

**《大数据导论》项目报告**

**题目：基于spark平台的校徽社区发现与可视化**

**课程名称： 大数据导论**

**专业班级： CS1703**

**学 号： U201714668**

**姓 名： 葛松**

**同组成员： 此处略去**

**指导教师： 石宣化**

**报告日期： 2019年10月27日**

**计算机科学与技术学院**

目 录

[一.测试环境说明 3](#_Toc755594861)

[1.亚马逊云（1 Master 2 Slave) 3](#_Toc1416023182)

[2.spark集群 3](#_Toc298125972)

[二.数据集特征说明 3](#_Toc972605807)

[三.测试应用说明 4](#_Toc115206657)

[1.kmeans算法 4](#_Toc70298129)

[2.WordCount 4](#_Toc243110911)

[四.工作动机 4](#_Toc2016747294)

[五.问题挑战 4](#_Toc272327630)

[1.集群搭建 4](#_Toc1156565744)

[2.数据获取转化以及PySpark代码编写 5](#_Toc376875057)

[六.应用测试的结果数据 6](#_Toc601698078)

[1.莎士比亚全集测试结果 6](#_Toc1476967137)

[2.校徽测试结果 7](#_Toc1578037126)

[3.典型样例 10](#_Toc602569078)

[七.小组分工与心得体会 11](#_Toc269365760)

[1.小组分工 11](#_Toc340606628)

[2.个人感想 12](#_Toc1502774186)

# 一.测试环境说明

## 1.亚马逊云（1 Master 2 Slave)

图1-1为亚马逊云平台硬件信息：

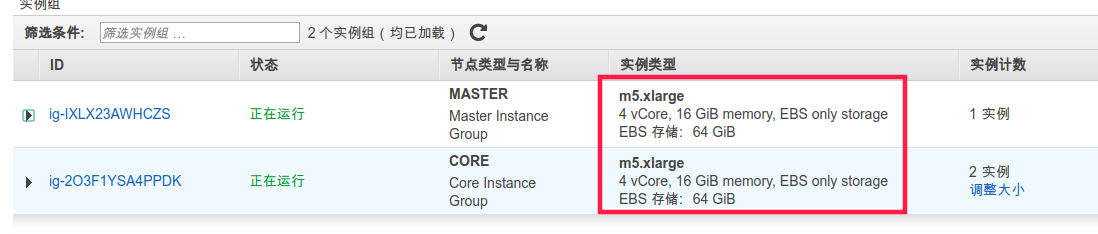


图1-1 亚马逊云平台硬件信息

## 2.spark集群

CPU：intel(R) Celeron(R) CPU 3865U @ 1.80GHz

内存：8.00GB

虚拟化环境：windows上的vmware上安装3台ubuntu虚拟机

分别为一台master（2G)、两台slave(1G+1G)

图1-2为启动spark-shell画面：

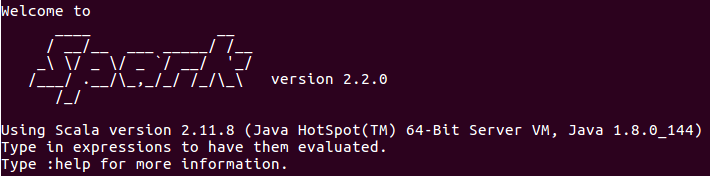


图1-2 启动spark-shell画面

# 二.数据集特征说明

（1）首先我们选取了莎士比亚全集，大小为5.5M，来测试词频。

（2）还有6000张校徽的数据集，筛选之后剩下4000张，每张resize到200 \* 200 像素，通过 AutoEncoder得到的向量为 512 × 6 × 6 = 18432维(单张图片)，来进行校徽分类。

# 三.测试应用说明

#### 1.kmeans算法

算法思想大致为：先从样本集中随机选取 k个样本作为簇中心，并计算所有样本与这 k个“簇中心”的距离，对于每一个样本，将其划分到与其距离最近的“簇中心”所在的簇中，对于新的簇计算各个簇的新的“簇中心”。其中 ， K 设为20， 将整个数据集初步分割为20个cluster(由于一共来自14个国家、地区) 。

#### 2. WordCount

作为基准测试，性能比较。

# 四.工作动机

随着云时代的来临，大数据也吸引了越来越多的关注。我们小组希望能运用大数据导论课堂以及自己所学的知识做一些有趣的项目。例如，我们通过《莎士比亚全集》词频搜索的尝试，利用大数据进行一些应用。同时，基于以前做过的校徽分类项目的一部分，我们计划通过大数据相关理论知识更好地完成这个项目。

# 五.问题挑战

## 1. 集群搭建

**本地集群**

集群搭建比我们想象中的困难很多，开始的过程中遇到很多问题，我们查阅了很多网络资料来解决这些问题，下面列举过程中的一些困难：

（1）ubuntu教程版本与实际操作版本的差别导致有些步骤难以实现，例如配置虚拟机的ip中edit connectinos的差别。

（2）对于linux的指令不熟练，修改interfaces文件与resolv.conf文件时遇到困难。

（3）网络连接问题导致一些下载不成功且没有在连续时间完成导致反复尝试。

（4）配置过程中出现permission denied问题，查阅相关资料后，通过chmod -R 777 修改文件权限解决。

（5）虚拟机克隆但是没有修改内存和网络配置导致后来内存不足。

**亚马逊云**

虽然说Amazon已经提供了一套解决方案，相对本地集群搭建比较简单，但是在实际操作的过程中也是遇到了一些问题。以下同样列举一些遇到的问题

1. 由于Amazon集群使用的是基于yarn的解决方案，所以在pyspark的脚本中设置Master时只需要简单的设置为yarn即可，但是官方的确没有给出一个完整的样例，导致诸如此类的一些参数需要在网络上查找很久才知道应该设置为什么。
2. Amazon的服务器由于位置被限制在美国东部，操作以及上传文件的时候常常会遇到链接突然断开导致需要重新来过的情况

