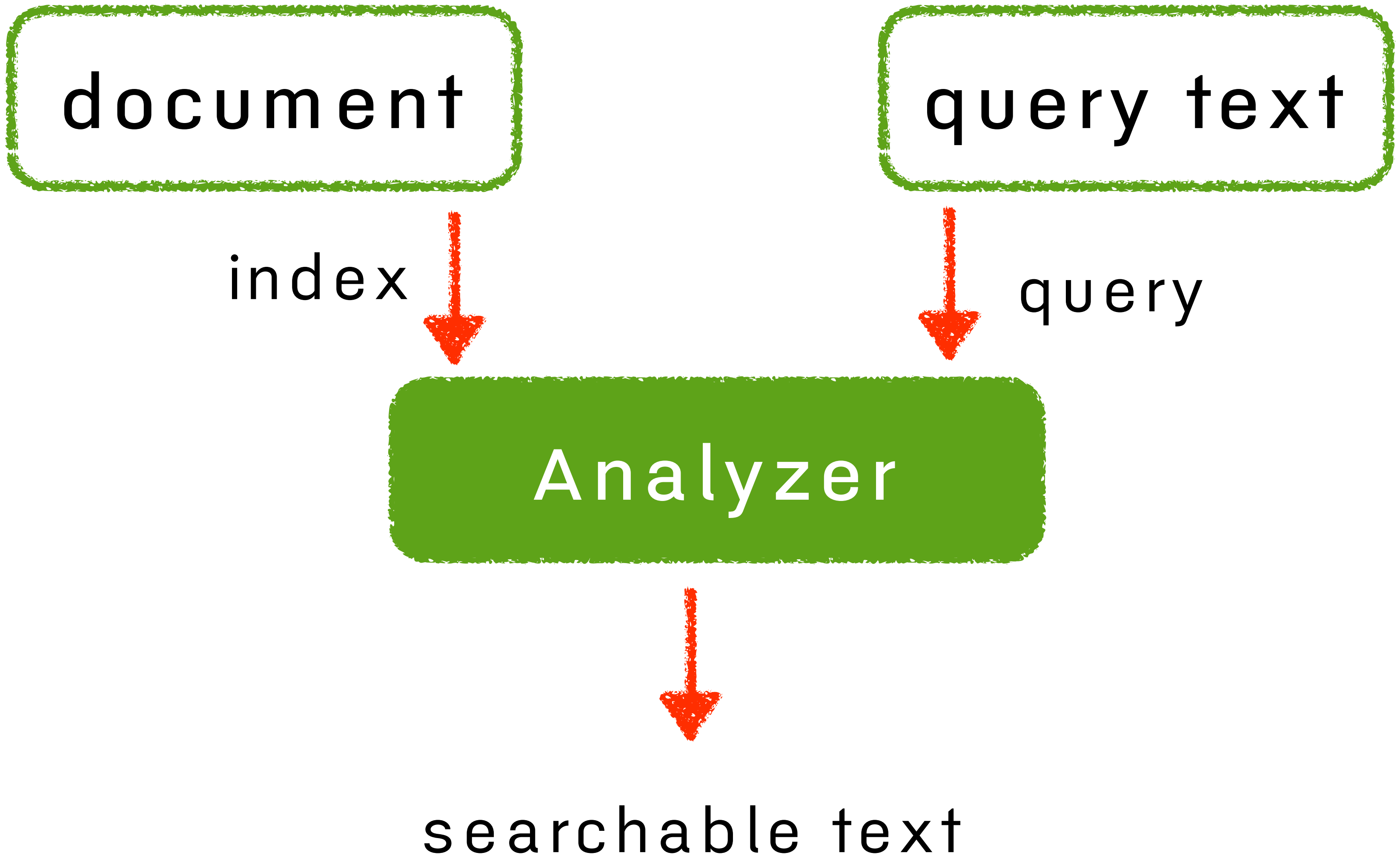
elasticsearch for developer

Analyzer

book_04.txt

Analyzer



Try to retrive data

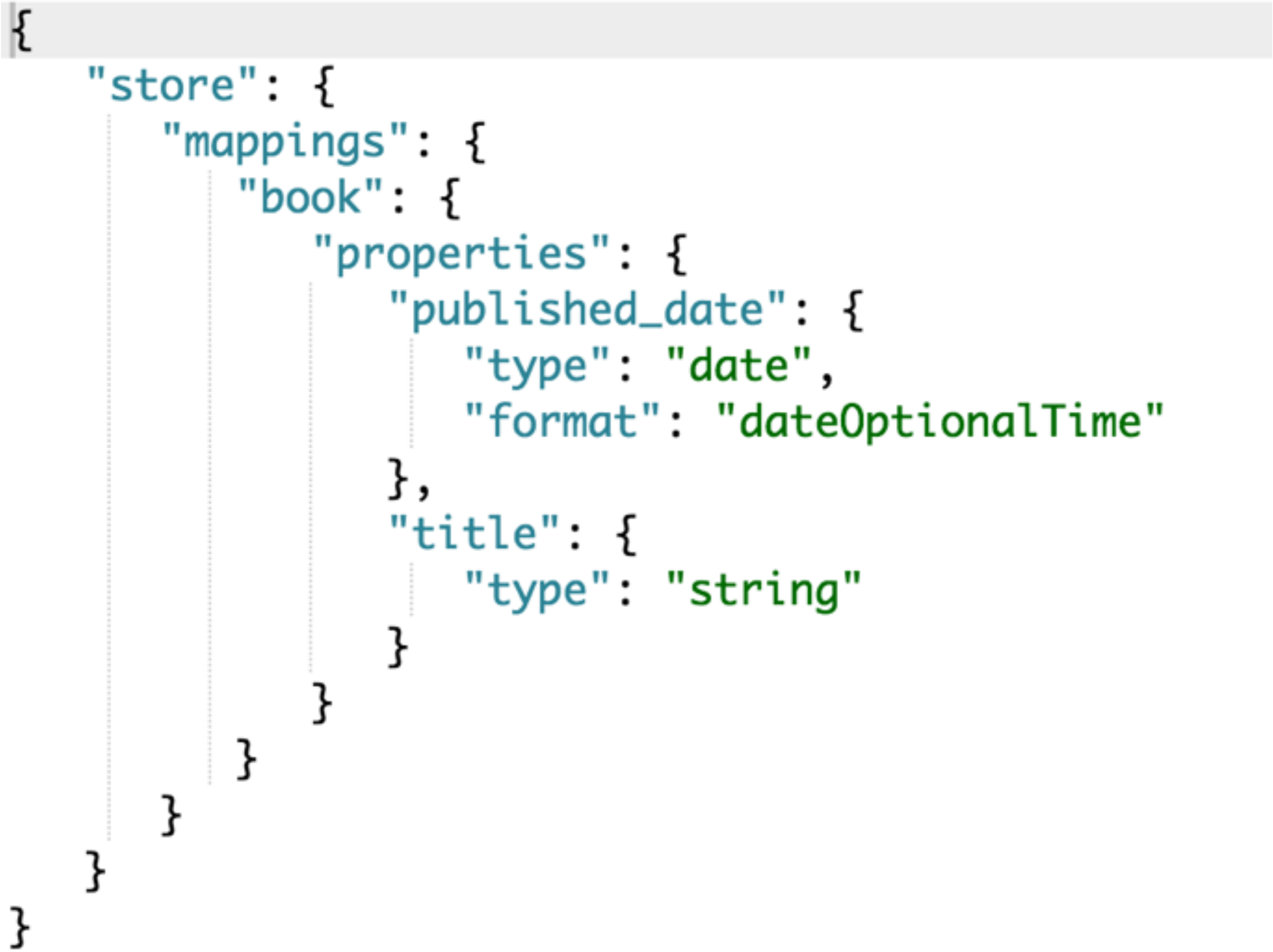
```
GET /store/book/_search?q=2015

GET /store/book/_search?q=2015-01-15

GET /store/book/_search?q=published_date:2015-01-15

GET /store/book/_search?q=published_date:2015
```

Why ?

```
GET /store/book/_search?q=2015 3
GET /store/book/_search?q=2015-01-15 3
GET /store/book/_search?q=published_date:2015-01-15 1
GET /store/book/_search?q=published_date:2015 0
```

