# IK分词器的拓展和停用词典

## IK分词器原理

刚刚已经安装了IK分词器，并且可以看到，使用IK分词器能成功地对中文进行分词，而且分出来的词还挺好的，大家可以思考以下，IK分词器底层分词的原理是什么呢？比如IK分词器怎么知道”你好”这两个字何在一起就是一个词，“程序员”合在一起就是一个词，我觉得在IK分词器的底层一定有一个字典：

IK分词器分词的原理是什么？

在底层一定有一个字典，在字典里一定会有各种各样的词语，提前罗列好了，当要分词的时候，一定会把要分词的内容拿去字典匹配，看字典里有没有，如果匹配到在词典里有，那么说明是一个词，那就分词出来，你想一下是不是这个逻辑，一字那么浅！所以任何中文分词器，都会依赖于一个字典来去做分词。

字典是不会包含全世界所有的词，甚至是中文词也不可能全部包含，因为一是太多，二是词一直在发展变化，不停有很多新词出来，譬如网络用语“你XX”、“潘噶之交”、“白嫖”、“老六”、“鸡你太美”。

那么问题来了，我们要怎么去拓展分词器的字典，比如把“你XX”、“老六”、“鸡你太美”这些词加到字典中；另外对一些无意义的词可以不去分词，即停用，譬如：“的”、“呢”。。。。，因为分词了会占用内存；还有一些敏感词汇、禁忌词，这些也应该禁掉。

那么问题来了，我们的分词器能不能实现字典的个性化设置呢？当然可以！

## ik分词器-扩展词库

IK分词器支持词库的扩展，拓展方式只需要找到ik分词器目录中的config目录下的IKAnalyzer.cfg.xml文件，并对这个文件进行修改，目录如下：

**/var/lib/docker/volumes/es-plugins/\_data/ik/config**

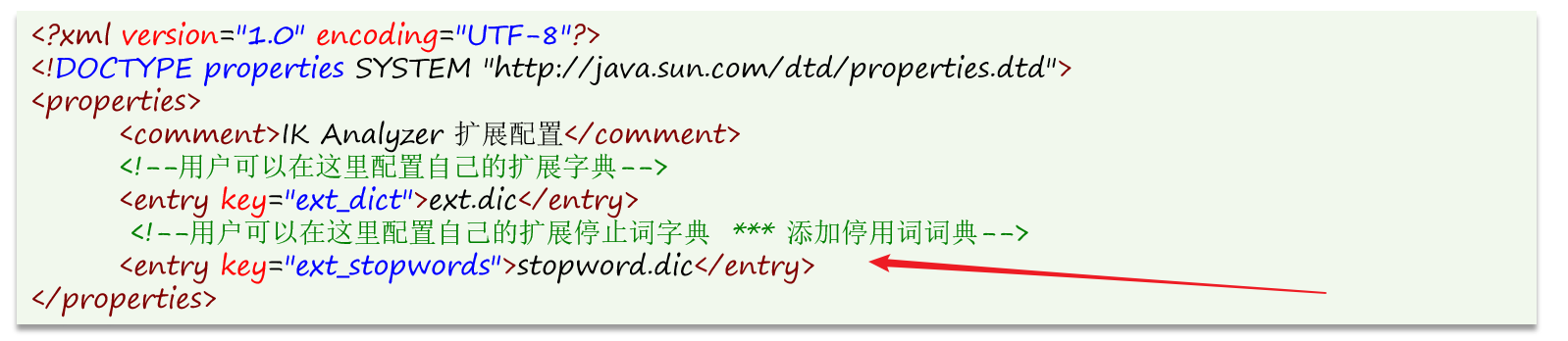


然后再这个ext.dic文件中添加想要拓展的词条！

## IK分词器-停用词库

IK分词器支持词库敏感词的停用，要禁用某些敏感词条，只需要找到ik分词器目录中的config目录下的IKAnalyzer.cfg.xml文件，并对这个文件进行修改，目录如下：

