IK分词器安装

1、环境准备

Elasticsearch 要使用 ik，就要先构建 ik 的 jar包，这里要用到 maven 包管理工具，而 maven 需要java 环境，而 Elasticsearch 内置了jdk， 所以可以将JAVA_HOME设置为Elasticsearch 内置的jdk

1）设置JAVA_HOME(这个可以省略)

```
vim /etc/profile
# 在profile文件末尾添加
#java environment
export JAVA_HOME=/opt/elasticsearch-7.4.0/jdk
export PATH=$PATH:${JAVA_HOME}/bin

# 保存退出后，重新加载profile
source /etc/profile
```

2、安装IK分词器

1）下载IK

```
https://github.com/medcl/elasticsearch-analysis-ik/archive/v7.4.0.zip
```

执行如下图：

2）解压IK

```
1.创建文件夹，在elasticSearch中的：mkdir elastic/plugin/ik
2.将文件移入到：elasticSeach/plugins/ik
```

由于这里是zip包不是gz包，所以我们需要使用unzip命令进行解压，如果本机环境没有安装unzip，请执行：

```
yum install zip
yum install unzip
```

解压IK

```
unzip v7.4.0.zip
```

3、使用IK分词器

IK分词器有两种分词模式：ik_max_word和ik_smart模式。

1、**ik_max_word**

会将文本做最细粒度的拆分，比如会将“乒乓球明年总冠军”拆分为“乒乓球、乒乓、球、明年、总冠军、冠军。

```
#方式一ik_max_word
GET /_analyze
{
  "analyzer": "ik_max_word",
  "text": "乒乓球明年总冠军"
}
```

ik_max_word分词器执行如下：

```
{
  "tokens" : [
    {
      "token" : "乒乓球",
      "start_offset" : 0,
      "end_offset" : 3,
      "type" : "CN_WORD",
      "position" : 0
    },
    {
      "token" : "乒乓",
      "start_offset" : 0,
      "end_offset" : 2,
      "type" : "CN_WORD",
      "position" : 1
    },
    {
      "token" : "球",
      "start_offset" : 2,
      "end_offset" : 3,
      "type" : "CN_CHAR",
      "position" : 2
    },
    {
      "token" : "明年",
      "start_offset" : 3,
      "end_offset" : 5,
      "type" : "CN_WORD",
      "position" : 3
    },
    {
      "token" : "总冠军",
```

```
      "start_offset" : 5,
      "end_offset" : 8,
      "type" : "CN_WORD",
      "position" : 4
    },
    {
      "token" : "冠军",
      "start_offset" : 6,
      "end_offset" : 8,
      "type" : "CN_WORD",
      "position" : 5
    }
  ]
}
```

2、**ik_smart**

会做最粗粒度的拆分，比如会将“乒乓球明年总冠军”拆分为乒乓球、明年、总冠军。

```
#方式二ik_smart
GET /_analyze
{
  "analyzer": "ik_smart",
  "text": "乒乓球明年总冠军"
}
```

ik_smart分词器执行如下：

```
{
  "tokens" : [
    {
      "token" : "乒乓球",
      "start_offset" : 0,
      "end_offset" : 3,
      "type" : "CN_WORD",
      "position" : 0
    },
    {
      "token" : "明年",
      "start_offset" : 3,
      "end_offset" : 5,
      "type" : "CN_WORD",
      "position" : 1
    },
    {
      "token" : "总冠军",
      "start_offset" : 5,
      "end_offset" : 8,
```

```
      "type" : "CN_WORD",
      "position" : 2
    }
  ]
}
```

由此可见 使用ik_smart可以将文本"text": "乒乓球明年总冠军"分成了【乒乓球】【明年】【总冠军】

这样看的话，这样的分词效果达到了我们的要求。