Get mapping

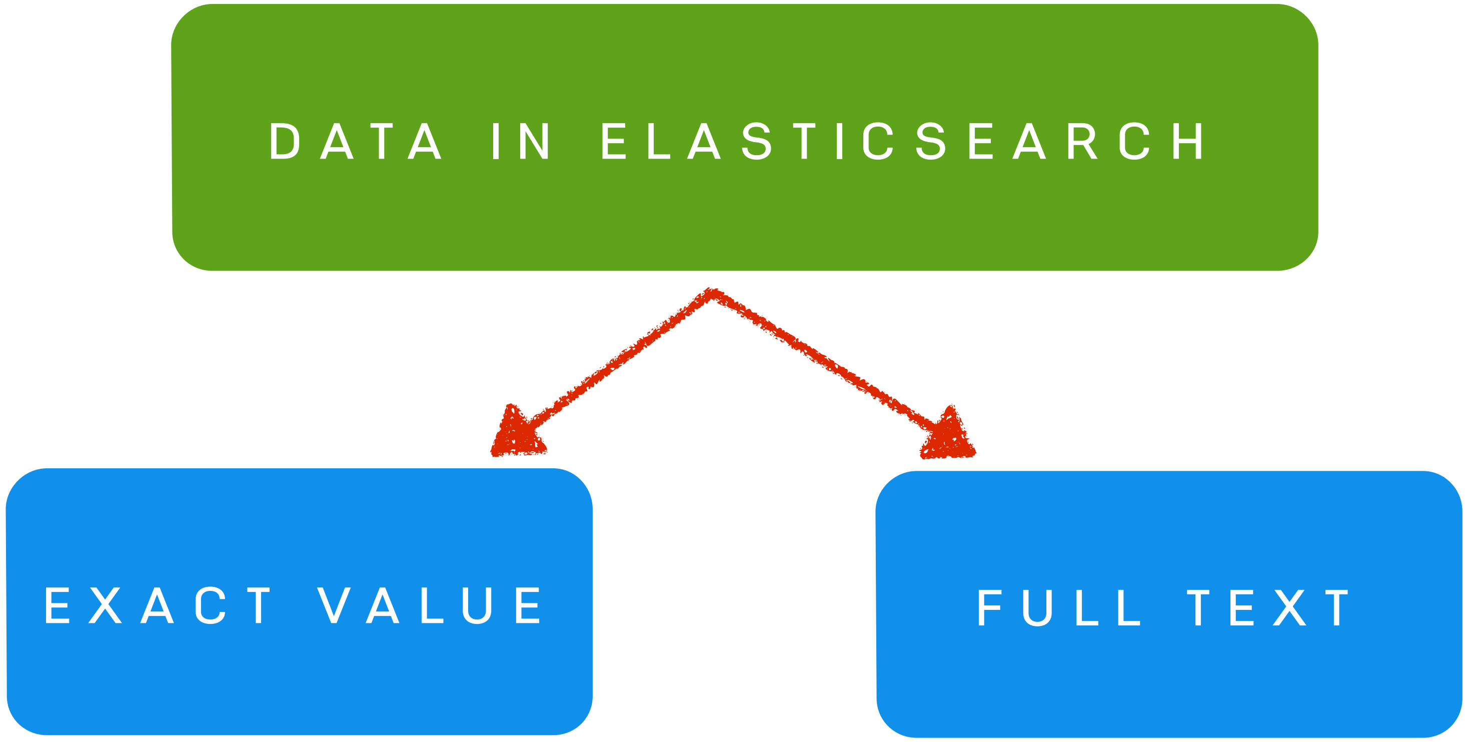
GET /store/_mapping/**book**



```
{
    "store": {
        "mappings": {
            "book": {
                "properties": {
                    "published_date": {
                        "type": "date",
                        "format": "dateOptionalTime"
                    },
                    "title": {
                        "type": "string"
                    }
                }
            }
        }
    }
}
```

Exact value vs Full text

คือเหตุผลที่เราควรแยก search engine ออกจาก database



DATA IN ELASTICSEARCH

EXACT VALUE

FULL TEXT

Exact value

“Foo” != “foo”

“2015” != “2015-01-15”

Full text

Human language

Natural language

Relevant with the query

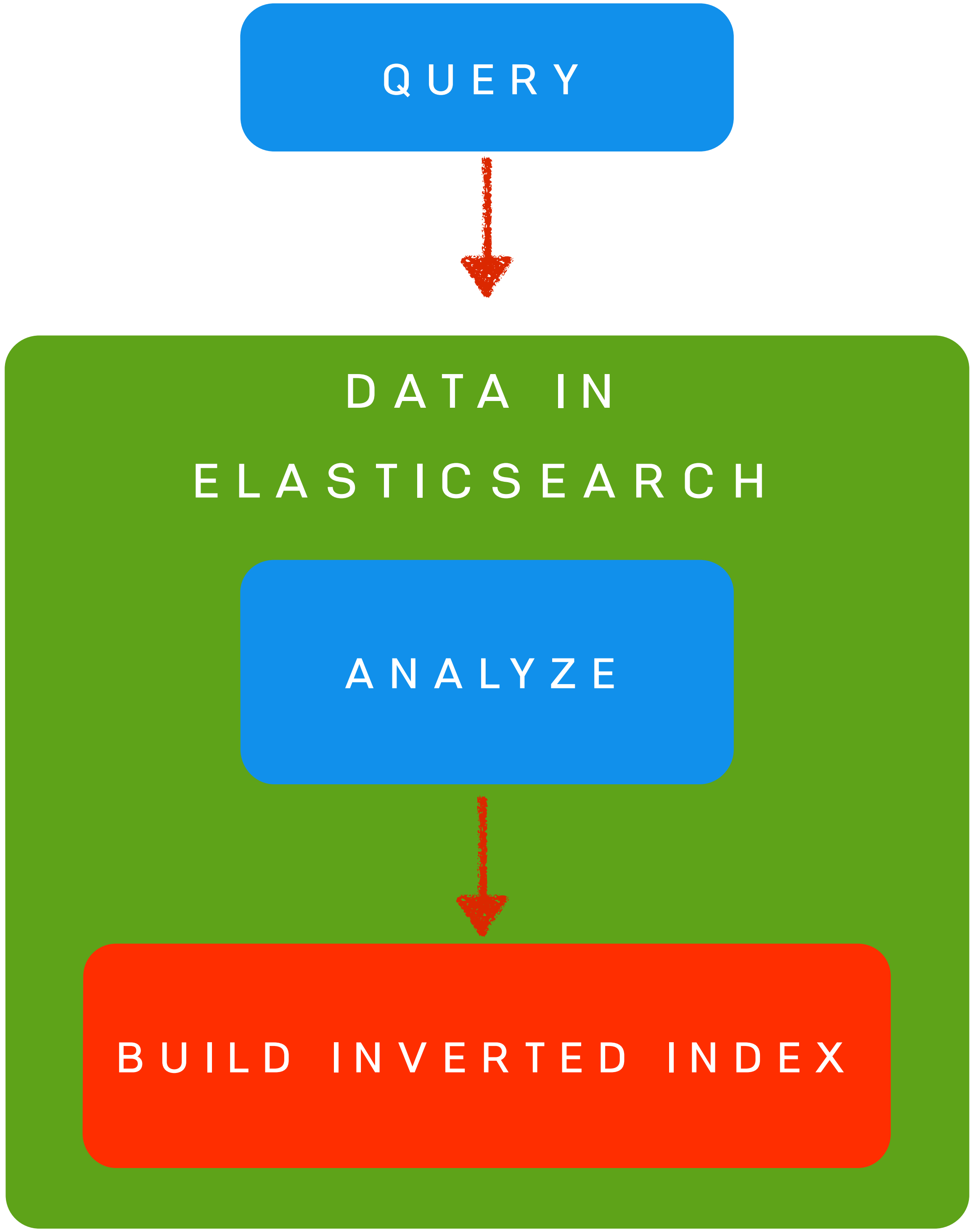
Full text

TH => Thailand

jump => jumped, jumps, jumping

search => Elasticsearch

How Elasticsearch Do ?



QUERY

DATA IN ELASTICSEARCH

ANALYZE

BUILD INVERTED INDEX

SPRINT3R

Siam Chamnankit Co., Ltd., Odd-e (Thailand) Co., Ltd. and Alliance

Inverted index

ออกแบบมาเพื่อ Fast full text search

THE QUICK BROWN FOX JUMPED OVER THE LAZY DOG

Inverted index ?

Inverted index

1. The quick brown fox jumped over the lazy dog

2. Quick brown foxes leap over lazy dogs in summer

Term	Document 1	Document 2
Quick		X
the	X	
quick	X	
brown	X	X
fox	X	
foxes		X
jumped	X	
over	X	
the	X	

