数据挖掘大作业

n-gram input text predictor

钱钧 2120151028

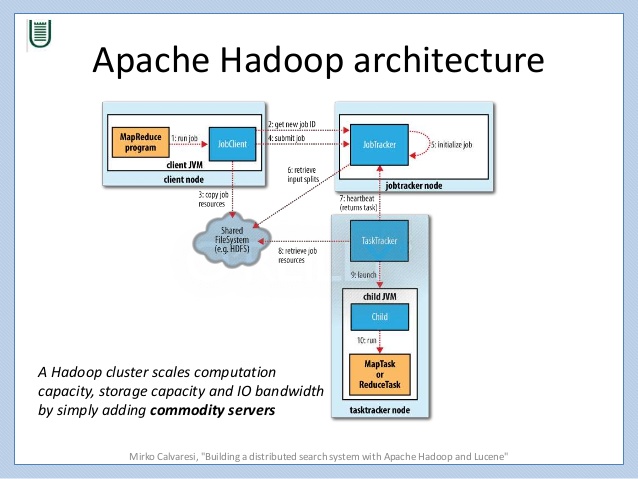
# 问题描述

使用hadoop集群建立n-gram语言模型，并使用该语言模型制作一个简单的用户输入预测系统。

# 背景知识

## MapReduce

Hadoop是Google MapReduce的开源实现项目。在MapReduce程序中，数据以键值对的形势存储，然后通过Mapper和Reducer进行处理：



InputFormat定义了Mapper如何从文件读入数据，并写入Writable类型：

Mapper：Map(k1, v1) 🡪 list(k2, v2)

然后会进行Shuffle和sort(按照键值)，接着到Reducer，会针对同一个Key的所有value进行处理

Reducer：Reduce(k2, list(v2)) 🡪 list(v3)

## N-gram

# 任务目标

我们需要构造一个输入文本预测器，通过n-gram及对应的语言模型，预测用户之后可能会输入的内容，具体步骤如下：

1. 利用wiki数据计算ngram
2. 构造语言模型
3. 代码
4. 语言自动完成

环境：

1. AWS Elastic MapReducer(ENR), 用m3开头的机器
2. MapReduce的java程序需要用jre1.7编译

构造n-gram模型：  
数据集：s3://cmucc-datasets/enwiki-20160204-pages

n-gram格式<phrase><\t><count>

算出所有的n-gram之后，需要选出出现次数最多的100个n-gram，如果次数一样就按照字母序排序，完成之后保存到文件中，之后会用来评分。使用Hive的sql语法选择。

具体步骤：

1. 因为原始数据是xml格式，所以需要进行数据清洗
   1. 移除<ref>和</ref>：使用正则表达式
   2. 移除所有URL，也就是以HTTP/HTTPS/FTP开头的内容
   3. 单词之间空格只有一个
   4. 所有的单词转化为小写字母

2．使用清洗后的数据，在一个MapReduce Job中生成1-gram, 2-gram, 3-gram, 4-gram, 5-gram

1. 现在小数据上测试
2. 不输出空字符
3. 把处理完的数据保存在s3中，在MapReduce中使用s3cmd进行s3的写入

操作步骤：

1. 开启一个EMR集群，确保Hive，HBase和Hadoop都安装好了，使用AMI version 3.10.0
   1. 连接：ssh –I ../key.pem hadoop@ec2-54-86-122-167
   2. 安装tmux
   3. 复制代码到服务器：scp –I ../key.pem ./WordCount.java hadoop@ec2-54-86-122-167:~/ngram/
2. 计算完成后，把前100个次数最多的ngram结果存储在ngrams文件中

命令代码在附件emr.cmd中

用hive进行统计hive.cmd

最后把ngrams文件和MapReduce代码放到同一个文件夹中用submit.cmd中的命令来提交

# 构造语言模型

模型公式

例子：

this 1000

this is 500

this is a 125

this is a blue 60

this is a blue house 20

于是我们需要使用MapReduce Job，从HDFS读取前一个阶段生成的ngram，计算所有单词短语模型，直接写入HBase中。

# Mapper

同样使用上面的例子：

this 1000

that is 1

this is 500

this is a 125

我们要做的实际上就是把目前的键值对，拆成phrase和word的形式，有两个地方需要注意：

1. 如果key拆分之后只有一个单词，需要过滤掉
2. 如果value的值为1，需要过滤掉

假设key拆分之后有n个词，那么把前n-1个词拼成phrase，最后一个词为word，不过需要注意的是，还要加上之前的计数。

# Reducer

Reducer的主要工作就是排序，另外使用TableReducer直接写入到HBase中，具体步骤如下：

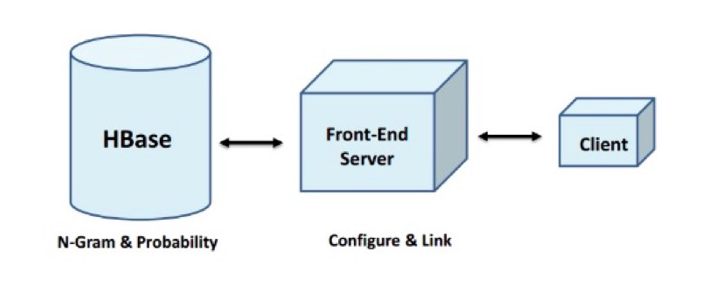
1. 遍历一个key的value，计算出现的总次数
2. 给value中不同的word进行排序
3. 计算概率，然后写入到记录中

# 工作日志

代码详见log.cmd

# 使用Web展示语言模型

总体架构如下所示：



ssh连接到master节点，开启HBase的RESTful API *hbase-daemon.sh start rest*，开启／重启Apache服务sudo service httpd restart。

开启服务代码见：start.cmd