Analyzer 101

how to work with analyzer in elasticsearch

Medcl

About me

- Follow Elasticsearch Since v0.5, 2010
- Joined Elastic Since September, 2015
- @medcl
- http://github.com/medcl

We love Elasticsearch

- Elasticsearch is build on top of Lucene!
- Json IN Json OUT!
- Fancy Distributed and Scalability!
- …

But What if …

- “为什么查不到,明明有的”
- “怎么出来这个鬼,数据怎么出来的”

什么情况

- “这个字段可以用模糊匹配么”
- “这个字段可以用Aggregation么”
- “为什么索引这么大呀”

If you don’t know why, you won’t know how to fix

TELL ME WHY?

Agenda

- What’s the Analyzer
- Why Analyzer Matters
- How Analyzer Works
- Analyzer for Chinese
- Analyzer In Elasticsearch
- How to Choose Analyzer

What’s the Analyzer

什么是ANALYZER

Let’s go back to basis

- Lucene & Invert Index
- How index works?
- How search works?

Inverted Index

- Doc1:
  - The quick brown fox jumped over the lazy dogs.
- Doc2:
  - The yellow dog is mine.
- Doc3:
  - I don’t have brown bag!

Inverted Index

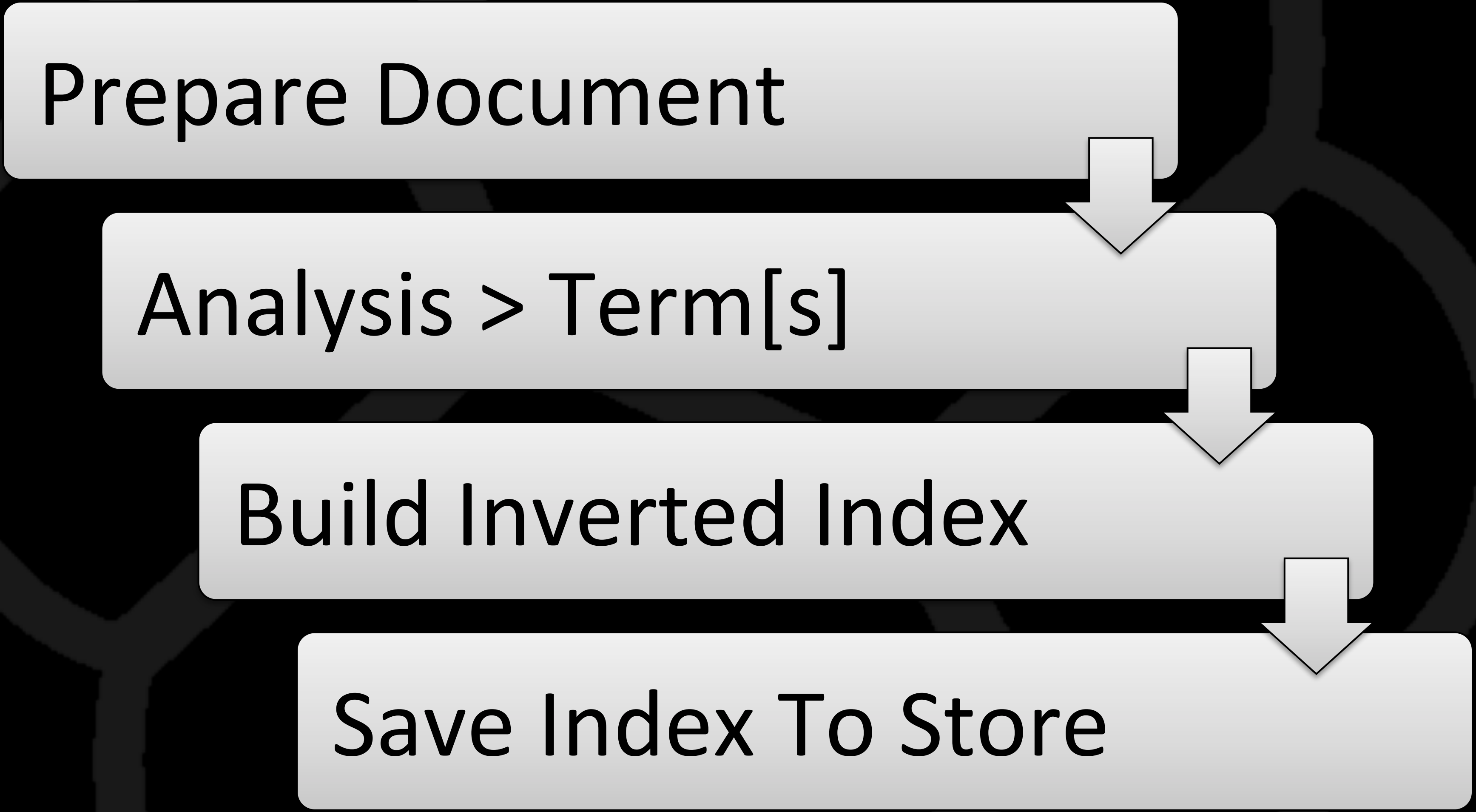
Term Name	Document ID
The	Doc1,Doc2
quick	Doc1,
Brown	Doc1,Doc3
Fox	Doc1
… …	… …

Search Index

- “The”
  - Doc1,Doc2
- “The Fox” => “The” AND “Fox”
  - Doc1

Term Name	Document ID
The	Doc1,Doc2
quick	Doc1,
Brown	Doc1,Doc3
Fox	Doc1
… …	… …

Index workflow



Prepare Document

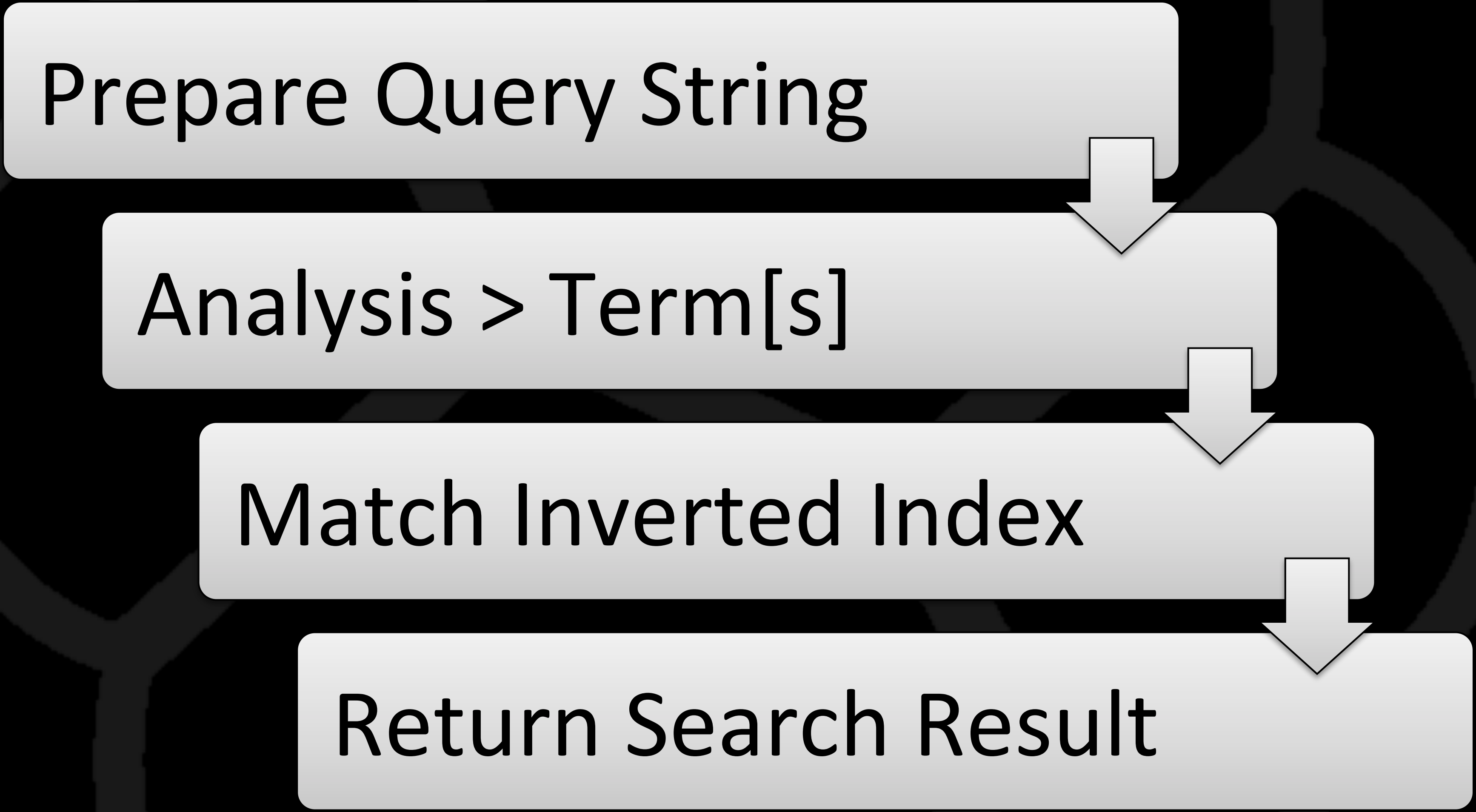
Analysis > Term[s]