Inverted index

เมื่อค้นหาคำว่า quick brown

Term	Document 1	Document 2
quick	X	
brown	X	X
Total	**2**	**1**

ปัญหาของ Inverted index

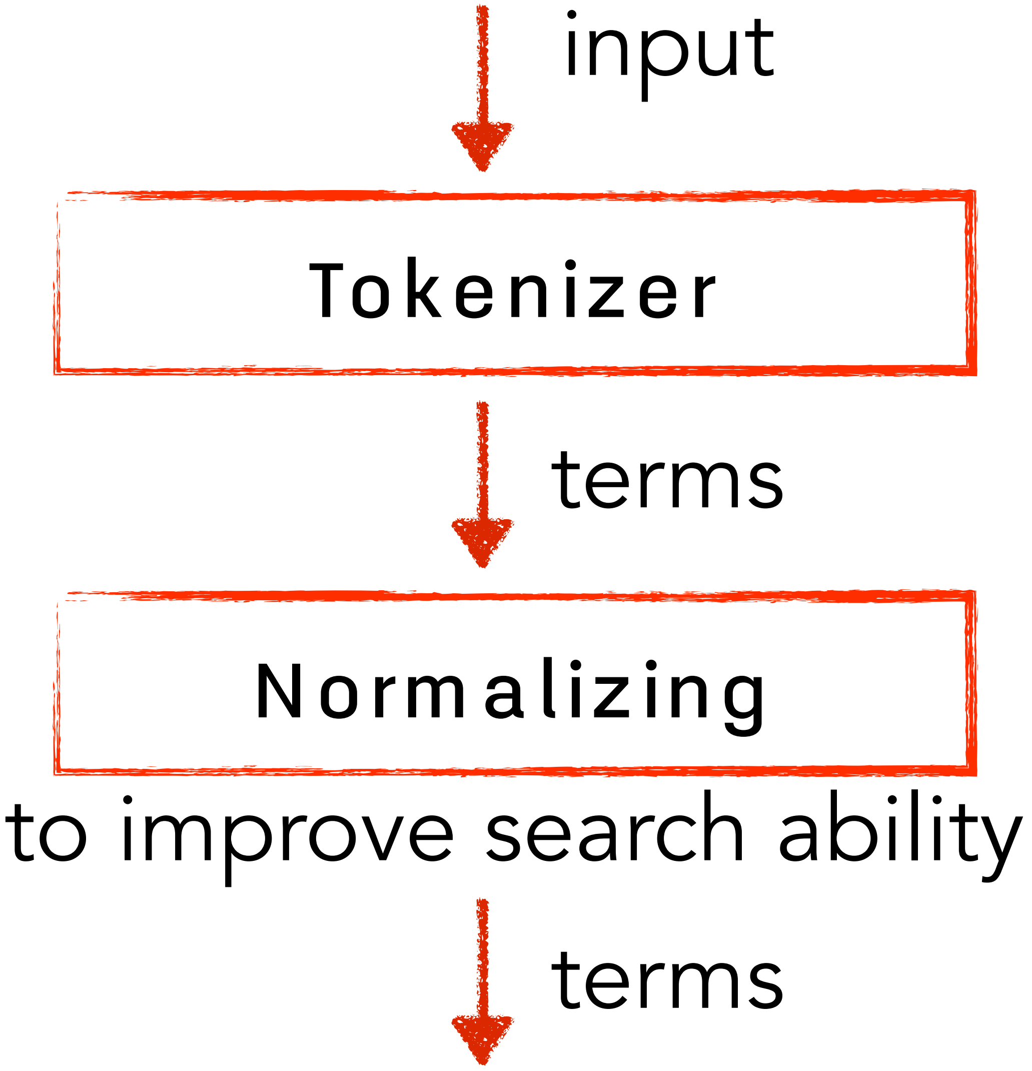
Term	Document 1	Document 2
Quick		X
quick	X	
brown	X	X
fox	X	
foxes		X
over	X	
jumped	X	
leap	X	

ข้อมูลถูกจัดการอย่างไร ?

