语料库检索技术

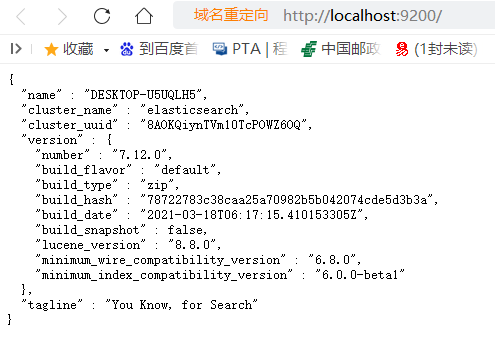
1. **Elasticsearch安装与连接Python**

（1）、第一步，我们需要去官网下载Elasticsearch，解压后运行.\elasticsearch-7.12.0（版本号）\bin\elasticsearch.bat这一文件。在之后的处理过程中需要一直保持这个程序打开。

（2）、在保持打开的状态下，我们查看本地9200端口，如果显示被占用，表示成功开启这项功能。



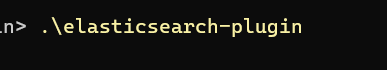
同时访问<http://localhost:9200/>。浏览器输出信息如下：



（3）、使用pip命令安装elasticsearch包，并且初始化操作对象es。



（4）、安装中文分词插件。在.\elasticsearch-7.12.0（版本号）\bin\目录下运行elasticsearch-plugin（注意，命令前需要加.\），如下图所示：



然后键入命令.\elasticsearch-plugin install https://github.com/medcl/elasticsearch-analysis-ik/releases/download/v7.12.0/elasticsearch-analysis-ik-7.12.0.zip



由于elasticsearch可以在各操作系统下运行，因此网上的教程中不同命令行的语法有略微的不同之处，按照提示即可换为正确的语法。

1. **创建索引**

创建两类标识，一类是新闻，一类是数学：

*es.indices.create(index='news', ignore=400)*

*es.indices.create(index='math', ignore=402)*

1. **导入**

按照语法导入自己编的少量新闻

es.index(index=**'news'**, doc\_type=**'science'**, body=data)

1. **查询结果与分析**





部分语料截图。

（1）以下以index=’news’为例；

1. 当输入“芯片”时，返回结果如下：



第一条新闻的得分为2.19，大于第二条的得分1.00，所以排在第一位；

1. 当输入关键词“显卡”时，结果如下：



（2）、在index=‘math’内搜索“求导”，结果如下所示：



一共有4个结果，得分最高的1.9分者排在第一位，得分后两位者内容如下所示：



1. **注意**

代码运行了一遍之后（即建立好索引与文档之后，相关代码就应该注释掉，不然会每一次运行都重复建立索引，或插入相同文档。）