Build Inverted Index

Save Index To Store

Brief version, we ignore details

Search workflow



Prepare Query String

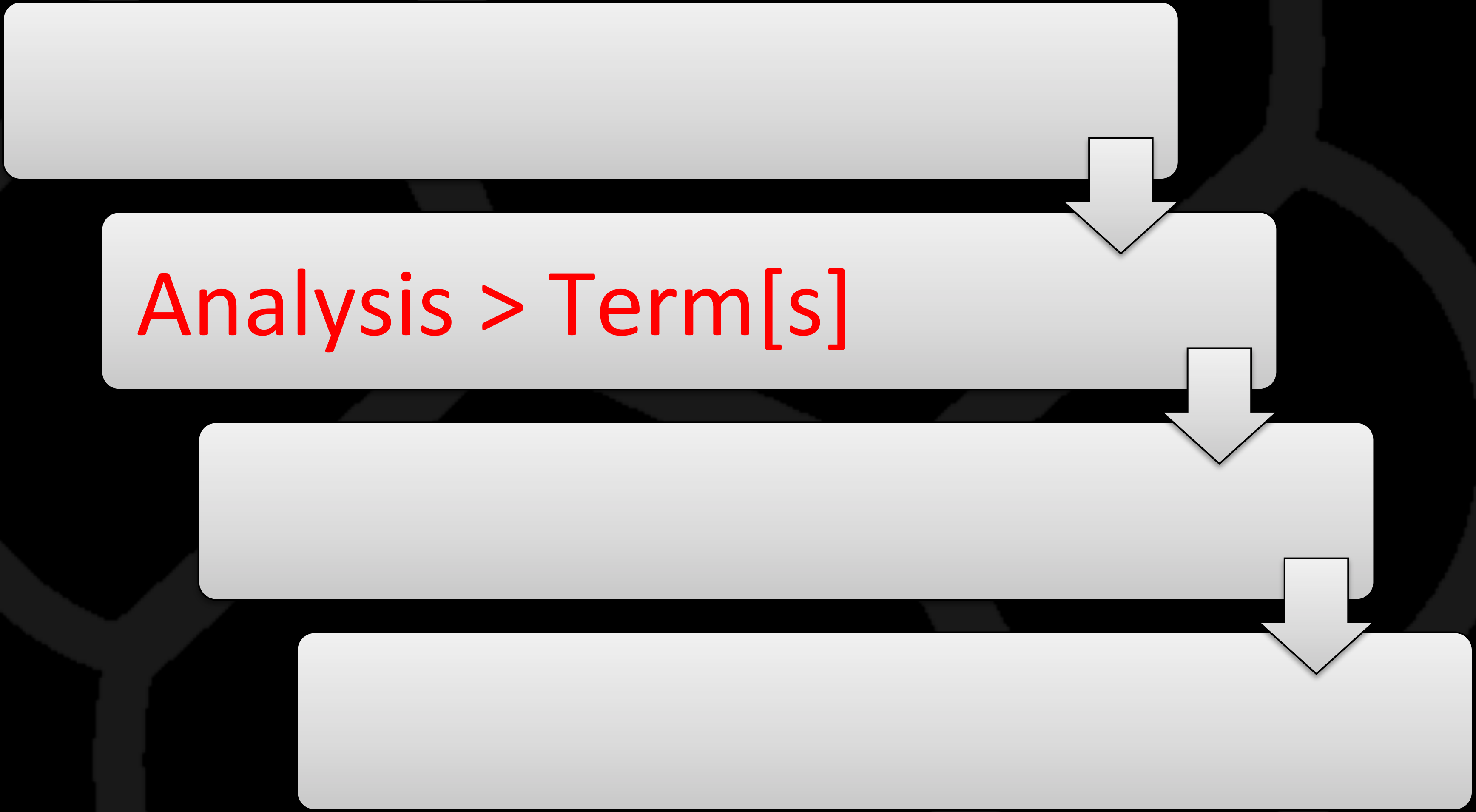
Analysis > Term[s]

Match Inverted Index

Return Search Result

Brief version, we ignore details

Highlighted workflow



Analysis > Term[s]

Brief version, we ignore details

Text->Terms?

Analysis

Analyzer

What is the Analysis

- Lucene is an indexing and search library, accepts only plain text input.
- Text analysis.
  - Lucene use **Analyzer** to Analysis,convert text into indexable/ searchable tokens.

What is the Analyzer

- Analyzers create tokens from the character stream.
- An ana analysis
  proces performing
  any nu could
  include unctuation,
  remov ercasing
  (also c mmon
  words,
  (stemming), or changing words into the basic form (lemmatization).

文本搅拌机

WHY ANALYZER MATTERS

The Key of Hit

Term[s]

Of Inverted Index
Of Analyzed Query String

- Match - > HIT
- Not Match -> MISS

The key of Condition

- Parameter:default_operator
  - AND: must match all terms
  - OR: match or not is OK

Example: _search?q=quick fox

The key of Condition

- BoolQuery
  - Must: must match the term
  - Must Not: must not match the term
  - Should: don’t really care match or not

Example:

_search?q=(“quick” AND “fox”) OR “dog”

The key of Term Query

- Term Query:
  - “QueryString” will be direct used as term to match the index

The key of Text Query

- Match_all / TextQuery / QueryStringQuery etc:
  - “QueryString” will be Analyzed to generate terms to match the index

The key of Range Query

- Range Query:
  - “QueryString” will also generated sequence terms to match the index

It Matters

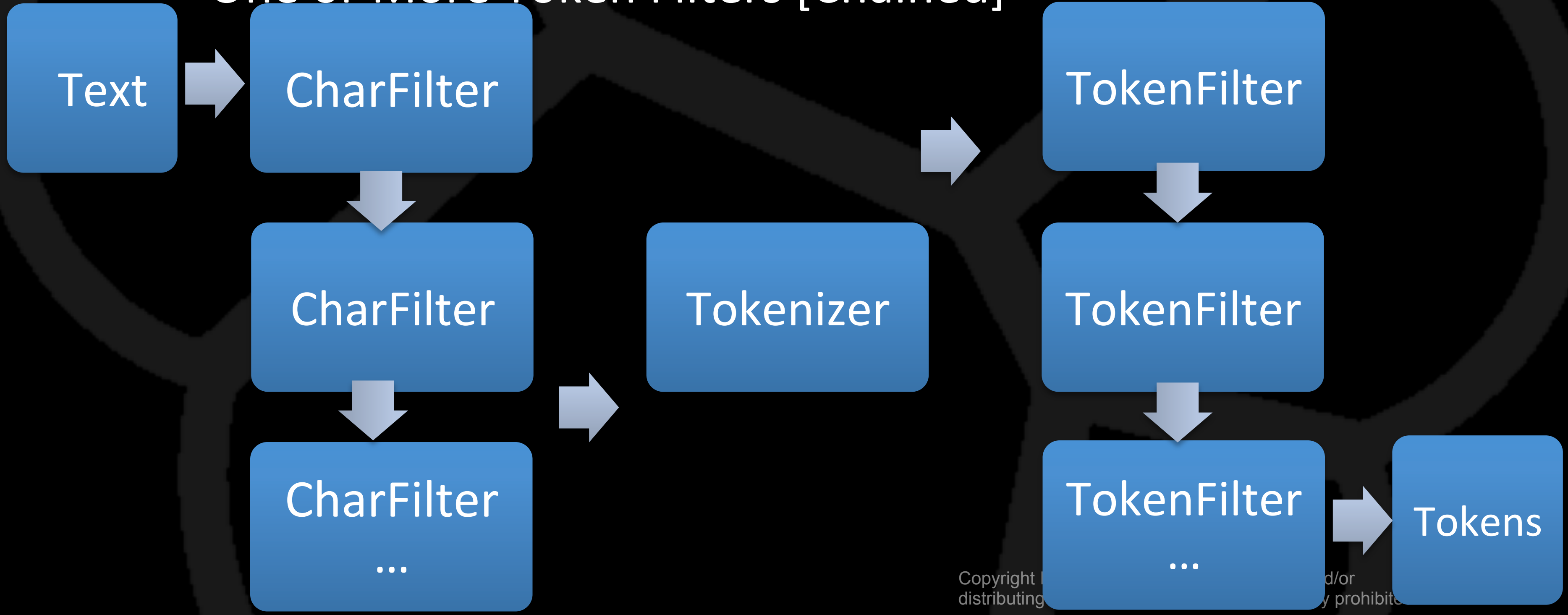
- Analyzer影响索引
- Analyzer影响查询

How Analyzer Works

ANALYZER如何工作

How Analyzer works?