การแบ่งคำ การจัดรูปแบบคำ การกรองข้อมูล

ANALYSIS PROCESS

Analysis process



input

Tokenizer

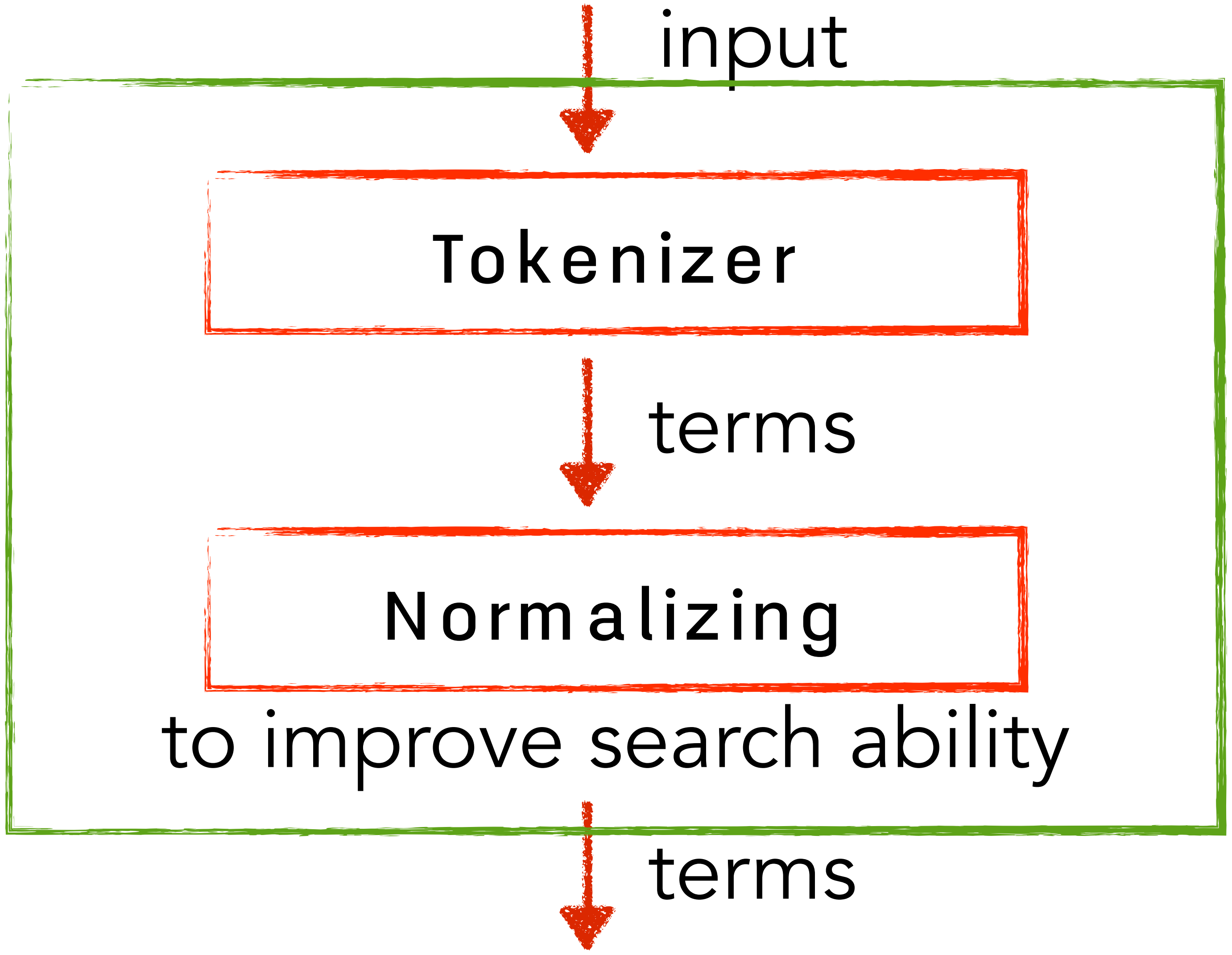
terms

Normalizing

to improve search ability

terms

Analyzer



input

Tokenizer

terms

Normalizing

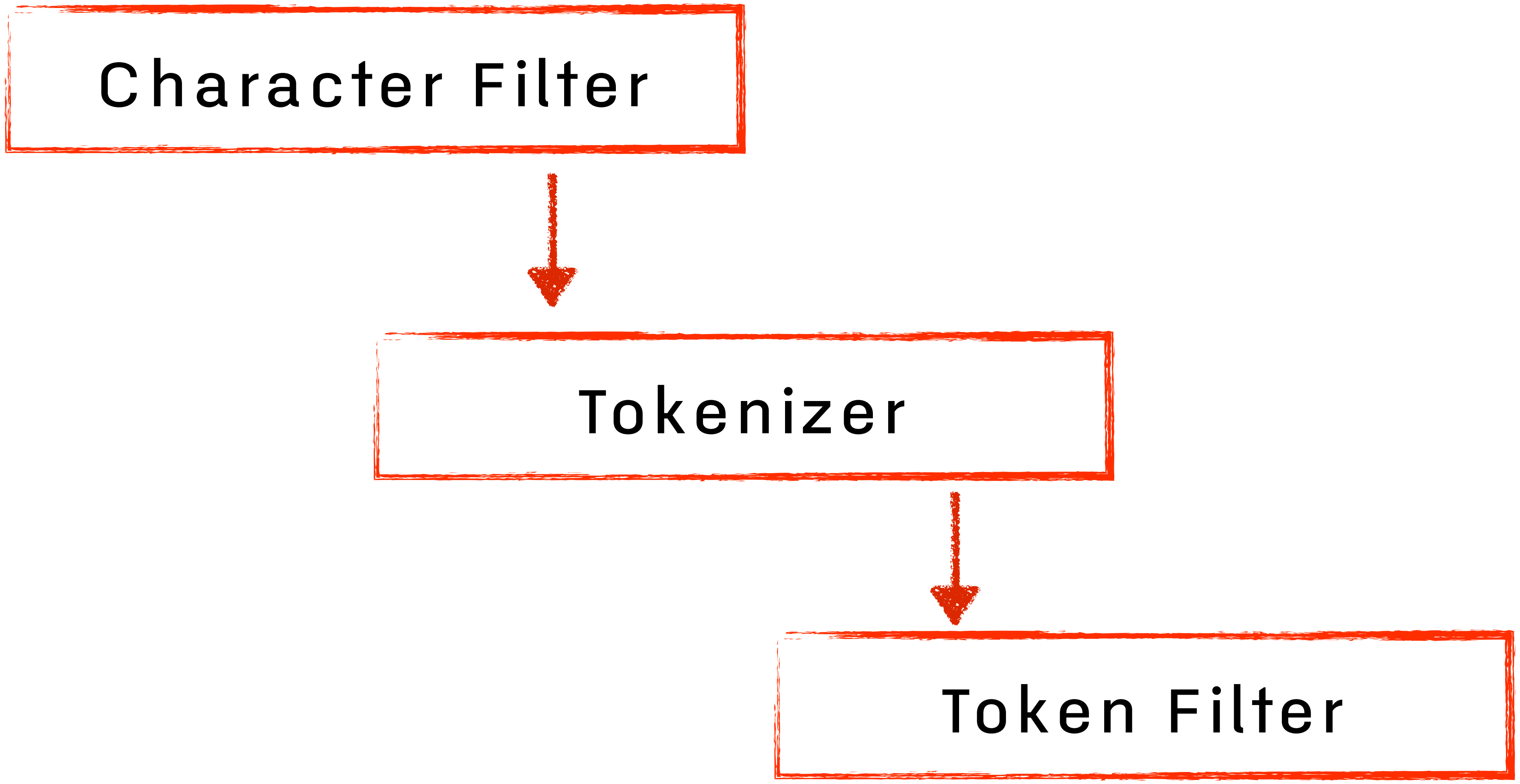
to improve search ability

terms

SPRINT3R

Siam Chamnankit Co., Ltd., Odd-e (Thailand) Co., Ltd. and Alliance

Analyzer

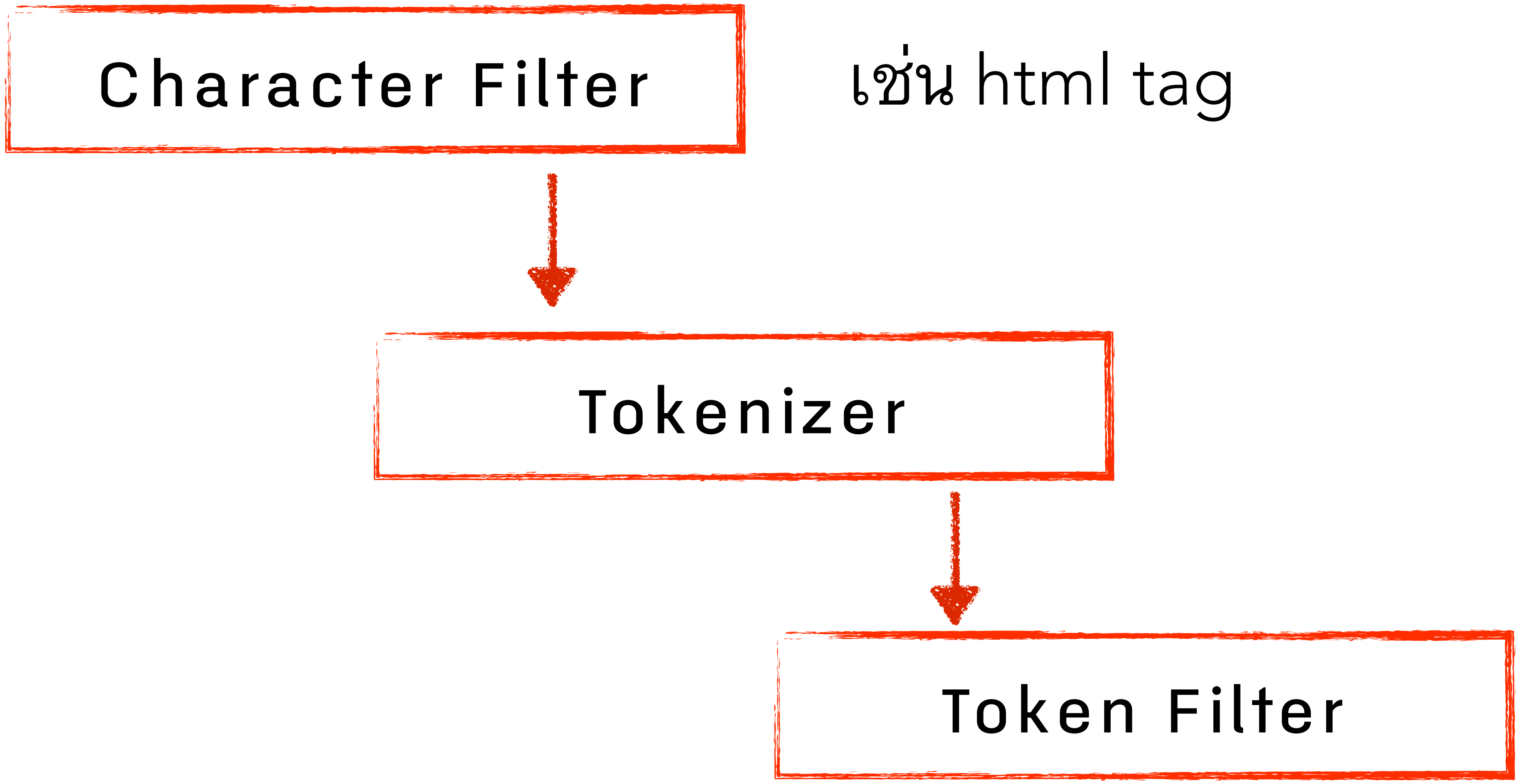


Character Filter

Tokenizer

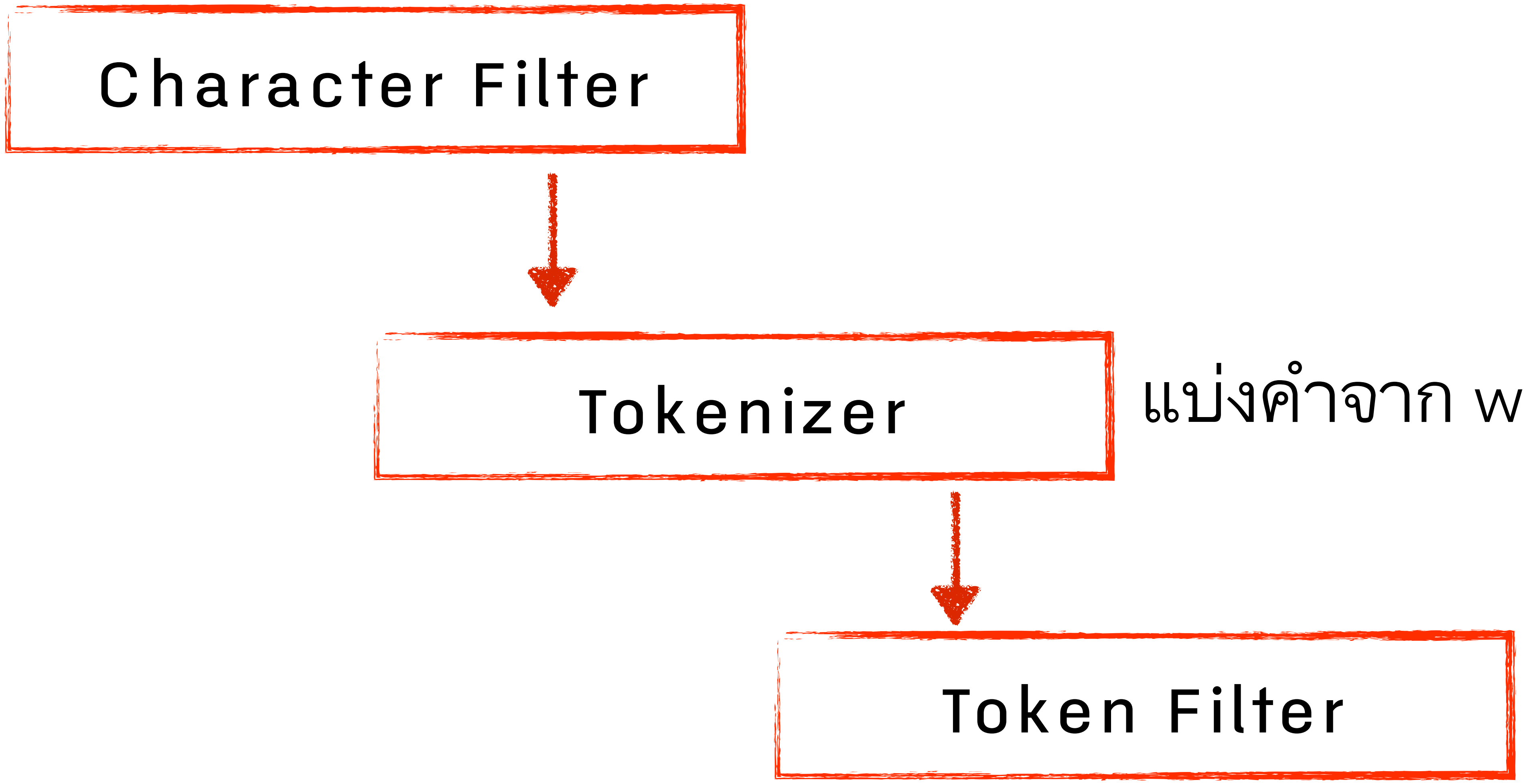