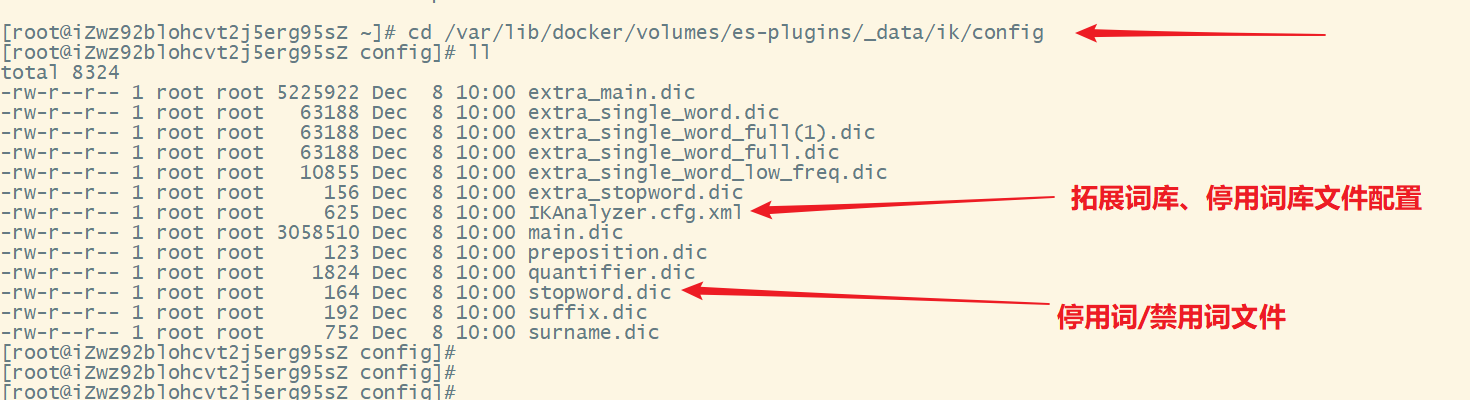
**/var/lib/docker/volumes/es-plugins/\_data/ik/config**