- Analyzer build with:
  - One or More Char Filters [Chained]
  - One Tokenizer
  - One or More Token Filters [Chained]



Text

CharFilter

CharFilter

CharFilter
…

Tokenizer

TokenFilter

TokenFilter

TokenFilter
…

Tokens

How Analyzer works?

- **Elastic search** IS REALLY Amazing!
- **Elasticsearch** IS REALLY Amazing!
- [Elasticsearch] [IS] [REALLY] [Amazing] [!]
- [elasticsearch] [**is**] [**really**] [**amazing**] [**!**]
- [elasticsearch] [really] [amazing]
- [elasticsearch] [really] [**amaze**]
- [elasticsearch] [really] [**indeed**] [amaze]

Built-In Analyzers

Analzying "The quick brown fox jumped over the lazy dogs”

- org.apache.lucene.analysis.WhitespaceAnalyzer:
  [The] [quick] [brown] [fox] [jumped] [over] [the] [lazy] [dogs]

- org.apache.lucene.analysis.SimpleAnalyzer:
  [the] [quick] [brown] [fox] [jumped] [over] [the] [lazy] [dogs]

- org.apache.lucene.analysis.StopAnalyzer:
  [quick] [brown] [fox] [jumped] [over] [lazy] [dogs]

- org.apache.lucene.analysis.standard.StandardAnalyzer:
  [quick] [brown] [fox] [jumped] [over] [lazy] [dogs]

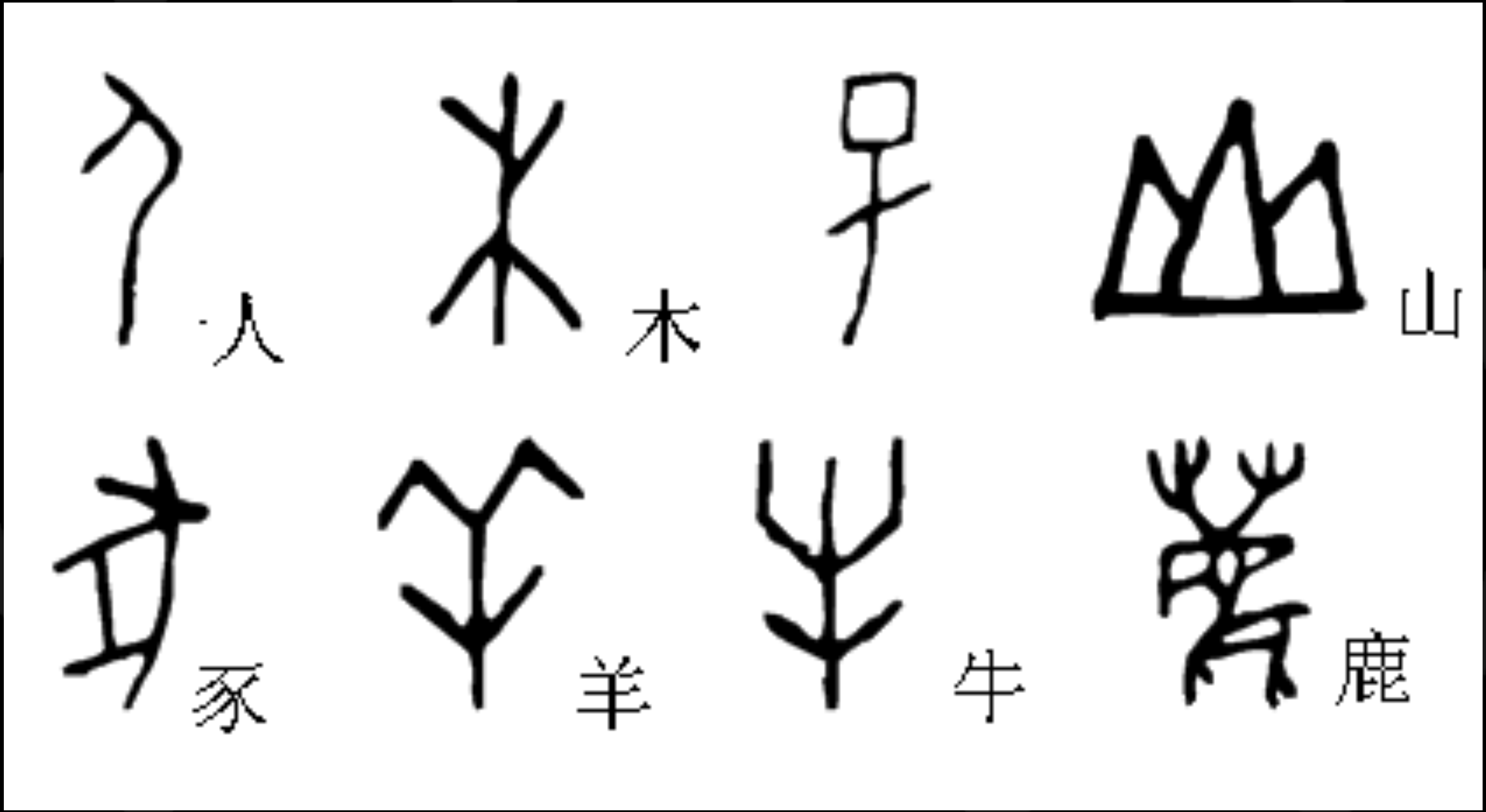
- org.apache.lucene.analysis.snowball.SnowballAnalyzer:
  [quick] [brown] [fox] [jump] [over] [lazi] [dog]

