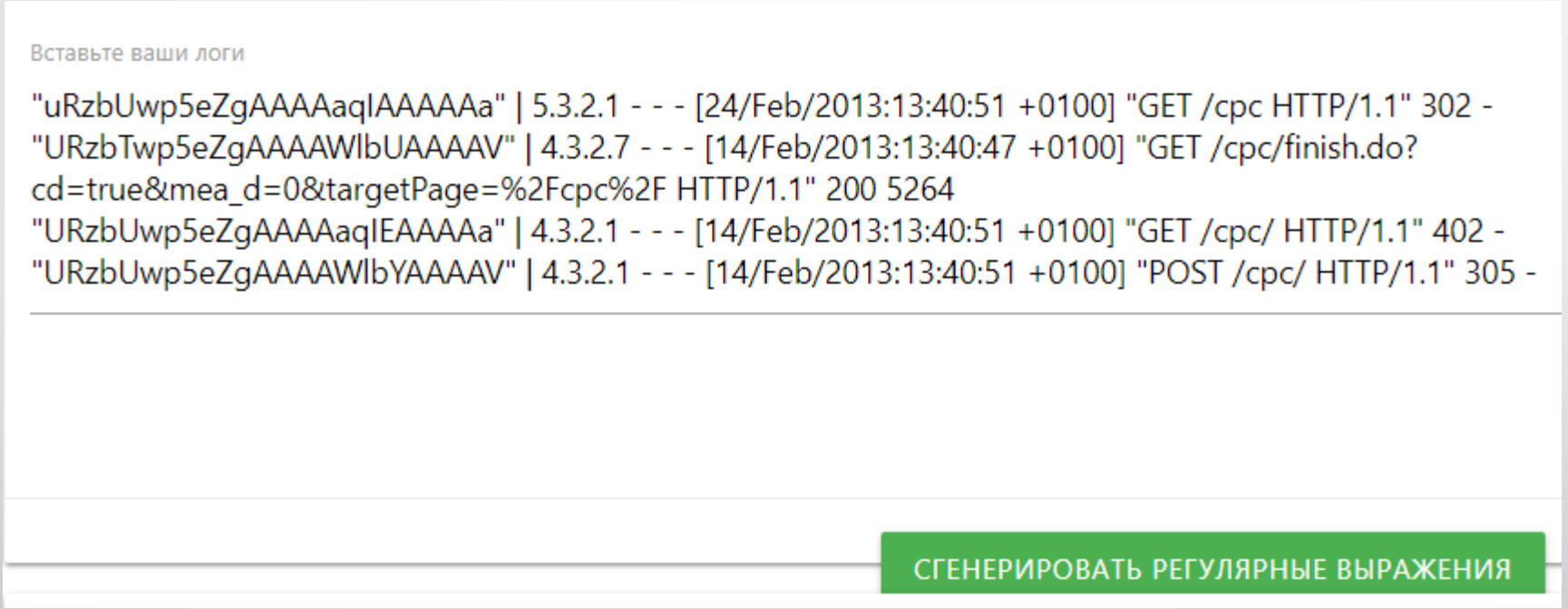
Порівняння GrokConstructor з розробленим додатком

- За структурою побудованих шаблонів результати схожі

```
[^A-Za-z\d][A-Za-z]+[\d][A-Za-z]+[^A-Za-z\d]@space@.*

[^A-Za-z\d][A-Za-z]+[\d][A-Za-z]+[^A-Za-z\d]@space@[^A-Za-z\d]@space@.*

[^A-Za-z\d][A-Za-z]+[\d][A-Za-z]+[^A-Za-z\d]@space@[^A-Za-z\d]@space@@ip@.*

[^A-Za-z\d][A-Za-z]+[\d][A-Za-z]+[^A-Za-z\d]@space@[^A-Za-z\d]@space@@ip@@space@.*

[^A-Za-z\d][A-Za-z]+[\d][A-Za-z]+[^A-Za-z\d]@space@[^A-Za-z\d]@space@@ip@@space@[^A-Za-z\d]@space@.*

[^A-Za-z\d][A-Za-z]+[\d][A-Za-z]+[^A-Za-z\d]@space@[^A-Za-z\d]@space@@ip@@space@[^A-Za-z\d]@space@[^A-Za-z\d]@space@.*

[^A-Za-z\d][A-Za-z]+[\d][A-Za-z]+[^A-Za-z\d]@space@[^A-Za-z\d]@space@@ip@@space@[^A-Za-z\d]@space@[^A-Za-z\d]@space@[^A-Za-z\d]@space@[^A-Za-z\d]@date@.*
```