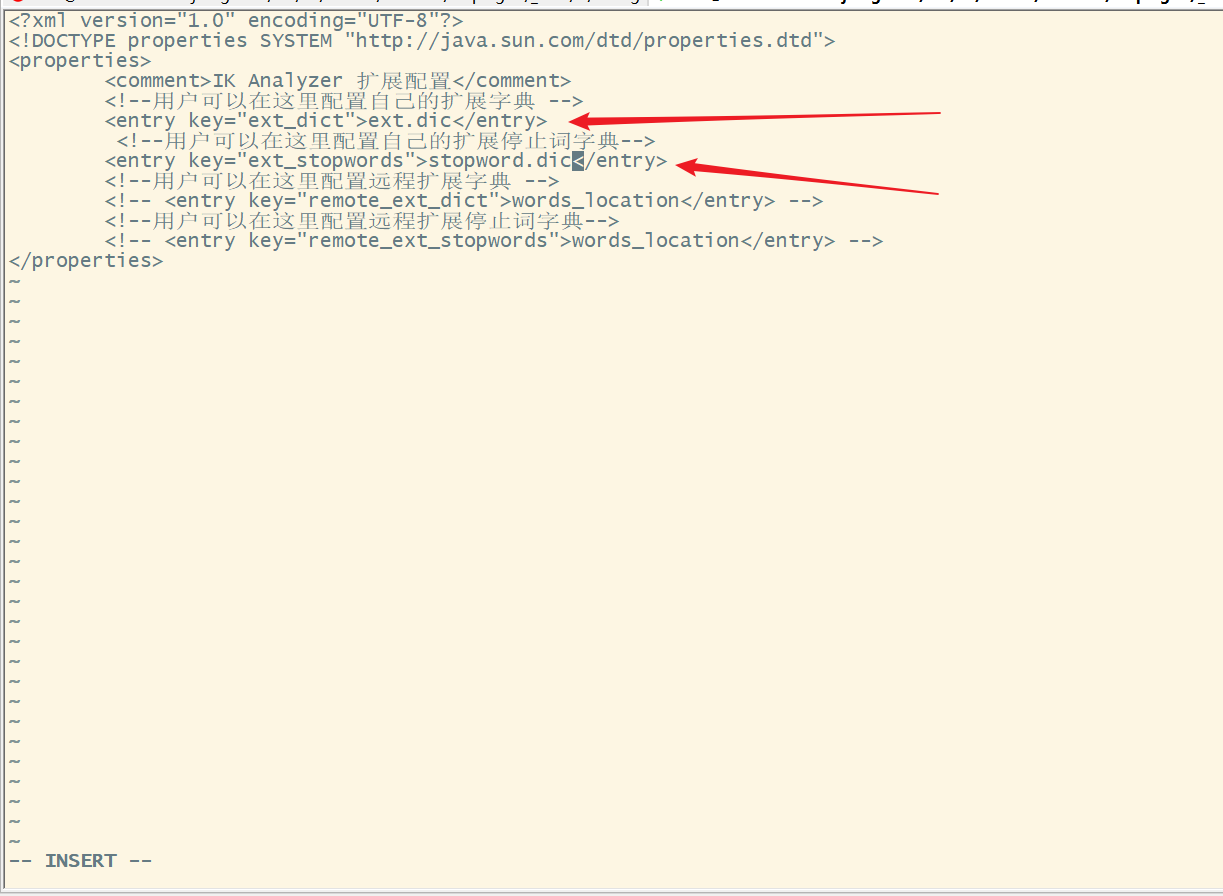
然后再这个stopword.dic文件中添加想要禁用、停用的词条！

## 1：进入目录ik分词器插件配置文件目录

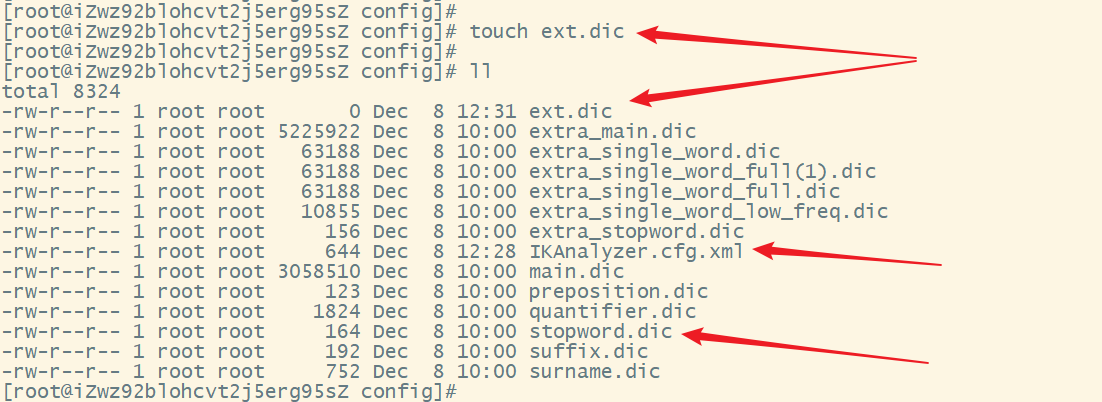
cd /var/lib/docker/volumes/es-plugins/\_data/ik/config