Token Filter

Analyzer



เช่น html tag

Analyzer



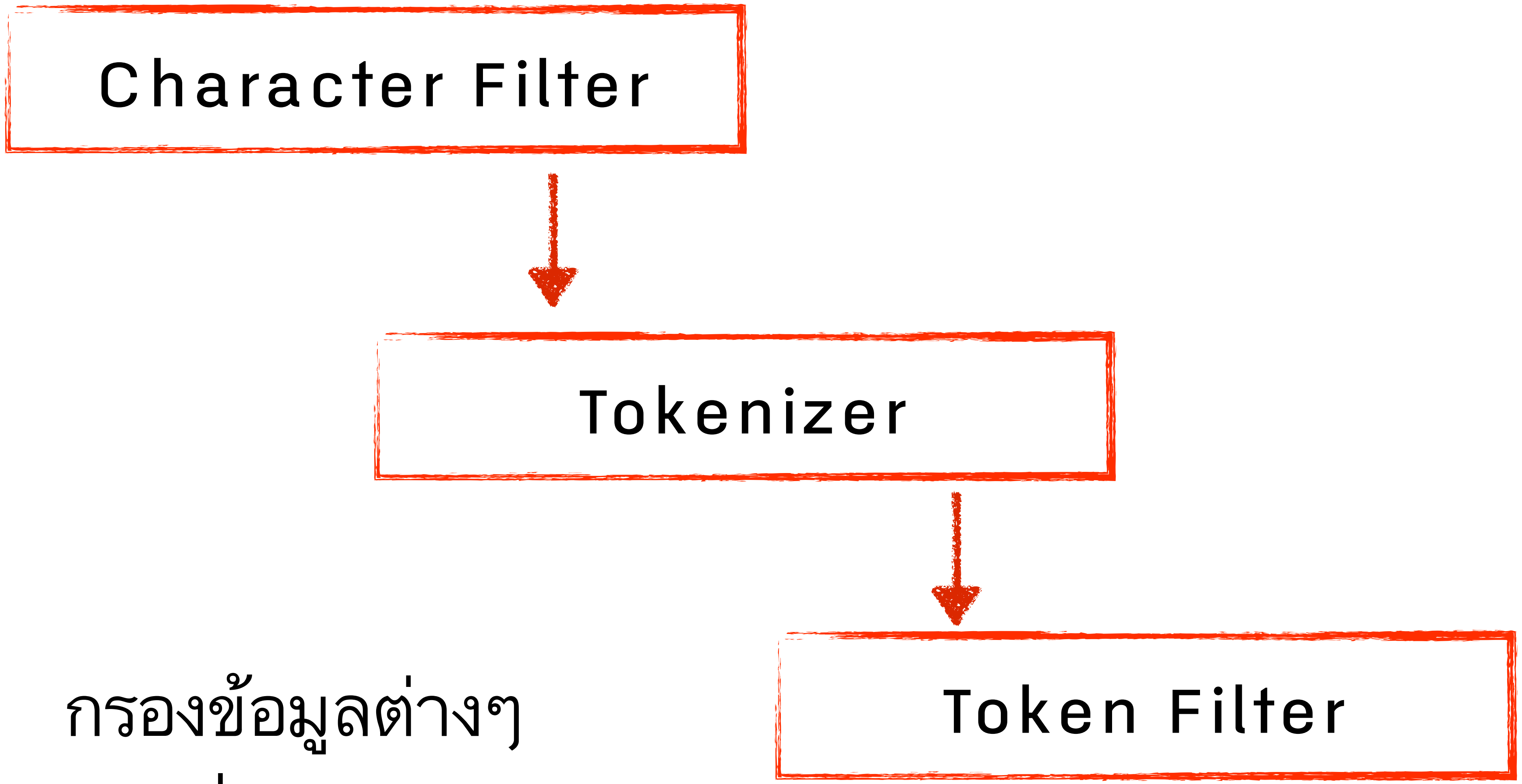
Character Filter

Tokenizer

แบ่งคำจาก white space

Token Filter

Analyzer



Character Filter

Tokenizer

Token Filter

กรองข้อมูลต่างๆ

- Lowercase
- Stopword
- Synonyms

Build-in analyzer

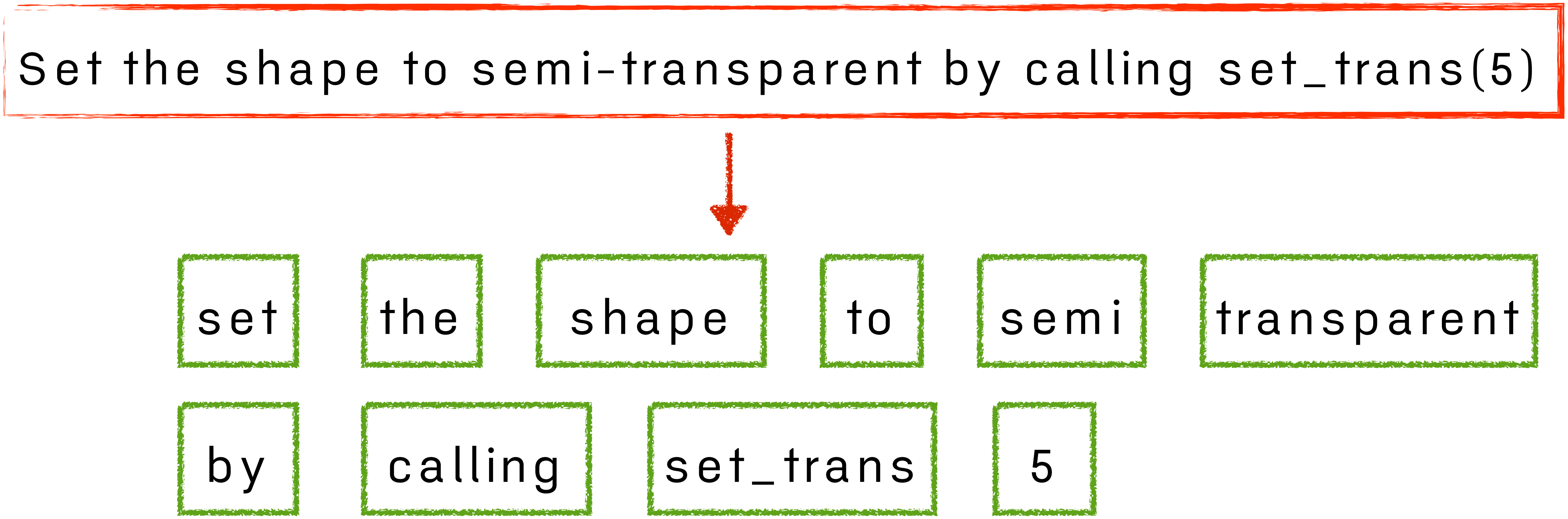
Standard analyzer [default]

Simple analyzer

Whitespace analyzer

Language analyzer

Standard analyzer



Set the shape to semi-transparent by calling set_trans(5)