Перевірка на тестових даних



Перевірка на тестових даних

Регулярные выражения

[^A-Za-z\d][A-Za-z]+[\d][A-Za-z]+[^A-Za-z\d]@space@

[^A-Za-z\d][A-Za-z]+[\d][A-Za-z]+[^A-Za-z\d]@space@[^A-Za-z\d]@space@

[^A-Za-z\d][A-Za-z]+[\d][A-Za-z]+[^A-Za-z\d]@space@[^A-Za-z\d]@space@@ip@

[^A-Za-z\d][A-Za-z]+[\d][A-Za-z]+[^A-Za-z\d]@space@[^A-Za-z\d]@space@@ip@@space@

[^A-Za-z\d][A-Za-z]+[\d][A-Za-z]+[^A-Za-z\d]@space@[^A-Za-z\d]@

[^A-Za-z\d][A-Za-z]+[\d][A-Za-z]+[^A-Za-z\d]@space@[^A-Za-z\d]@z\d]@space@[^A-Za-z\d]@space@

[^A-Za-z\d][A-Za-z]+[\d][A-Za-z]+[^A-Za-z\d]@space@[^A-Za-z\d]@z\d]@space@[^A-Za-z\d]@space@[^A-Za-z\d]@space@

Ваши логи

Ваше выражение

/[^A-Za-z\d][A-Za-z]+[\d][A-Za-z]+[^A-Za-z\d]@space@/

Вставьте ваши логи

"uRzbUwp5eZgAAAAaqIAAAAAa" | 5.3.2.1 - - - [24/Feb/2013:13:40:51 +0100] "GET /cpc HTTP/1.1" 302 -
"URzbTwp5eZgAAAAWlbUAAAAV" | 4.3.2.7 - - - [14/Feb/2013:13:40:47 +0100] "GET /cpc/finish.do?cd=true&mea_d=0&targetPage=%2Fcpc%2F HTTP/1.1" 200 5264
"URzbUwp5eZgAAAAaqIEAAAAa" | 4.3.2.1 - - - [14/Feb/2013:13:40:51 +0100] "GET /cpc/ HTTP/1.1" 402 -
"URzbUwp5eZgAAAAWlbYAAAAV" | 4.3.2.1 - - - [14/Feb/2013:13:40:51 +0100] "POST /cpc/ HTTP/1.1" 305 -