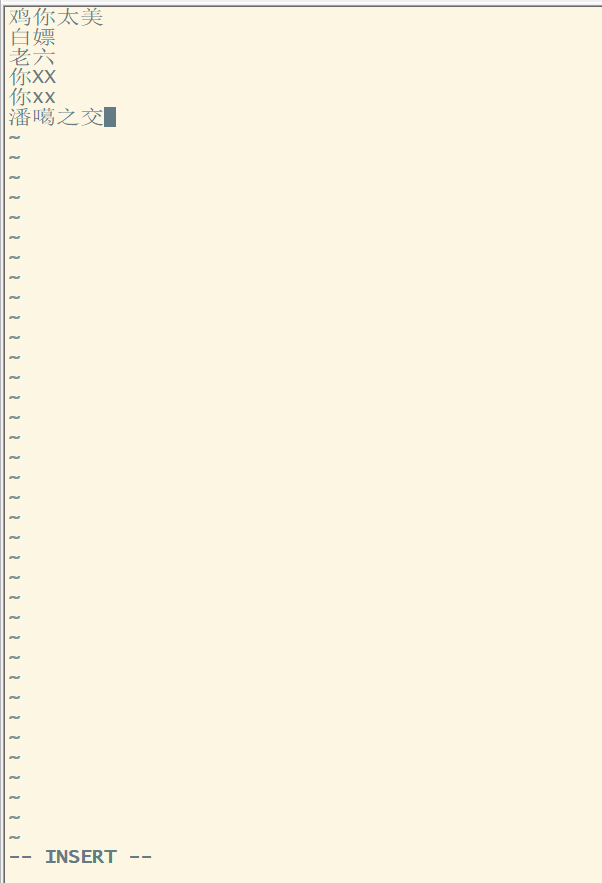
## 2：编辑文件IKAnalyzer.cfg.xml



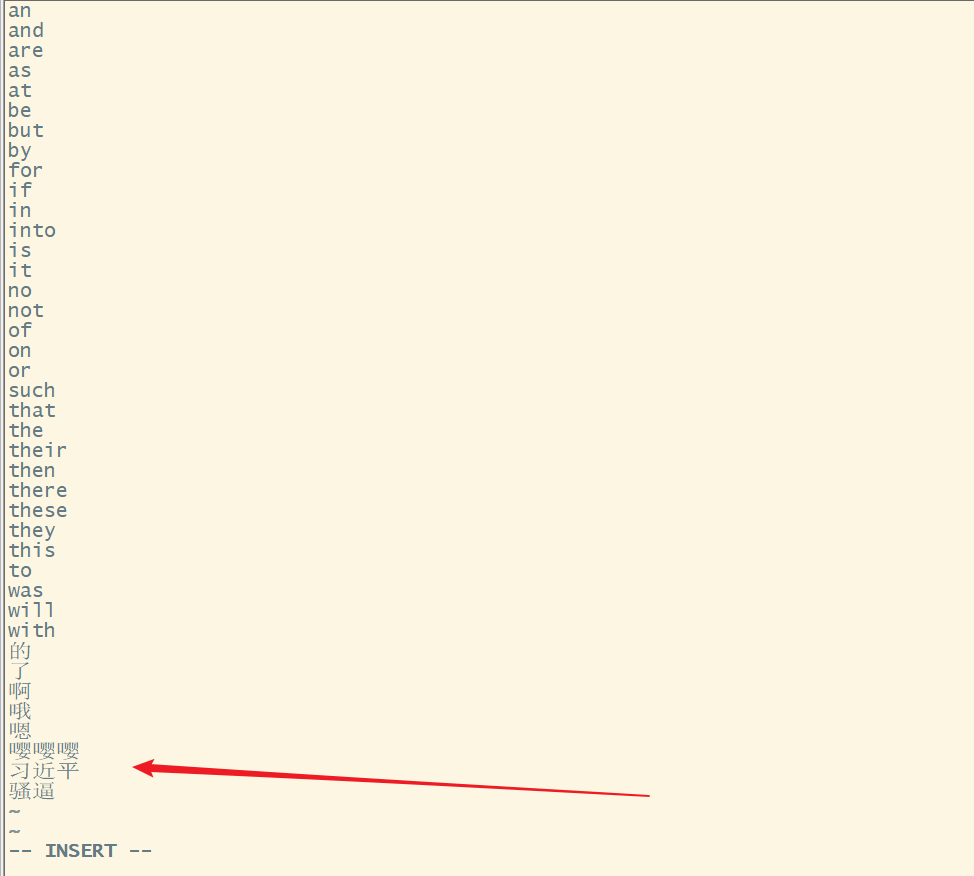
## 3：在IKAnalyzer.cfg.xml所在目录中，stopword.dic已经存在，不用创建；ext.dic不存在，要创建



## 4：在ext.dic中加入想拓展的词语 vi ext.dic



## 5：在stopword.dic中加入想停用的词语 vi stopword.dic



## 6：重启es，让配置生效



## 7：DevTool控制台测试一下：

POST /\_analyze

{

"analyzer": "ik\_max\_word",

"text": "你XX练习时长两年半的鸡你太美的习近平，老六你好想白嫖骚逼啊"

}

POST /\_analyze

{

"analyzer": "ik\_smart",

"text": "你XX练习时长两年半的鸡你太美的习近平，老六你好想白嫖骚逼啊"

}

测试可以看到，确实是生效了，不信你试试看嘛！

## 注意点

拓展词库是对字典扩容，禁用词库是对字典删除

**不管是对文档的字段生产倒排索引时的分词，还是对用户输入的搜索内容分词，分词是用分词器来分的，分词器是基与字典来的分词的！文档的字段或者搜索内容匹配到在字典中有的就分词，字典中没有或者字典中禁用的就部分词！**

## 总结：

### 分词器的作用是什么？

#### 两个场景需要用分词

1是将文档创建倒排索引的时候，要对文档的某些内容进行分词，得到一个个词条，然后对词条创建倒排索引；

2是当用户进行搜索的时候，要对输入的搜索内容进行分词。

所以分词器的作用主要有两个

#### 1：创建倒排索引时对文档分词；

#### 2：用户搜索时，对输入的内容分词。

### IK分词器有几种模式？

1：Ik\_smart：智能切分，粗粒度

2：Ik\_max\_word：最细切分，细粒度

### IK分词器如何拓展词条？如何停用词条？

1：利用config目录中的IKAnalyzer,cfg.xml文件添加拓展词典文件和停用词典文件

2：在词典中添加拓展词条或停用词条