set

the

shape

to

semi

transparent

by

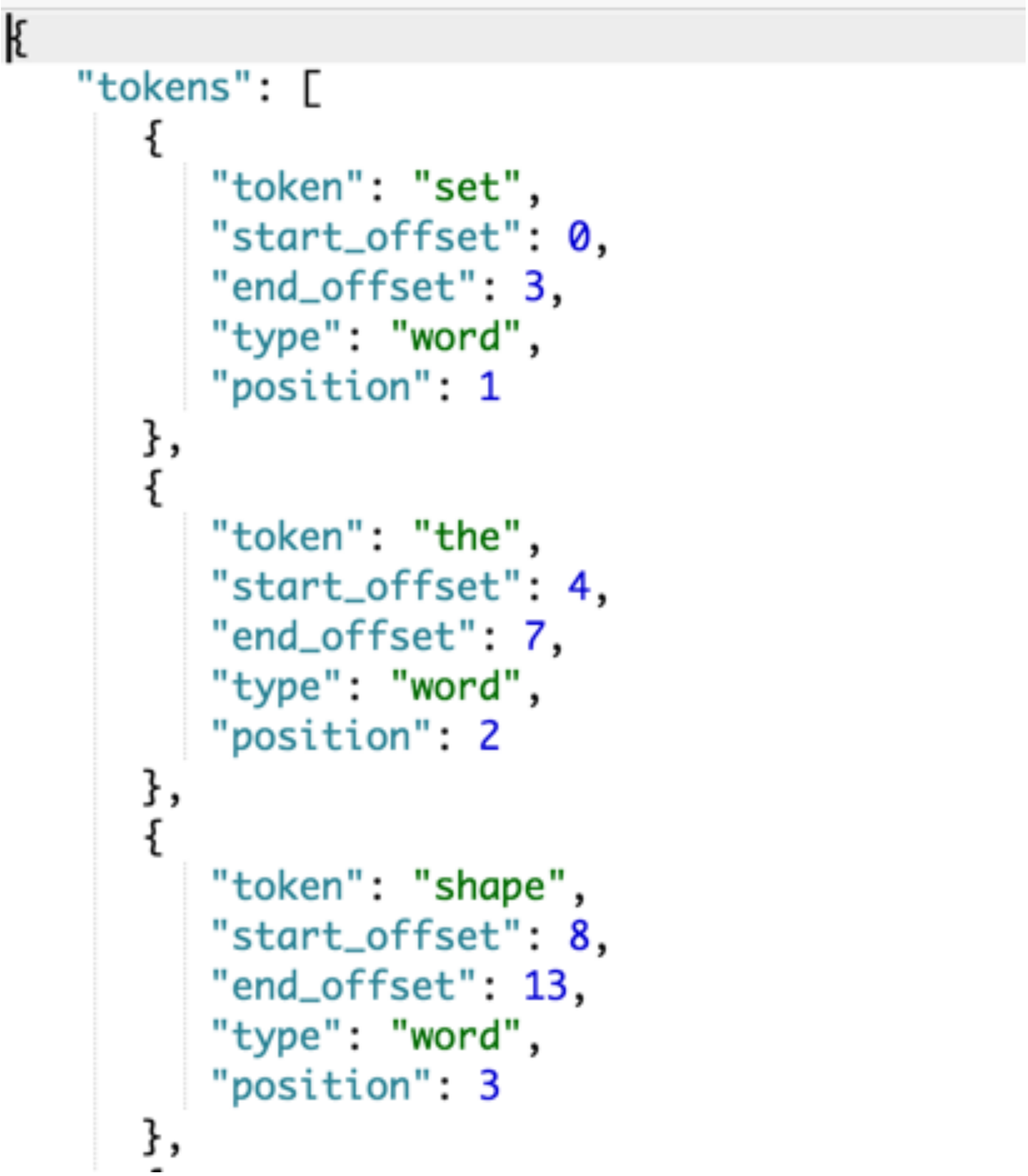
calling

set_trans

5

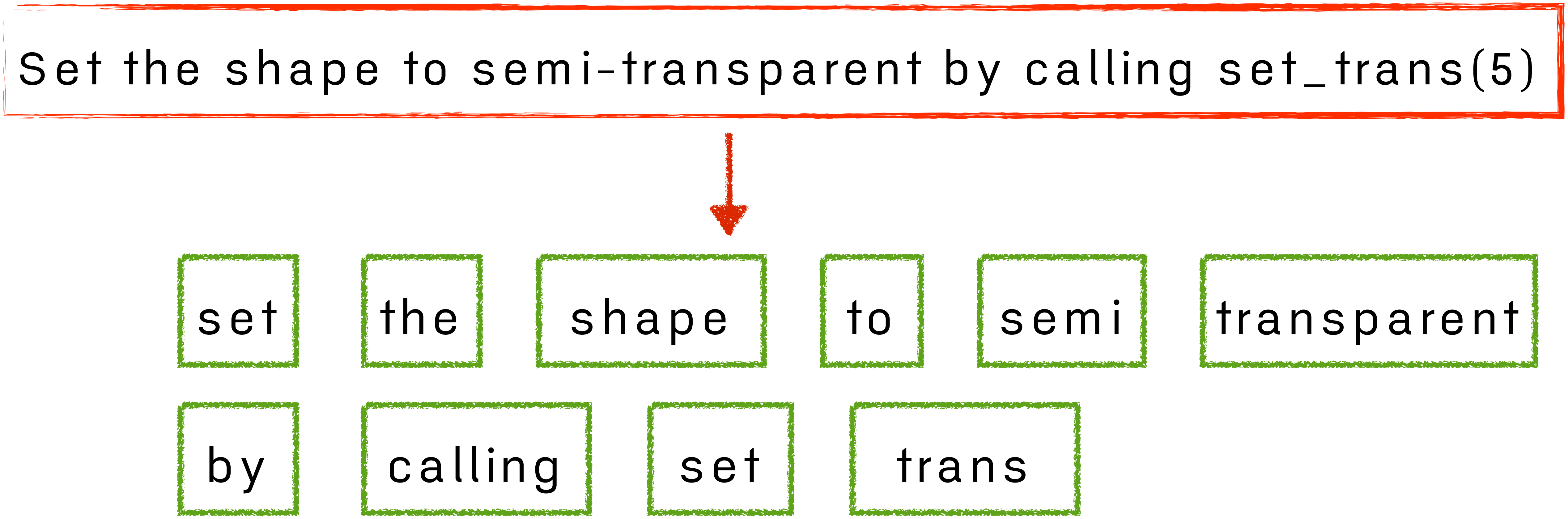
Test standard analyzer

```
GET /_analyze?analyzer=standard&text=xxx
```



```
{
  "tokens": [
    {
      "token": "set",
      "start_offset": 0,
      "end_offset": 3,
      "type": "word",
      "position": 1
    },
    {
      "token": "the",
      "start_offset": 4,
      "end_offset": 7,
      "type": "word",
      "position": 2
    },
    {
      "token": "shape",
      "start_offset": 8,
      "end_offset": 13,
      "type": "word",
      "position": 3
    },
```

Simple analyzer



Set the shape to semi-transparent by calling set_trans(5)