Lucene has many built-in analyzers

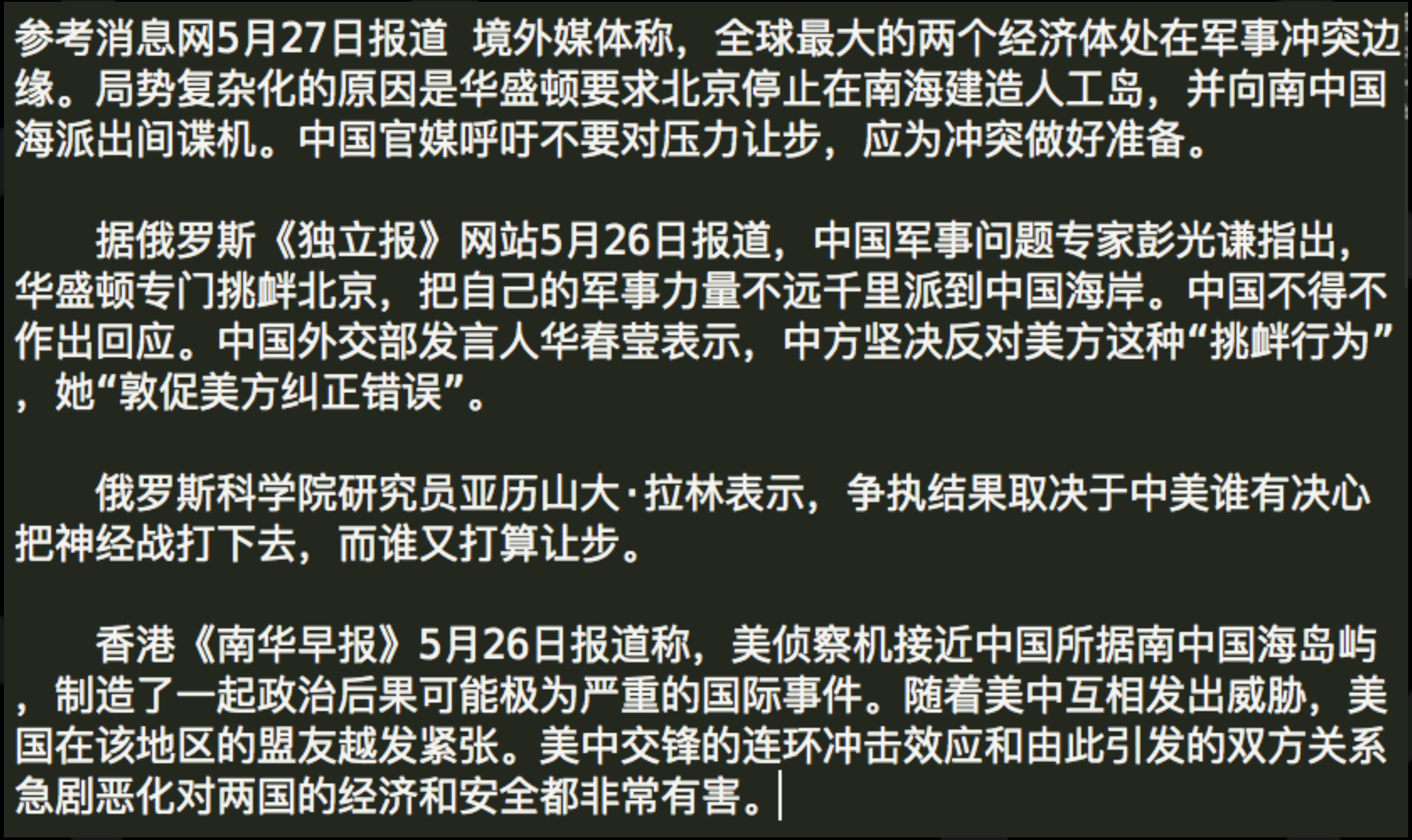
Analyzer for Chinese

中文分析处理

中文博大精深



一则新闻



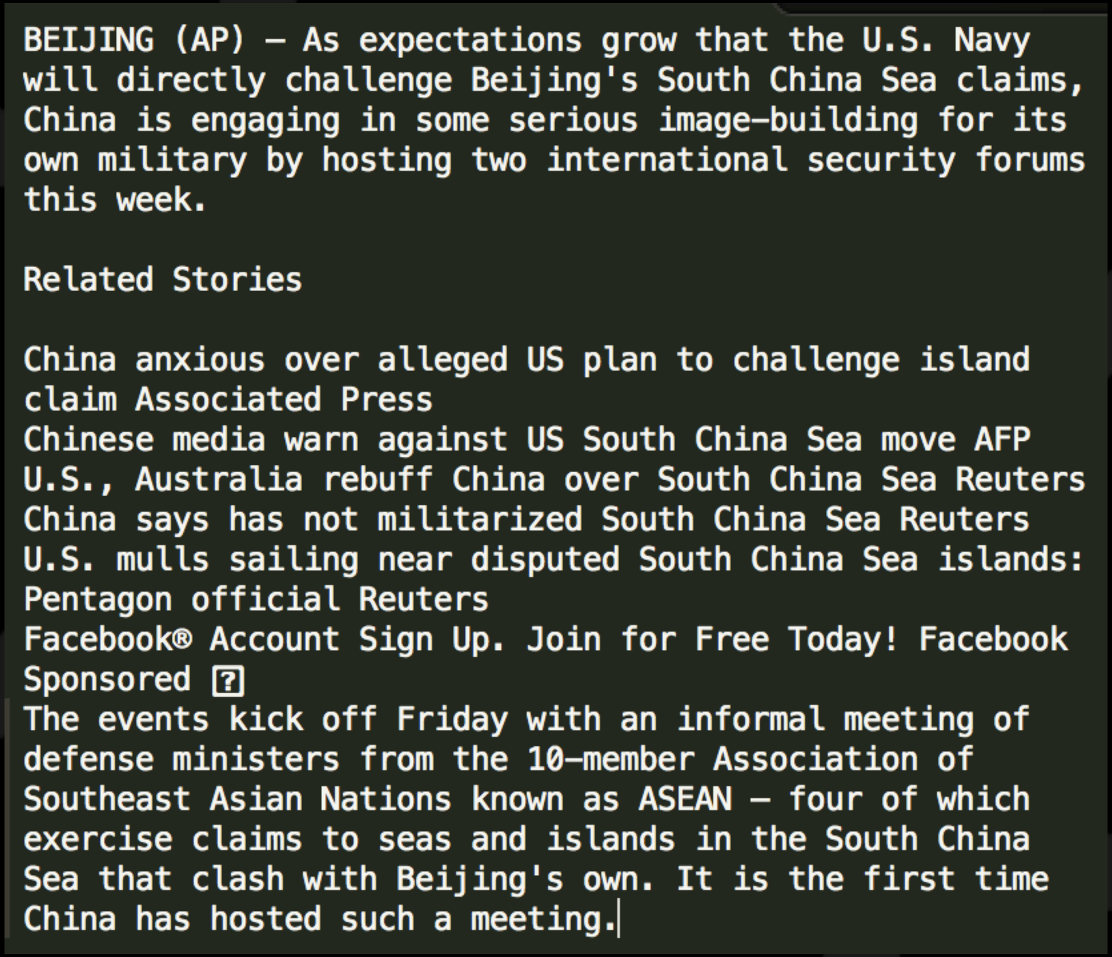
参考消息网5月27日报道 境外媒体称，全球最大的两个经济体处在军事冲突边缘。局势复杂化的原因是华盛顿要求北京停止在南海建造人工岛，并向南中国海派出间谍机。中国官媒呼吁不要对压力让步，应为冲突做好准备。

据俄罗斯《独立报》网站5月26日报道，中国军事问题专家彭光谦指出，华盛顿专门挑衅北京，把自己的军事力量不远千里派到中国海岸。中国不得不作出回应。中国外交部发言人华春莹表示，中方坚决反对美方这种“挑衅行为”，她“敦促美方纠正错误”。

俄罗斯科学院研究员亚历山大·拉林表示，争执结果取决于中美谁有决心把神经战打下去，而谁又打算让步。

香港《南华早报》5月26日报道称，美侦察机接近中国所据南中国海岛屿，制造了一起政治后果可能极为严重的国际事件。随着美中互相发出威胁，美国在该地区的盟友越发紧张。美中交锋的连环冲击效应和由此引发的双方关系急剧恶化对两国的经济和安全都非常有害。

