### 安装UBUNTU系统

本项目采用的UBUNTU 16.04 LTS系统，安装过程较繁琐，但是比较简单，因此不做赘述。

### 2. 配置JDK环境

1.下载jdk-7u79-linux-x64.tar.gz

2.输入指令sudo mkdir /usr/lib/jvm创建一个目录

3.输入指令sudo tar –zxvf jdk-7u79-linux-x64.tar.gz –d /usr/lib/jvm将jdk压缩包解压到/usr/lib/jvm文件夹下。

4.输入指令sudo gedit /etc/profile打开全局配置文件，在末尾添加图2.3中的命令

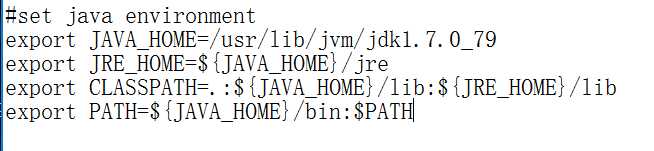


图2.3 配置JDK环境配置

### 3. 配置Tomcat环境

1.下载apache-tomcat-7.0.57.tar.gz

2.输入指令sudo tar –zxvf apache-tomcat-7.0.57.tar.gz–d /opt将Tomcat压缩包解压到/opt文件夹下

3.输入指令sudo gedit /etc/profile打开全局配置文件，在末尾添加图2.4中的命令

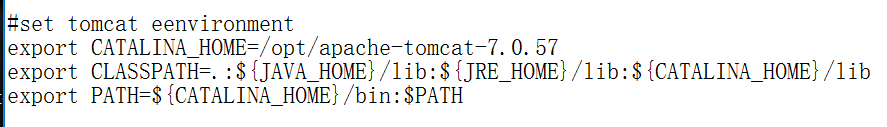


图2.4 配置Tomcat环境变量

### 4. 配置ANT环境

1.下载apache-ant-1.9.6-src.tar.gz

2.输入指令sudo tar –zxvf apache-ant-1.9.6-src.tar.gz–d /opt将ANT压缩包解压到/opt文件夹下。

3.输入指令sudo gedit /etc/profile打开全局配置文件，在末尾添加图2.5中的命令。

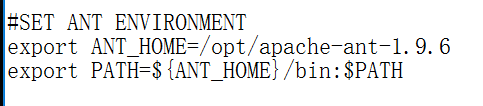


图2.5 配置ANT环境变量

### 5. 配置MySQL

通过指令sudo apt-get install mysql-server mysql-client

libmysqlclient-dev即可自动下载完成配置

### 6. 配置Nutch

1） 下载apache-nutch-2.2.1-src.tar.gz

2） 输入指令sudo tar –zxvf apache-nutch-2.2.1-src.tar.gz–d /opt将Nutch压缩包解压到/opt文件夹下

3） 输入指令sudo gedit /etc/profile打开全局配置文件，在末尾图2.6所示指令[7]。

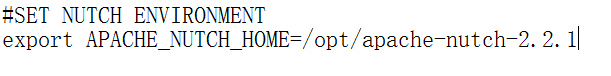


图2.6 配置NUTCH根目录

4） 打开${APACHE\_NUTCH\_HOME}/ivy/ivy.xml，将

<dependency org="org.apache.gora" name="gora-core" rev="0.3" conf="\*->default"/>

改成

<dependency org="org.apache.gora" name="gora-core" rev="0.2.1" conf="\*->default"/>

5） 将以下代码取消注释[7]

<dependency org="mysql" name="mysql-connector-java" rev="5.1.18" conf="\*->default"/>

6） 打开${APACHE\_NUTCH\_HOME}/conf/gora.properties，添加以下代码使Nutch以MySQL作为数据库。

###############################  
# MySQL properties #

###############################  
gora.sqlstore.jdbc.driver=com.mysql.jdbc.Driver  
gora.sqlstore.jdbc.url=jdbc:mysql://localhost:3306/nutch?createDatabaseIfNotExist=true  
gora.sqlstore.jdbc.user=root  
gora.sqlstore.jdbc.password=123456

7) 打开 ${APACHE\_NUTCH\_HOME}/conf/gora-sql-mapping.xml，找到

<primarykey column="id" length="512"/>，将其改成

<primarykey column="id" length="767"/>

8) 打开${APACHE\_NUTCH\_HOME}/conf/nutch-site.xml，在<configuration>双标符中的末尾添加以下内容：

|  |
| --- |
| <property> <name>http.agent.name</name> <value>YourNutchSpider</value> </property> <property> <name>http.accept.language</name> <value>ja-jp, en-us,en-gb,en,zh\_ch;q=0.7,\*;q=0.3</value>  <description>Value of the "Accept-Language" request header field. This allows selecting non-English language as default one to retrieve. It is a useful setting for search engines build for certain national group. </description> </property> <property> <name>parser.character.encoding.default</name> <value>utf-8</value> <description>The character encoding to fall back to when no other information is available</description>  </property> <property> <name>storage.data.store.class</name> <value>org.apache.gora.sql.store.SqlStore</value> <description>The Gora DataStore class for storing and retrieving data. Currently the following stores are available: .... </description> </property> |

在该目录下，找到以下双标

|  |
| --- |
| <name>file.content.limit</name>   <value>65536</value> |

将value值改大一点，例如6553600

找到

|  |
| --- |
| <property>  <name>parser.skip.truncated</name>   <value>true</value> |

将value改成false

找到

|  |
| --- |
| <property>  <name>plugin.folders</name>   <value>./src/plugin</value> |

将value的值改成plugins

9) 切换到${APACHE\_NUTCH\_HOME},输入ant后按回车进行编译

### 7. 配置ElasticSearch

1.下载elasticsearch-2.2.1.tar.gz

2.通过sudo tar –zxvf elasticseach-2.2.1.tar.gz –d /opt将elasticsearch解压到/opt目录下

3.输入命令bin/plugin –install mobz/elasticsearch-head安装elasticsearch的web界面插件**[8]**。

4.输入命令bin/elasticsearch即可运行elasticsearch[8]。

5.打开浏览器，输入<http://localhost:9200/_plugin/head>即可看到elasticsearch的相关信息。

### 8. 配置IKAnalyzer

配置IKAnalyzer之前，要先配置maven环境，通过sudo apt-get install maven即可配置完成。

在IKAnalyzer解压目录下，通过指令mvn clean package编译IKAnalyzer。编译完成后将生成的target目录下的release下的elasticsearch-analysis-ik-1.9.1-yxjay.zip解压到elasticsearch根目录下的plugins文件夹下。然后在elasticsearch根目录下的config/elasticsearch.yml增加以下代码

|  |
| --- |
| index.analysis.analyzer.ik.type:”ik” |

  重启elasticsearch即可配置完成。

### 9. 注意事项

* + 安装过程中可能有用到其他库软件如maven等，安装即可。
  + MySQL的字符设置为utf8mb4，与Nutch一致。
  + 在ElasticSearch-JDBC导入数据之前，可通过指令

curl –XPOST ‘http://localhost:9200/nutch’-d @mapping.json来重建映射，新映射中去掉了无关字段，相关文件在mapping.json中