set

the

shape

to

semi

transparent

by

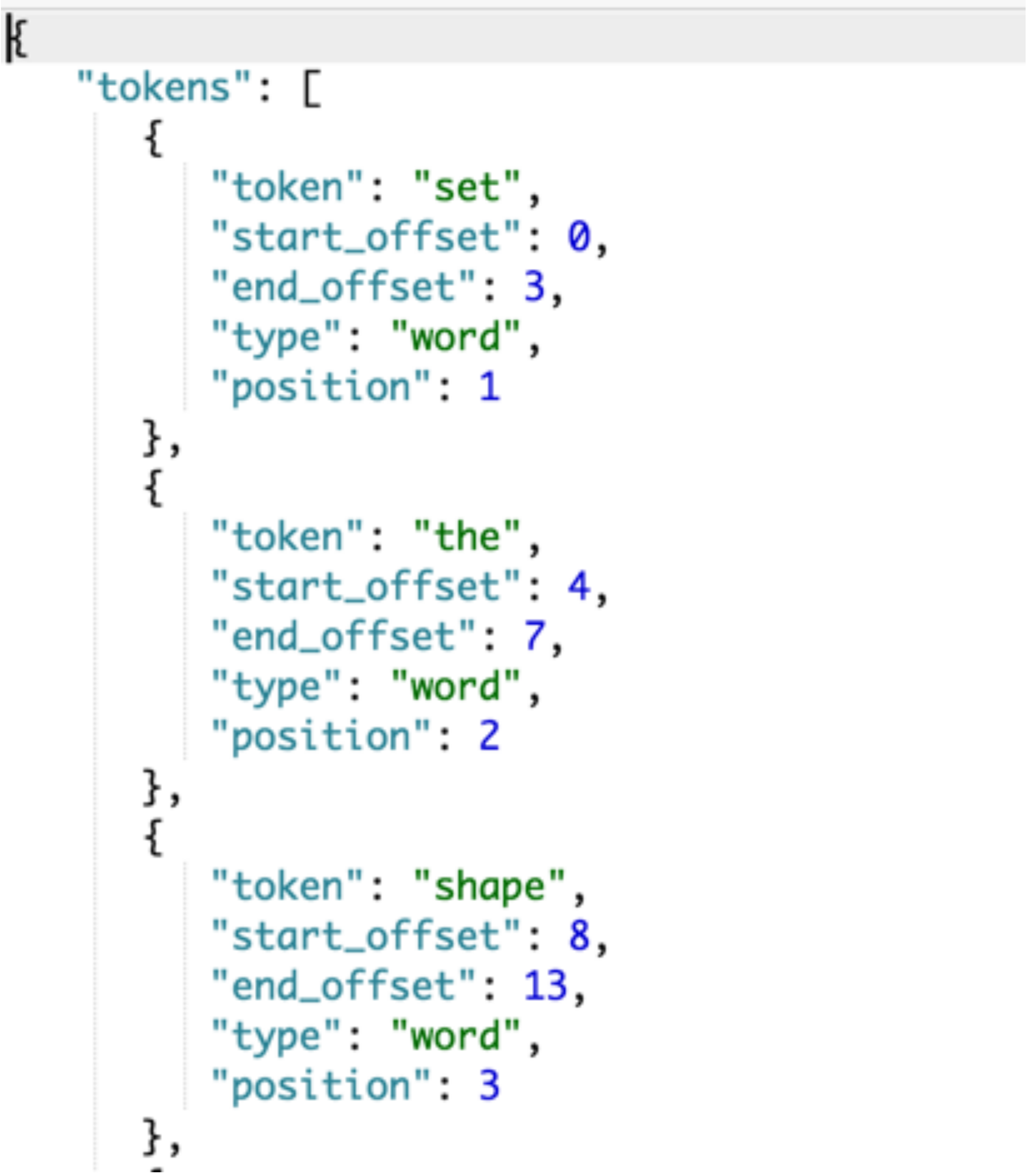
calling

set

trans

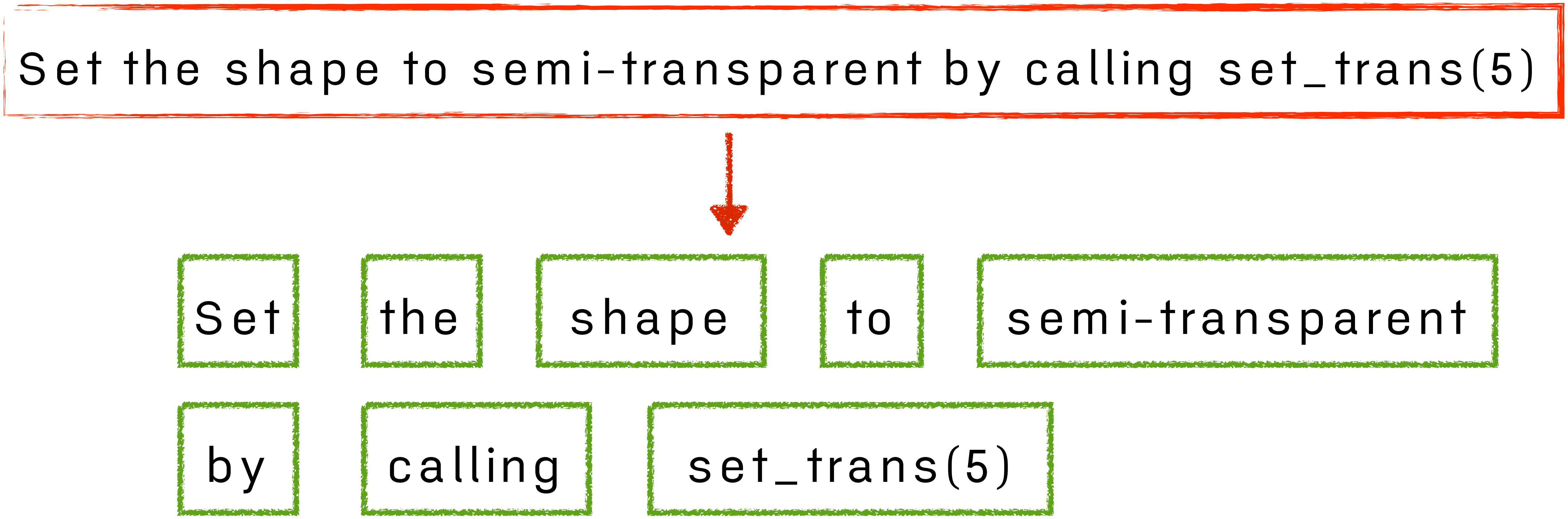
Test simple analyzer

GET /_analyze?analyzer=**simple**&text=xxx



```
{
    "tokens": [
      {
        "token": "set",
        "start_offset": 0,
        "end_offset": 3,
        "type": "word",
        "position": 1
      },
      {
        "token": "the",
        "start_offset": 4,
        "end_offset": 7,
        "type": "word",
        "position": 2
      },
      {
        "token": "shape",
        "start_offset": 8,
        "end_offset": 13,
        "type": "word",
        "position": 3
      },
```

Whitespace analyzer



Set the shape to semi-transparent by calling set_trans(5)

Set

the

shape

to

semi-transparent

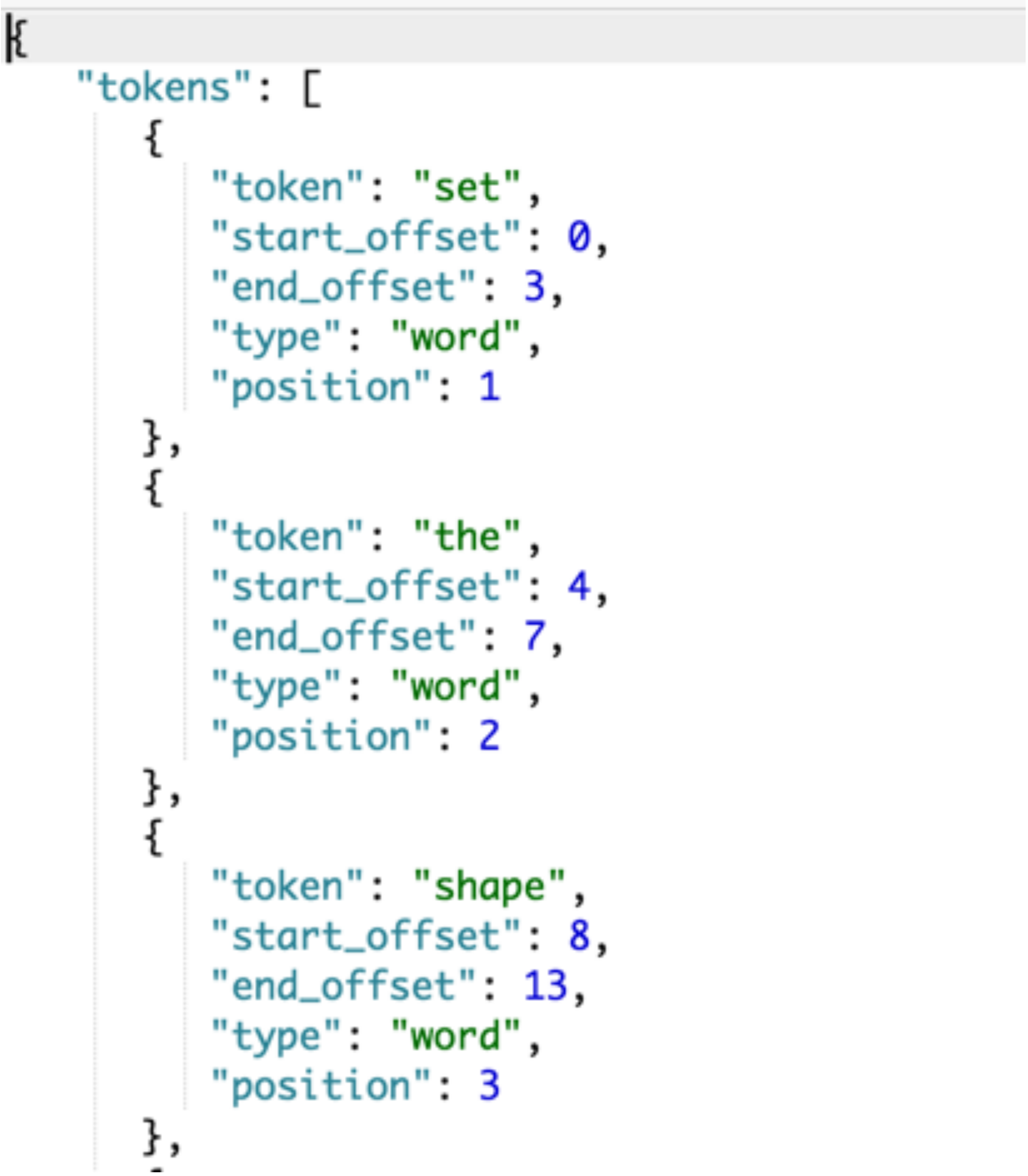
by

calling

set_trans(5)

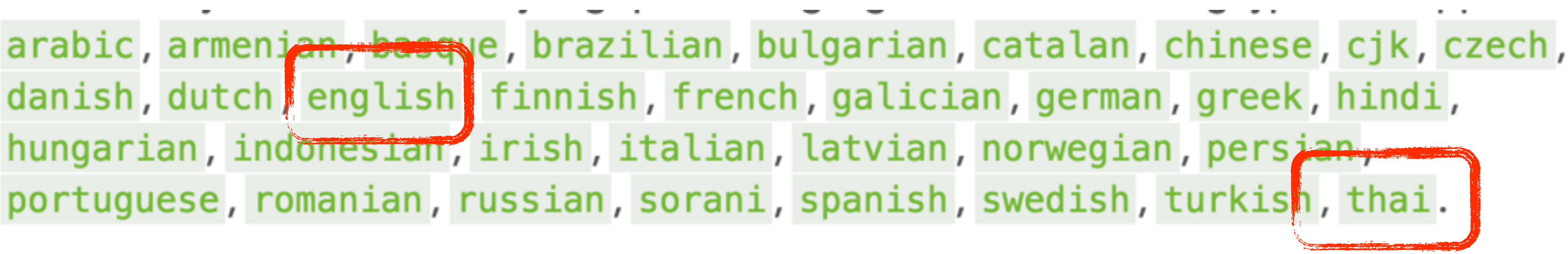
Test whitespace analyzer

GET /_analyze?analyzer=**whitespace**&text=xxxx



```
{
    "tokens": [
      {
        "token": "set",
        "start_offset": 0,
        "end_offset": 3,
        "type": "word",
        "position": 1
      },
      {
        "token": "the",
        "start_offset": 4,
        "end_offset": 7,
        "type": "word",
        "position": 2
      },
      {
        "token": "shape",
        "start_offset": 8,
        "end_offset": 13,
        "type": "word",
        "position": 3
      },
```

Language analyzer



arabic, armenian, basque, brazilian, bulgarian, catalan, chinese, cjk, czech, danish, dutch, english, finnish, french, galician, german, greek, hindi, hungarian, indonesian, irish, italian, latvian, norwegian, persian, portuguese, romanian, russian, sorani, spanish, swedish, turkish, thai.

