操作文档

# 配置环境

## 安装JAVA

添加相应的源并更新

sudo add-apt-repository ppa:webupd8team/java

sudo apt-get update

安装Oracle JDK 8

sudo apt-get install oracle-java8-installer

## 安装Maven

到apache maven官网下载 xxx-bin.tar.gz

在/usr/local中创建放置maven的文件夹

cd /usr/local

mkdir maven

解压该文件

sudo tar -zxvf apache-maven-3.3.9-bin.tar.gz –c /usr/local/maven

在~./bashrc下配置环境变量

M2\_HOME=/usr/local/maven/apache-maven-3.3.9

export MAVEN\_HOME=/usr/local/maven/apache-maven-3.3.9

export PATH=$PATH:$MAVEN\_HOME/bin

## 配置pom.xml文件

设置Hadoop的版本

<properties>

<hadoop.version>2.7.3</hadoop.version>

</properties>

添加与Hadoop相关的依赖

<dependency>

<groupId>org.apache.hadoop</groupId>

<artifactId>hadoop-common</artifactId>

<version>${hadoop.version}</version>

</dependency>

<dependency>

<groupId>org.apache.hadoop</groupId>

<artifactId>hadoop-hdfs</artifactId>

<version>${hadoop.version}</version>

</dependency>

<dependency>

<groupId>org.apache.hadoop</groupId>

<artifactId>hadoop-mapreduce-client-core</artifactId>

<version>${hadoop.version}</version>

</dependency>

<dependency>

<groupId>org.apache.hadoop</groupId>

<artifactId>hadoop-client</artifactId>

<version>${hadoop.version}</version>

</dependency>

<dependency>

<groupId>org.apache.hadoop</groupId>

<artifactId>hadoop-core</artifactId>

<version>1.2.1</version>

</dependency>

添加与分词工具Jieba相关的依赖

<dependency>

<groupId>com.huaban</groupId>

<artifactId>jieba-analysis</artifactId>

<version>1.0.2</version>

</dependency>

添加与爬取网页时所用的库的依赖

<dependency>

<groupId>org.jsoup</groupId>

<artifactId>jsoup</artifactId>

<version>1.10.3</version>

</dependency>

# 运行

## 运行爬虫爬取网页

设定com.executor.ExecutorMain.java中的CLUSTER\_SIZE可以修改每个类别的网页爬取的数量

执行com.executor.ExecutorMain.java中的main函数进行爬取网页，爬取的网页存放在./input/目录下

## 2.2 运行使用Hadoop的Map-Reduce框架实现的Kmeans聚类算法

运行com.bigdata.KmeansDriver.java中main函数，并且需要提供五个参数，分别是

(1). 输入文件的目录

(2). 输出文件的目录

(3). 缓存文件放置的目录

(4). Kmeans算法中K的值

(5). 最大迭代次数

当迭代结束后将结果输出到(2)中，打开该文件可以查看聚类结果

其显示格式为：

网页编号 对应簇的编号