再看看英文

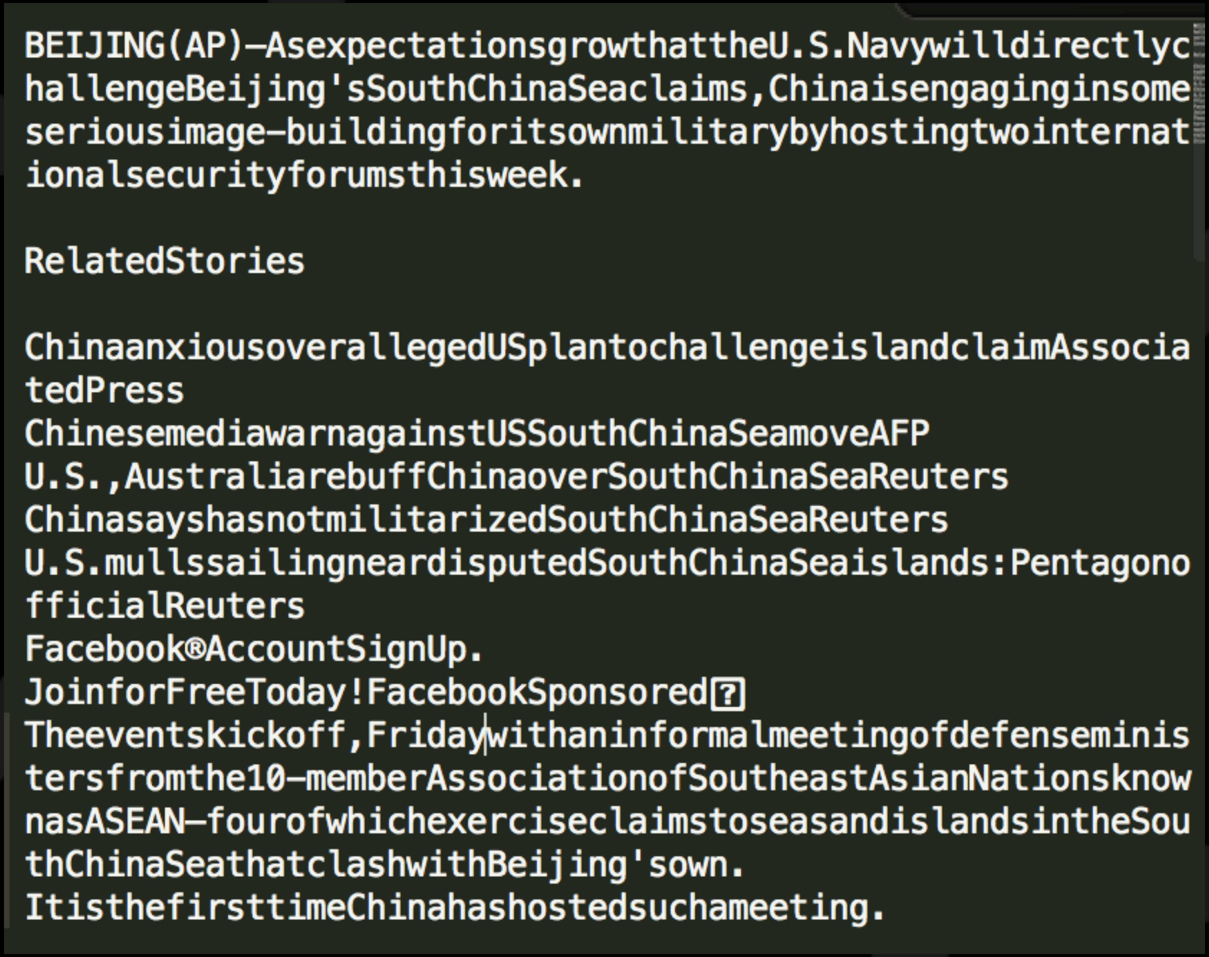


BEIJING (AP) — As expectations grow that the U.S. Navy will directly challenge Beijing's South China Sea claims, China is engaging in some serious image-building for its own military by hosting two international security forums this week.

Related Stories

China anxious over alleged US plan to challenge island claim Associated Press
Chinese media warn against US South China Sea move AFP
U.S., Australia rebuff China over South China Sea Reuters
China says has not militarized South China Sea Reuters
U.S. mulls sailing near disputed South China Sea islands: Pentagon official Reuters
Facebook® Account Sign Up. Join for Free Today! Facebook
Sponsored ?
The events kick off Friday with an informal meeting of defense ministers from the 10-member Association of Southeast Asian Nations known as ASEAN — four of which exercise claims to seas and islands in the South China Sea that clash with Beijing's own. It is the first time China has hosted such a meeting.

换个姿势



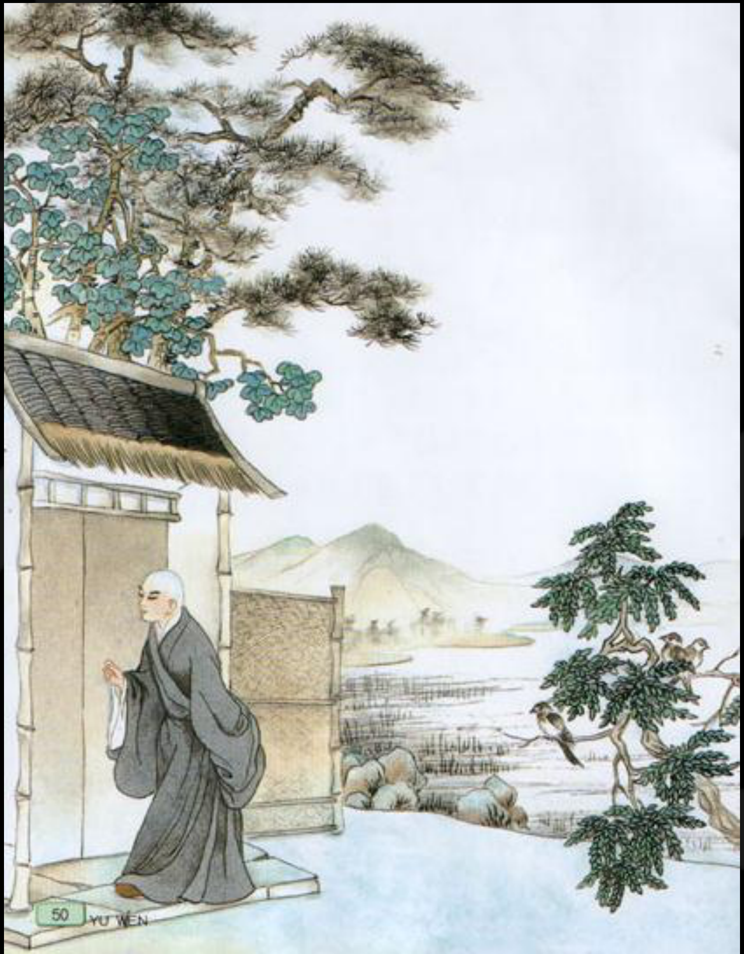
```
BEIJING(AP)—AsexpectationsgrowthattheU.S.Navywilldirectlyc
hallengeBeijing'sSouthChinaSeaclaims,Chinaisengaginginsome
seriousimage-buildingforitsownmilitarybyhostingtwointernat
ionalsecurityforumsthisweek.

RelatedStories

ChinaanxiousoverallegedUSplantochallengeislandclaimAssocia
tedPress
ChinesemediawarnagainstUSSouthChinaSeamoveAFP
U.S.,AustraliarebuffChinaoverSouthChinaSeaReuters
ChinasayshasnotmilitarizedSouthChinaSeaReuters
U.S.mullssailingneardisputedSouthChinaSeaislands:Pentagono
fficialReuters
Facebook®AccountSignUp.
JoinforFreeToday!FacebookSponsored⍰
Theeventskickoff,Fridaywithaninformalmeetingofdefensemin is
tersfromthe10-memberAssociationofSoutheastAsianNationsknow
nasASEAN—fourofwhichexerciseclaimstoseasandislandsintheSou
thChinaSeathatclashwithBeijing'sown.
ItisthefirsttimeChinahashostedsuchameeting.
```

中文很复杂

- 同音、多音、多义、兼类词和同形异构… …

1. 还欠款壹万元
2. 放弃美丽的女人让人心碎
3. 开刀的是他父亲

“推” “敲”