Stopwords

Excluding words

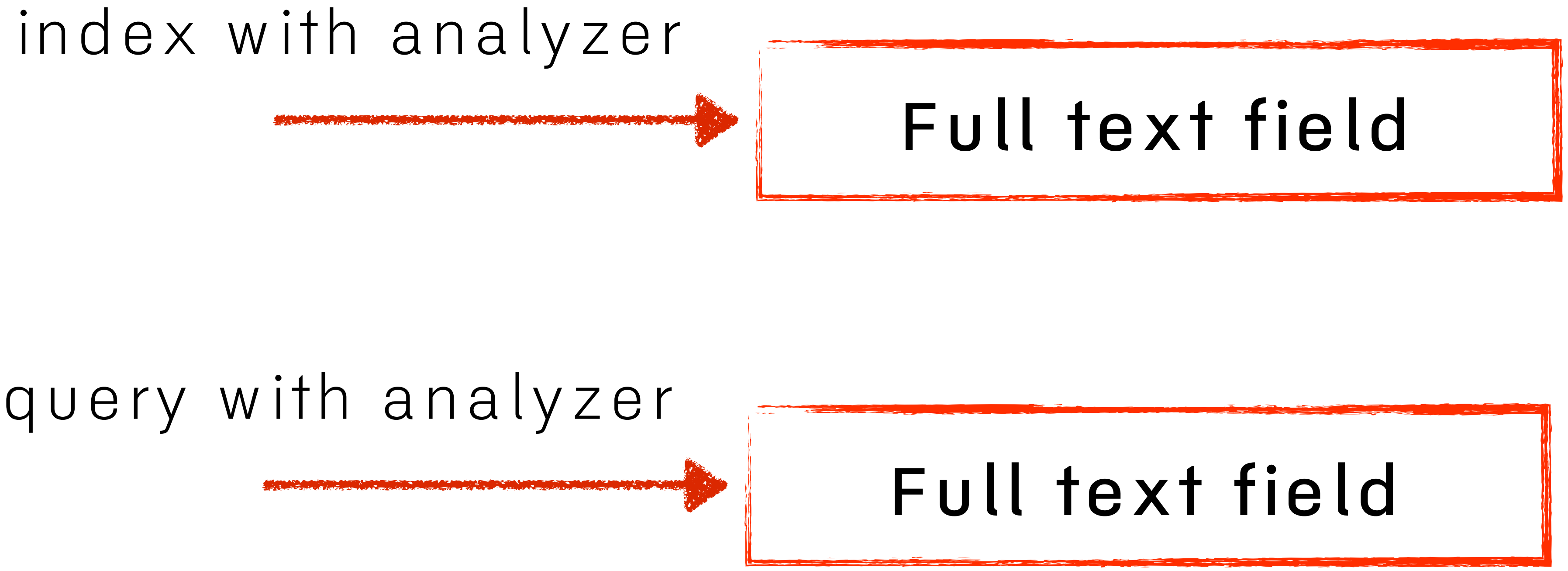
http://www.elasticsearch.org/guide/en/elasticsearch/reference/current/analysis-lang-analyzer.html

When to use Analyzer ?

Index a document

Query string

When to use Analyzer ?



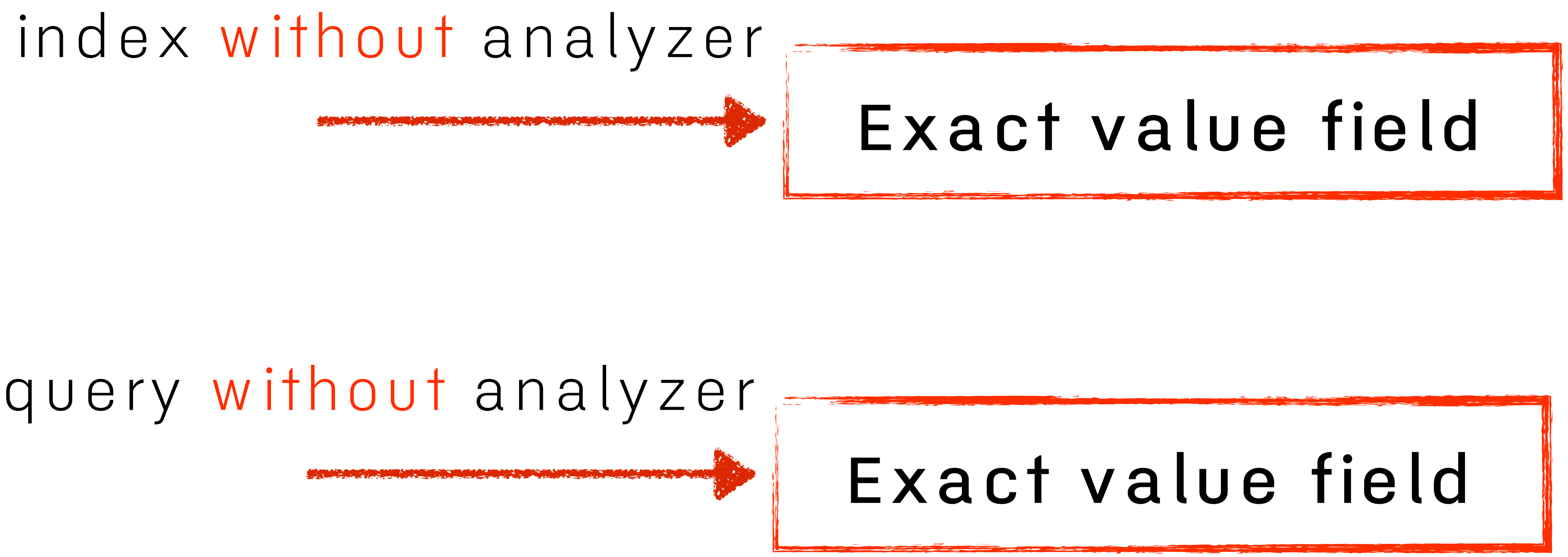
index with analyzer

Full text field

query with analyzer

Full text field

When to use Analyzer ?



index without analyzer

Exact value field

query without analyzer

Exact value field

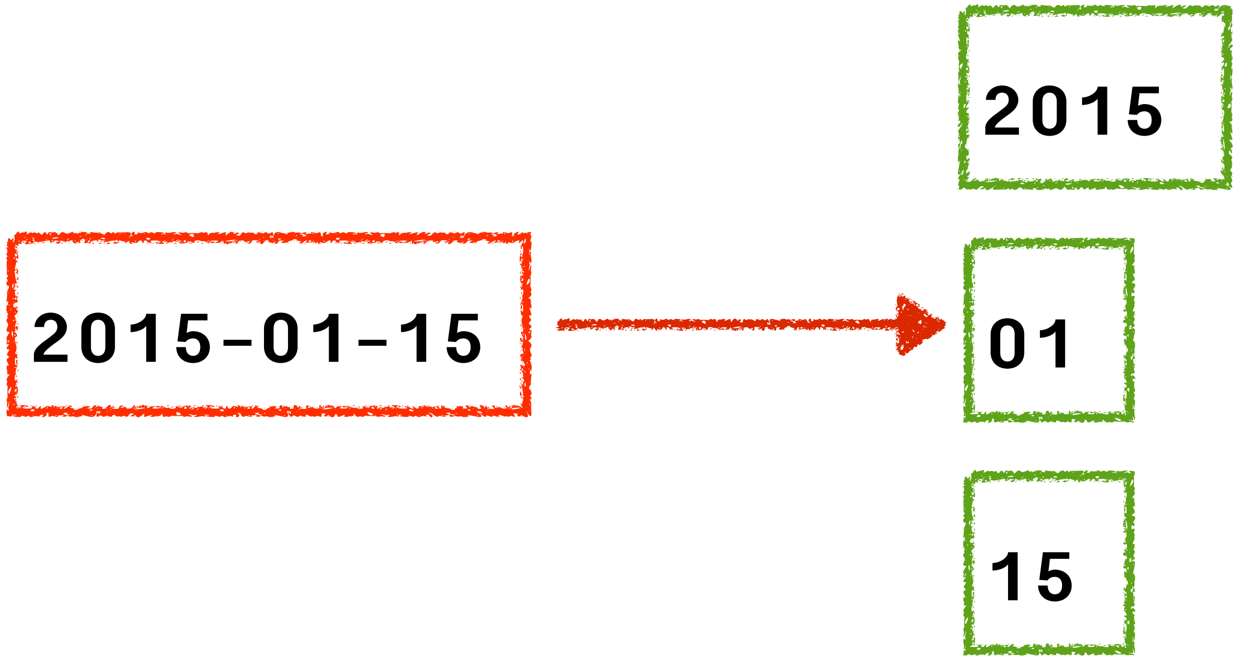
Why ?

```
GET /store/book/_search?q=2015 3
GET /store/book/_search?q=2015-01-15 3
GET /store/book/_search?q=published_date:2015-01-15 1
GET /store/book/_search?q=published_date:2015 0
```

Why ?

date field เก็บข้อมูลแบบ Exact value

_all field คือ full text field



2015-01-15

2015

01

15