



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

AX



REGEX Standard

Grok(плагін ElasticSearch)

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

At most 200 possible grok regex combinations that match all lines	
»"«	%{GREEDYDATA}
»"«	%{NOTSPACE} » « %{GREEDYDATA}
»"«	%{NOTSPACE} » » « %{HOSTNAME} » --- [« %{GREEDYDATA}
»"«	%{NOTSPACE} » » « %{HOSTNAME} » » --- [« %{HTTPDATE} »] "« %{GREEDYDATA}
»"«	%{NOTSPACE} » » « %{HOSTNAME} » » --- [« %{HTTPDATE} »] "« %{EMAILLOCALPART
»"«	%{NOTSPACE} » » « %{HOSTNAME} » » --- [« %{HTTPDATE} »] "« %{EMAILLOCALPART
»"«	%{NOTSPACE} » » « %{HOSTNAME} » » --- [« %{HTTPDATE} »] "« %{EMAILLOCALPART
»"«	%{NOTSPACE} » » » « %{GREEDYDATA}
»"«	%{NOTSPACE} » » « %{HOSTNAME} » » --- [« %{HTTPDATE} »] "« %{EMAILLOCALPART
»"«	%{NOTSPACE} » » » « %{BASE10NUM} » » « %{GREEDYDATA}
»"«	%{NOTSPACE} » » « %{HOSTNAME} » » --- [« %{HTTPDATE} »] "« %{EMAILLOCALPART
»"«	%{NOTSPACE} » » » « %{BASE10NUM} » » « %{NOTSPACE} » » « %{GREEDYDATA}
»"«	%{NOTSPACE} » » « %{HOSTNAME} » » --- [« %{HTTPDATE} »] "« %{EMAILLOCALPART

Створення додатку для побудови регулярного виразу

Ваши логи

Ваше выражение

```
/[^\A-Za-z\d\s][\d]+[-\V.][\d\w]+[-\V.]0-9+[^A-Za-z\d\s]{2}[\d]{2}[^\A-Za-z\d\s][\d]+[\s]+[^\A-Za-z\d\s][\d]+[^\A-Za-z\d\s]/
```

Вставьте ваши записи

```
10.121.123.104 - - [01/Nov/2012:21:01:04 +0100]  
10.121.123.104 - - [01/Nov/2012:21:01:17 +0100]  
10.121.123.104 - - [01/Nov/2012:21:01:18 +0100]  
10.121.123.104 - - [01/Nov/2012:21:01:18 +0100]
```

через запятую запишите части которые хотели бы получить

```
[01/Nov/2012:21:01:04 +0100]
```

Шаблонизированных вывод

Совпадения

```
[01/Nov/2012:21:01:04 +0100]
```

```
[01/Nov/2012:21:01:17 +0100]
```

```
[01/Nov/2012:21:01:18 +0100]
```

```
[01/Nov/2012:21:01:18 +0100]
```

СГЕНЕРИРОВАТЬ РЕГУЛЯРНЫЕ ВЫРАЖЕНИЯ

Регулярні вирази показані без використання шаблонів

vasia.danilyuk@gmail.com
+380966815412
www.iasaCad.com

через запятую запишите части которые хотели бы получить

Шаблонизированных вывод

СГЕНЕРИРОВАТЬ РЕГУЛЯРНЫЕ ВЫРАЖЕНИЯ

Регулярные выражения

Заданные совпадения

Общие совпадения

```
[\\w.]+@[a-z]+[.][a-z]{0,3}
```

```
[^A-Za-z\\d\\s][\\d]{10,13}
```

```
[w]{3}.[\\w]+.[a-z]{1,3}
```


Результат побудови регулярних виразів з використанням шаблонів

vasia.danilyuk@gmail.com
+380966815412
www.iasaCad.com

через запятую запишите части которые хотели бы получить

Шаблонизированных вывод

СГЕНЕРИРОВАТЬ РЕГУЛЯРНЫЕ ВЫРАЖЕНИЯ

Регулярные выражения

Заданные совпадения

Общие совпадения

@email@

[^A-Za-z\d\s]@numberPhone@

@url@

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][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@date@.*
```