Community Analyzers

- ICTCLAS
- ANSJ
- CC-CEDICT
- SCWS
- FudanNLP
- IK
- MMSEG
- JIEBA
- …

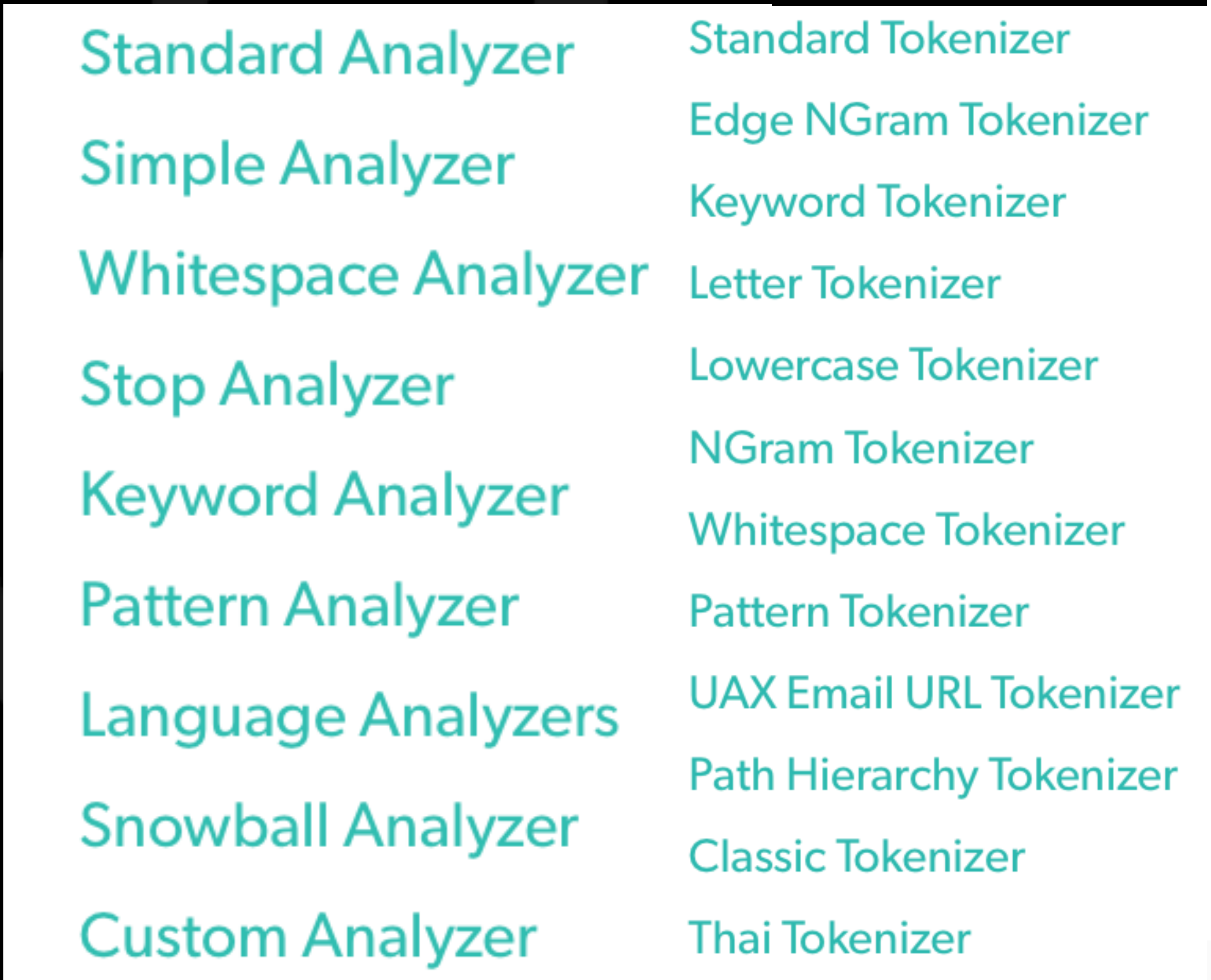
More

- 繁体
- 拼音

Analyzer In Elasticsearch

ELASTICSEARCH中的ANALYZER

内置丰富



Standard Analyzer

Simple Analyzer

Whitespace Analyzer

Stop Analyzer

Keyword Analyzer

Pattern Analyzer

Language Analyzers

Snowball Analyzer

Custom Analyzer

Standard Tokenizer

Edge NGram Tokenizer

Keyword Tokenizer

Letter Tokenizer

Lowercase Tokenizer

NGram Tokenizer

Whitespace Tokenizer

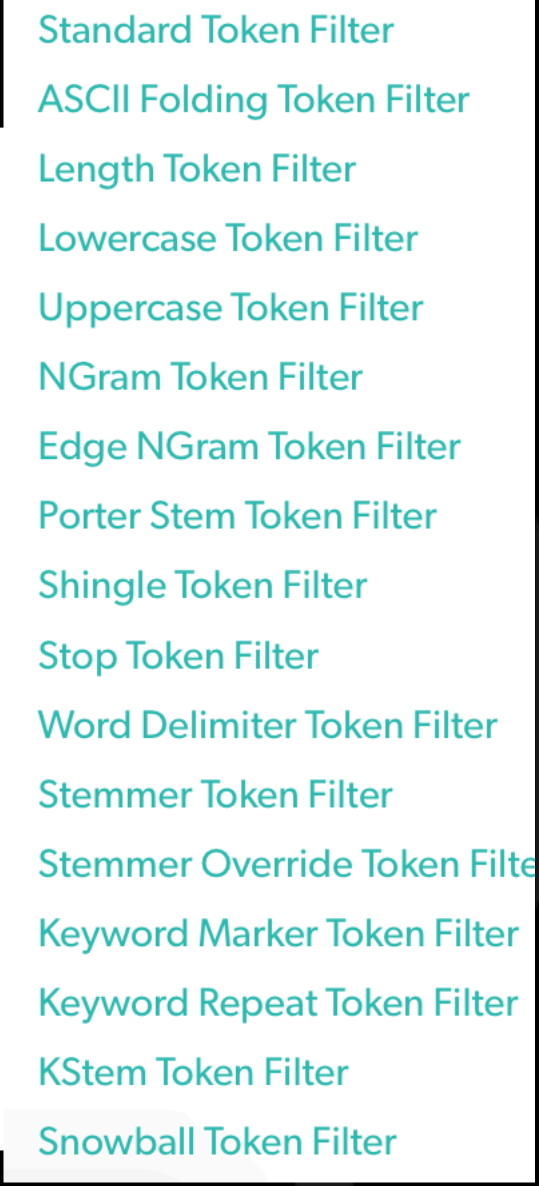
Pattern Tokenizer

UAX Email URL Tokenizer

Path Hierarchy Tokenizer

Classic Tokenizer

Thai Tokenizer



Standard Token Filter

ASCII Folding Token Filter

Length Token Filter

Lowercase Token Filter

Uppercase Token Filter

NGram Token Filter

Edge NGram Token Filter

Porter Stem Token Filter

Shingle Token Filter

