Pages / ... / elasticsearch

5-es分词

Created by 杨超, last modified on 2018 Nov 15

常用分词

列举出目前es中的几种常用分词

分词器	作用	使用场景
standard 也就是默认分词器	主要是英文分词，中文分词就是逐字分词，基本没用	纯英文?
ik分词 - ik_max_word & ik_smart	中文分词 ik_max_word 会按照最小维度将所有的可能分词都分出来，ik_smart就是智能分一次，有点听天由命的味道	中文分词的不二选择
pinyin	拼音分词 就是将中文转成拼音	有特殊需求的时候，目前场景不多
keyword	不分词	模糊搜索的时候 实现类似 like功能时候，对性能损耗较大

示例

1 - standard

```
GET /_analyze
{
  "text": "中华人民共和国"
}
```

```
{
  "tokens": [
    {
      "token": "中",
      "start_offset": 0,
      "end_offset": 1,
      "type": "<IDEOGRAPHIC>",
      "position": 0
    },
    {
      "token": "华",
      "start_offset": 1,
      "end_offset": 2,
      "type": "<IDEOGRAPHIC>",
      "position": 1
    },
    {
      "token": "人",
      "start_offset": 2,
      "end_offset": 3,
      "type": "<IDEOGRAPHIC>",
      "position": 2
    },
    {
      "token": "民",
      "start_offset": 3,
      "end_offset": 4,
      "type": "<IDEOGRAPHIC>",
      "position": 3
    },
    {
      "token": "共",
      "start_offset": 4,
      "end_offset": 5,
      "type": "<IDEOGRAPHIC>",
      "position": 4
    },
    {
      "token": "和",
      "start_offset": 5,
      "end_offset": 6,
      "type": "<IDEOGRAPHIC>",
      "position": 5
    },
    {
      "token": "国",
      "start_offset": 6,
      "end_offset": 7,
      "type": "<IDEOGRAPHIC>",
      "position": 6
    }
  ]
}
```

2 - ik_max_word

```
GET /_analyze
{
  "analyzer": "ik_max_word",
  "text": "中华人民共和国"
}
```

```
{
  "tokens": [
    {
      "token": "中华人民共和国",
      "start_offset": 0,
      "end_offset": 7,
      "type": "CN_WORD",
      "position": 0
    },
    {
      "token": "中华人民",
      "start_offset": 0,
      "end_offset": 4,
      "type": "CN_WORD",
      "position": 1
    },
    {
      "token": "中华",
      "start_offset": 0,
      "end_offset": 2,
      "type": "CN_WORD",
      "position": 2
    },
    {
      "token": "华人",
      "start_offset": 1,
      "end_offset": 3,
      "type": "CN_WORD",
      "position": 3
    },
    {
      "token": "人民共和国",
      "start_offset": 2,
      "end_offset": 7,
      "type": "CN_WORD",
      "position": 4
    },
    {
      "token": "人民",
      "start_offset": 2,
      "end_offset": 4,
      "type": "CN_WORD",
      "position": 5
    },
    {
      "token": "共和国",
      "start_offset": 4,
      "end_offset": 7,
      "type": "CN_WORD",
      "position": 6
    },
    {
      "token": "共和",
      "start_offset": 4,
      "end_offset": 6,
      "type": "CN_WORD",
      "position": 7
    },
    {
      "token": "国",
      "start_offset": 6,
      "end_offset": 7,
      "type": "CN_CHAR",
      "position": 8
    }
  ]
}
```

3 pinyin

```
POST /_analyze
```

```
{
    "analyzer":"pinyin",
    "text":"中华人民共和国"
}
```

```
{
  "tokens": [
    {
      "token": "zhong",
      "start_offset": 0,
      "end_offset": 1,
      "type": "word",
      "position": 0
    },
    {
      "token": "hua",
      "start_offset": 1,
      "end_offset": 2,
      "type": "word",
      "position": 1
    },
    {
      "token": "ren",
      "start_offset": 2,
      "end_offset": 3,
      "type": "word",
      "position": 2
    },
    {
      "token": "min",
      "start_offset": 3,
      "end_offset": 4,
      "type": "word",
      "position": 3
    },
    {
      "token": "gong",
      "start_offset": 4,
      "end_offset": 5,
      "type": "word",
      "position": 4
    },
    {
      "token": "he",
      "start_offset": 5,
      "end_offset": 6,
      "type": "word",
      "position": 5
    },
    {
      "token": "guo",
      "start_offset": 6,
      "end_offset": 7,
      "type": "word",
      "position": 6
    },
    {
      "token": "zhrmghg",
      "start_offset": 0,
      "end_offset": 7,
      "type": "word",
      "position": 6
    }
  ]
}
```

4 keyword

```
POST /_analyze
{
    "analyzer":"keyword",
    "text":"中华人民共和国"
}
```

```
{
  "tokens": [
    {
      "token": "中华人民共和国",
      "start_offset": 0,
      "end_offset": 7,
      "type": "word",
      "position": 0
    }
  ]
}
```

Like 2 people like this

No labels

1 Comment

纪浩

求调研score权重这块啊