```
> * * {GREEDYDATA}
> * * {NOTSPACE} > * | < {GREEDYDATA}
> * * {NOTSPACE} > * | < {HOSTNAME} > > --- [ < {GREEDYDATA}
> * * {NOTSPACE} > * | < {HOSTNAME} > > --- [ < {HTTPDATE} > * * {GREEDYDATA}
> * * {NOTSPACE} > * | < {HOSTNAME} > > --- [ < {HTTPDATE} > * * {EMAILLOCALPART} > > / < {GREEDYDATA}
> * * {NOTSPACE} > * | < {HOSTNAME} > > --- [ < {HTTPDATE} > * * {EMAILLOCALPART} > > / < {NOTSPACE} > > < {GREEDYDATA}
> * * {NOTSPACE} > * | < {HOSTNAME} > > --- [ < {HTTPDATE} > * * {EMAILLOCALPART} > > / < {NOTSPACE} > > <
{NOTSPACE} > > < {GREEDYDATA}
> * * {NOTSPACE} > * | < {HOSTNAME} > > --- [ < {HTTPDATE} > * * {EMAILLOCALPART} > > / < {NOTSPACE} > > <
{NOTSPACE} > > < {BASE10NUM} > > < {GREEDYDATA}
> * * {NOTSPACE} > * | < {HOSTNAME} > > --- [ < {HTTPDATE} > * * {EMAILLOCALPART} > > / < {NOTSPACE} > > <
{NOTSPACE} > > < {BASE10NUM} > > < {NOTSPACE} > > {GREEDYDATA} >
> * * {NOTSPACE} > * | < {HOSTNAME} > > --- [ < {HTTPDATE} > * * {EMAILLOCALPART} > > / < {NOTSPACE} > > <
{NOTSPACE} > > < {ISO8601_SECOND} > > {GREEDYDATA}
> * * {NOTSPACE} > * | < {HOSTNAME} > > --- [ < {HTTPDATE} > * * {EMAILLOCALPART} > > / < {NOTSPACE} > > <
{NOTSPACE} > > < {ISO8601_SECOND} > > {BASE10NUM} > > < {GREEDYDATA}
> * * {NOTSPACE} > * | < {HOSTNAME} > > --- [ < {HTTPDATE} > * * {EMAILLOCALPART} > > / < {NOTSPACE} > > <
{NOTSPACE} > > < {ISO8601_SECOND} > > {BASE10NUM} > > < {NOTSPACE} > > {GREEDYDATA} >
> * * {NOTSPACE} > * | < {HOSTNAME} > > --- [ < {HTTPDATE} > * * {EMAILLOCALPART} > > / < {NOTSPACE} > > <
{NOTSPACE} > > < {MONTHDAY} > > {GREEDYDATA}
> * * {NOTSPACE} > * | < {HOSTNAME} > > --- [ < {HTTPDATE} > * * {EMAILLOCALPART} > > / < {NOTSPACE} > > <
{NOTSPACE} > > < {MONTHDAY} > > {BASE10NUM} > > < {GREEDYDATA}
> * * {NOTSPACE} > * | < {HOSTNAME} > > --- [ < {HTTPDATE} > * * {EMAILLOCALPART} > > / < {NOTSPACE} > > <
{NOTSPACE} > > < {MONTHDAY} > > {BASE10NUM} > > < {NOTSPACE} > > {GREEDYDATA} >
> * * {NOTSPACE} > * | < {HOSTNAME} > > --- [ < {HTTPDATE} > * * {EMAILLOCALPART} > > / < {NOTSPACE} > > <
{NOTSPACE} > > < {HOURL} > > {GREEDYDATA}
> * * {NOTSPACE} > * | < {HOSTNAME} > > --- [ < {HTTPDATE} > * * {EMAILLOCALPART} > > / < {NOTSPACE} > > <
{NOTSPACE} > > < {HOURL} > > {BASE10NUM} > > < {GREEDYDATA}
```

Перевірка на тестових даних

Вставьте ваши логи

"uRzbUwp5eZgAAAAaqlAAAAAa" | 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

"URzbUwp5eZgAAAAaqlEAAAAa" | 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

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@

Ваши логи

Ваше выражение

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

Вставьте ваши логи

[^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@

"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=%2Fcpcc%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 -

Совпадения

"uRzbUwp5eZgAAAAaqIAAAAAa"

"URzbTwp5eZgAAAAWlbUAAAAV"

"URzbUwp5eZgAAAAaqIEAAAAa"

"URzbUwp5eZgAAAAWlbYAAAAV"

Побудовані вирази та їх перевірка

Ваши логи

Ваше выражение

```
/[d]+[s]+[^A-Za-z\d\s][s]+[^A-Za-z\d\s][s]+[^A-Za-z\d\s][d]+/
```

Совпадения

104 -- [01]

105 -- [02]

108 -- [09]

107 -- [30]

Вставьте исходные данные

```
19.121.163.104 -- [01/Nov/2012:21:01:04 +0100]
19.12.15.105 -- [02/Nov/2012:21:01:17 +0100]
18.118.115.108 -- [09/Nov/2012:21:01:18 +0100]
12.12.135.107 -- [30/Nov/2012:21:01:18 +0100]
```

через запятую закройте части которые хотели бы получить

```
104 -- [01
```

Шаблонизированный вывод

[СГЕНЕРИРОВАТЬ РЕГУЛЯРНЫЕ ВЫРАЖЕНИЯ](#)

Регулярные выражения

Заданные совпадения

```
[d]+[s]+[^A-Za-z\d\s][s]+[^A-Za-z\d\s][s]+[^A-Za-z\d\s][d]+ - 104 -- [01
```

Общие совпадения

```
[d]+[.][d]+[.][d]+[.][d]+.*
```

```
[d]+[.][d]+[.][d]+[.][d]+.*[s]+.*
```

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

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


Висновки

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

Майбутні напрями для досліджень

- В майбутньому 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