СГЕНЕРИРОВАТЬ РЕГУЛЯРНЫЕ ВЫРАЖЕНИЯ

Совпадения

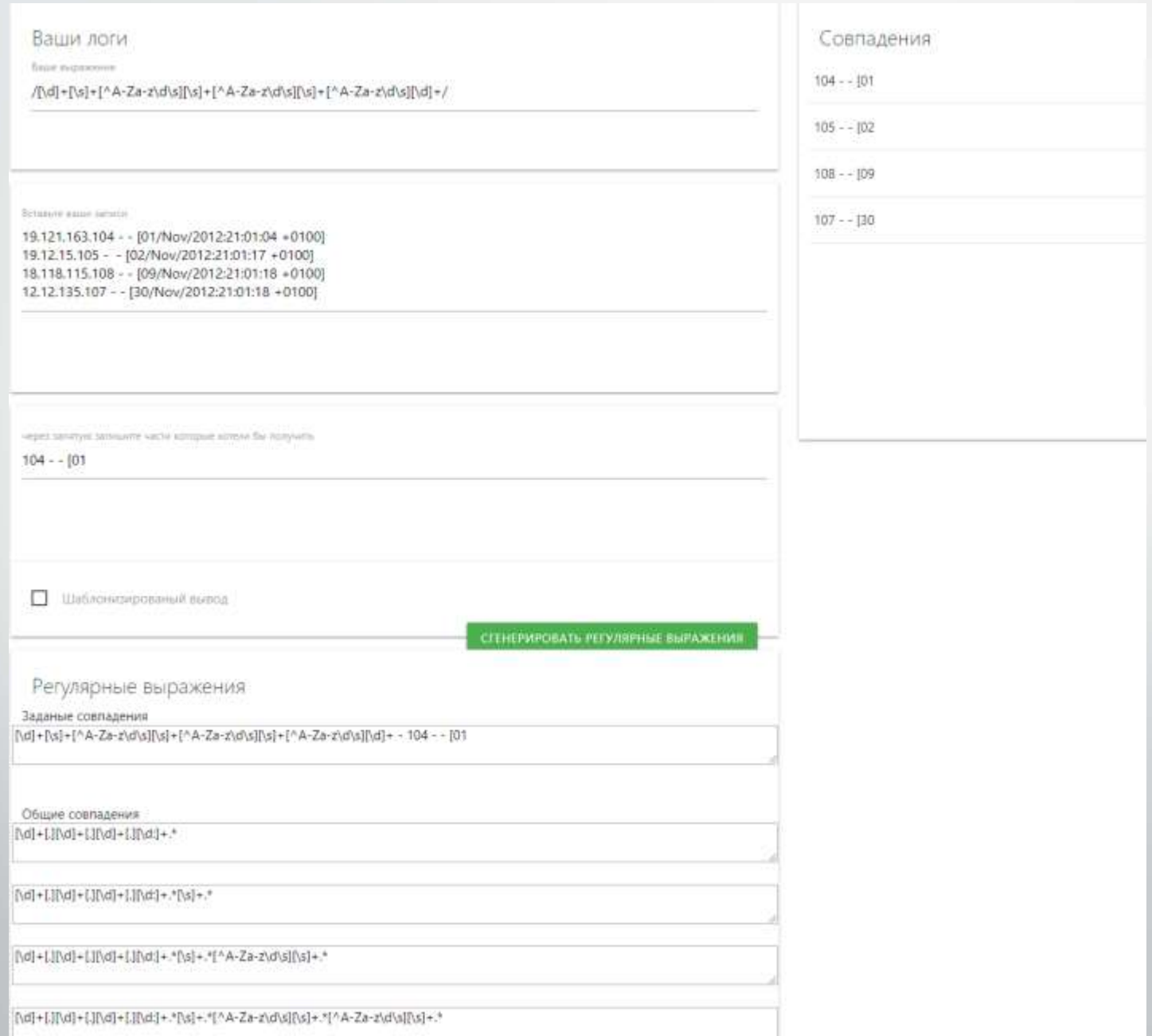
"uRzbUwp5eZgAAAAaqIAAAAAa"

"URzbTwp5eZgAAAAWlbUAAAAV"

"URzbUwp5eZgAAAAaqIEAAAAa"

"URzbUwp5eZgAAAAWlbYAAAAV"

Побудовані вирази та їх перевірка



Висновки

- Було досліджено специфіку роботи регулярних виразів
- Досліджено плагін ES GrokConstructor
- Створено Веб-додаток для автоматизованої побудови регулярних виразів

Майбутні напрями для досліджень

- В майбтньому було б гарно дослідити побудову регулярних виразів за допомогою ML.
- Створити бібліотеку шаблонів
- Розробка API для Веб-сервісу для інтеграція з cloud-сховищами даних.

Дякую за увагу