Stop Token Filter

Word Delimiter Token Filter

Stemmer Token Filter

Stemmer Override Token Filte

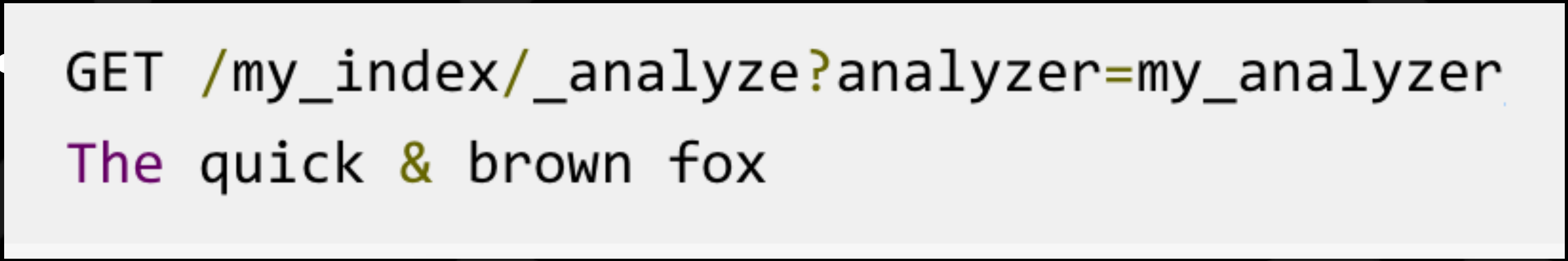
Keyword Marker Token Filter

Keyword Repeat Token Filter

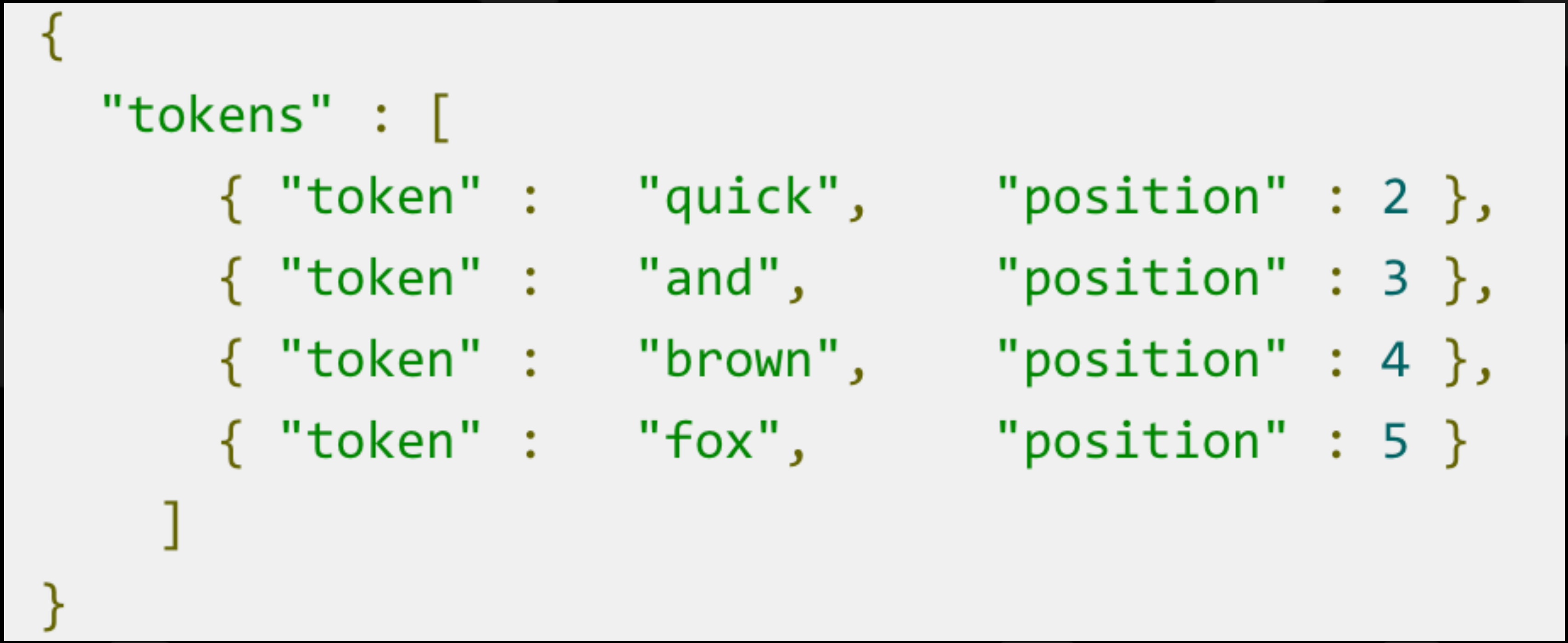
KStem Token Filter

Snowball Token Filter

调试Analyzer

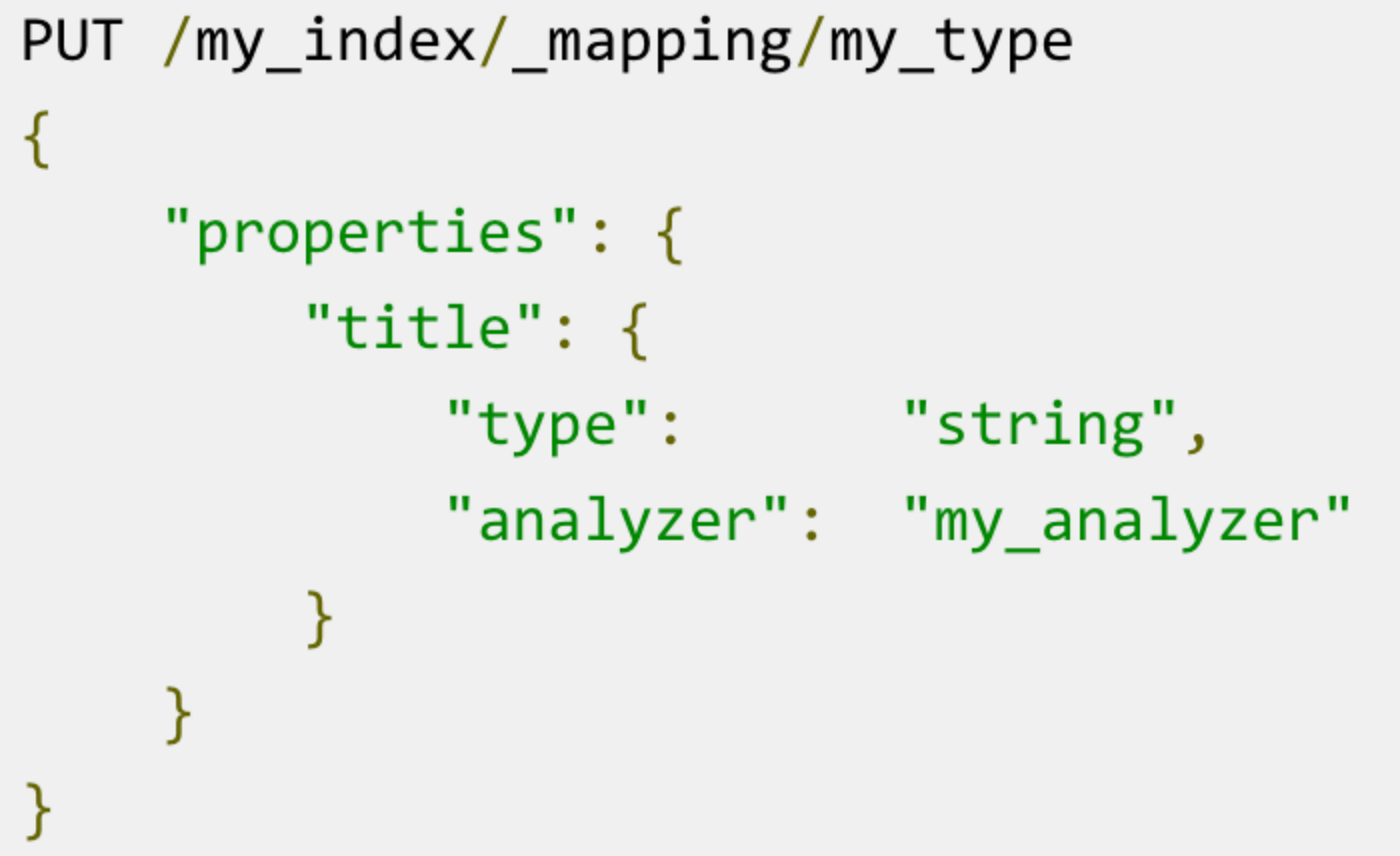


```
GET /my_index/_analyze?analyzer=my_analyzer
The quick & brown fox
```



```
{
  "tokens" : [
      { "token" :   "quick",     "position" : 2 },
      { "token" :   "and",       "position" : 3 },
      { "token" :   "brown",     "position" : 4 },
      { "token" :   "fox",       "position" : 5 }
    ]
}
```

Mapping

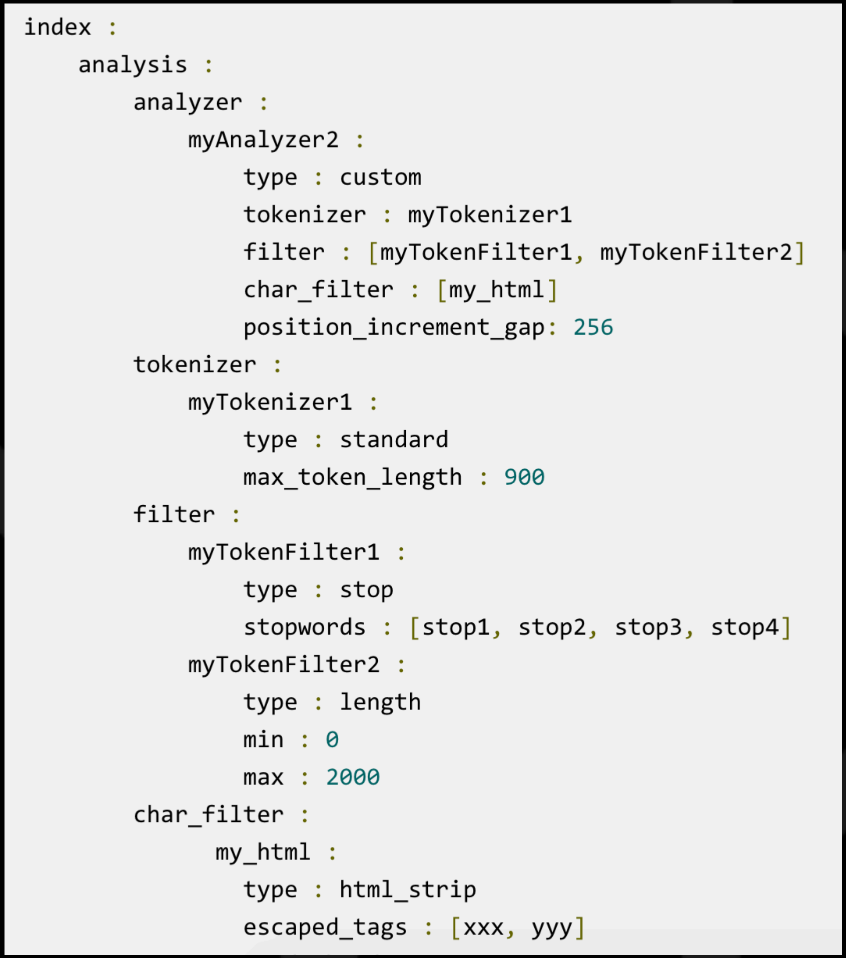


```
PUT /my_index/_mapping/my_type
{
    "properties": {
        "title": {
            "type":       "string",
            "analyzer":   "my_analyzer"
        }
    }
}
```

自定义Analyzer

- elasticsearch.yml

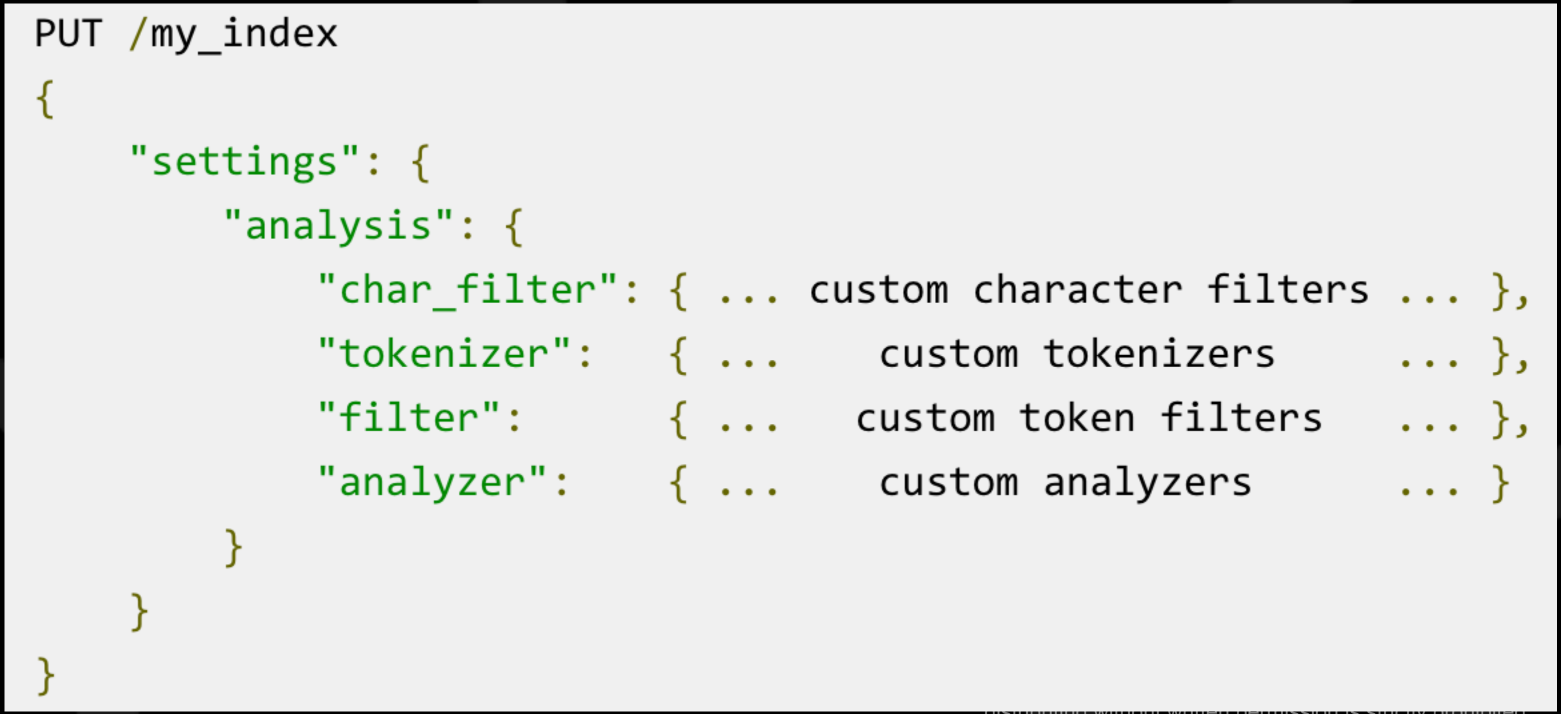
- 配置全局可见
- 修改需要重启集群



```
index :
    analysis :
        analyzer :
            myAnalyzer2 :
                type : custom
                tokenizer : myTokenizer1
                filter : [myTokenFilter1, myTokenFilter2]
                char_filter : [my_html]
                position_increment_gap: 256
        tokenizer :
            myTokenizer1 :
                type : standard
                max_token_length : 900
        filter :
            myTokenFilter1 :
                type : stop
                stopwords : [stop1, stop2, stop3, stop4]
            myTokenFilter2 :
                type : length
                min : 0
                max : 2000
        char_filter :
              my_html :
                type : html_strip
                escaped_tags : [xxx, yyy]
```

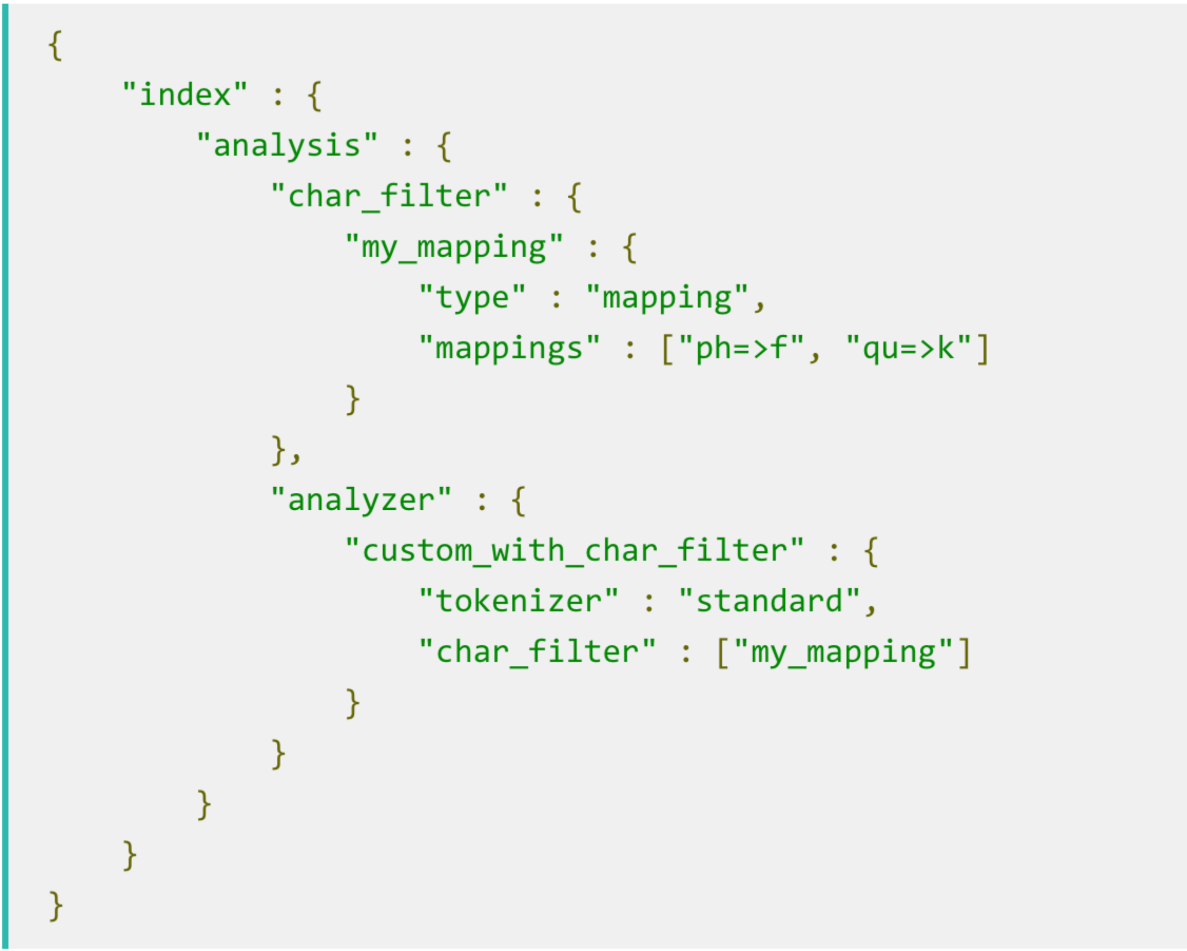
动态组合Analyzer

- 1.关闭Index
- 2.修改IndexSettings,创建Analyzer
- 3.打开Index



```
PUT /my_index
{
    "settings": {
        "analysis": {
            "char_filter": { ... custom character filters ... },
            "tokenizer":   { ...    custom tokenizers      ... },
            "filter":      { ...   custom token filters    ... },
            "analyzer":    { ...    custom analyzers       ... }
        }
    }
}
```

配置 Char-Filter



```
{
    "index" : {
        "analysis" : {
            "char_filter" : {
                "my_mapping" : {
                    "type" : "mapping",
                    "mappings" : ["ph=>f", "qu=>k"]
                }
            },
            "analyzer" : {
                "custom_with_char_filter" : {
                    "tokenizer" : "standard",
                    "char_filter" : ["my_mapping"]
                }
            }
        }
    }
}
```

How to Choose the Analyzer

如何选择ANALYZER

关于停用词

- “To be or not to be”

关于排序

- 字段能分词么？

Prefix/Wildcard Query& Analyzers

- 字段能分词么?
- 和Ngram的比较

Aggregation & Analyzers

- 字段能分词么?

Suggest

- 一个Analyzer搞定所有场景?
- 试试Multi-Field
- 试试多字段查询

Final words

Search well,

choose right analyzer!

Recommendation

- **Apache Lucene Documentation**

  http://lucene.apache.org/core/5_3_1/index.html

- Lucene In Action 2

Thank You!