总体来说，亚马逊的集群搭建好之后之后的每次使用都非常方便(除了文件传输)，并且计算速度等相比本地集群有着很大的优势。

## 2. 数据获取转化以及PySpark代码编写

1. **前期数据获取**

我们用于计算的校徽数据是通过模拟浏览的方式来变相使用Google Image 搜索的方式来获得的，所以遇到的第一个问题就是反爬虫问题，可能连续爬10到20张左右就会被Google检测到，最后在不断尝试中偶然发现中途关闭浏览器可以刷新状态，我们便利用这个特性实现了可以正常运行的爬虫。

1. **数据导入与转化**

由于从AutoEncoder得到的数据是.npy格式的(numpy)，而PySpark并没有提供针对.npy文件的读取的API，而在spark中引入额外的包又很麻烦，并且还有相关依赖等种种问题，所以我们便决定将数据保存为.csv格式(大小为1.7G)之后再用spark提供的针对txt文件格式的API进行读取，并通过阅读文档将其一步步转化为Spark内置的DataFrame的格式(我们没有使用RDD格式，由于Spark的DataFrame在各个方面都要优于RDD)，其中Spark对的Python的支持相对Scala以及Java都较弱，所以遇到的一些问题，官方文档并没有给出很好的解决方案，最终还是需要在各种论坛之中寻找蛛丝马迹。

1. **本地运行Community Detection算法以及前端可视化制作**

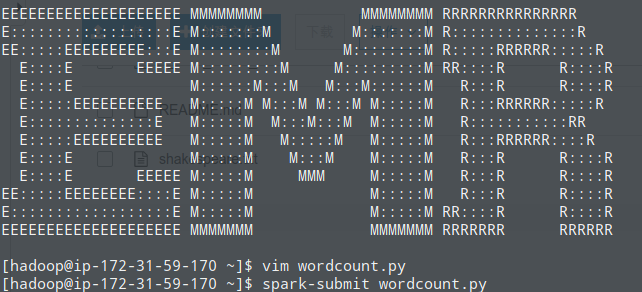
Spark中图计算相关的包对Python的支持到1.4就不再继续了，而我们使用的Spark版本是2.4.5，所以只好在本地运行社区发现相关算法，然后再根据得到的数据进行可视化前端的制作。其中可视化前端我们采用了Github上一个名为3d-force的仓库所提供的API，通过对其进行一些魔改，来制作我们的前端。在整个过程当中我们也遇到了很多困难，如Json数据的转化问题：由于一些学校的名字当中含有引号，导致整个Json文件无法被解析，所以我们专门编写了一段脚本用于将GraphX格式的数据转化为符合规范的Json数据，最终解决了Json数据的格式问题。

# 六.应用测试的结果数据

## 1. 莎士比亚全集测试结果

（1）在本地集群上运行时间：13s。

（2）在亚马逊云上运行时间：15s。



6-1运行开始

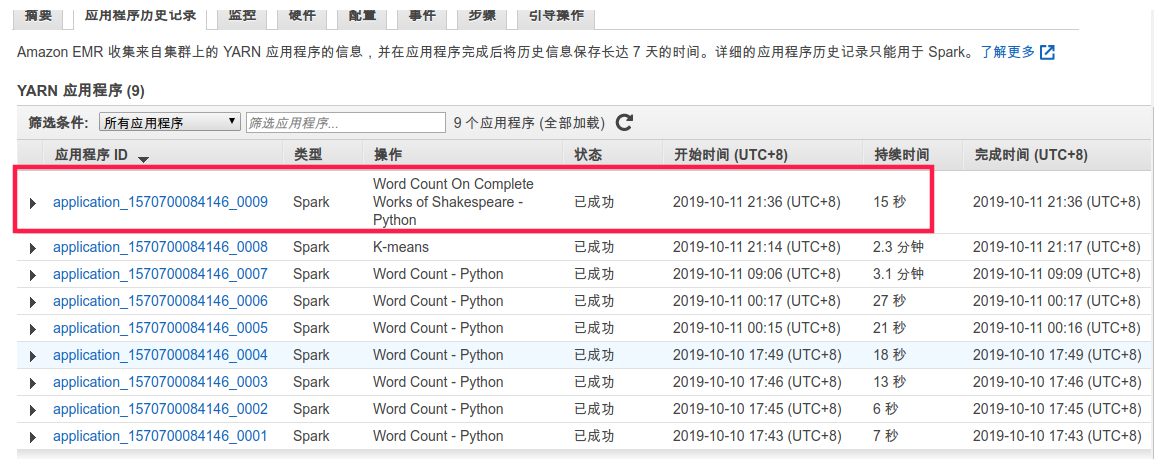


图6-2运行时间

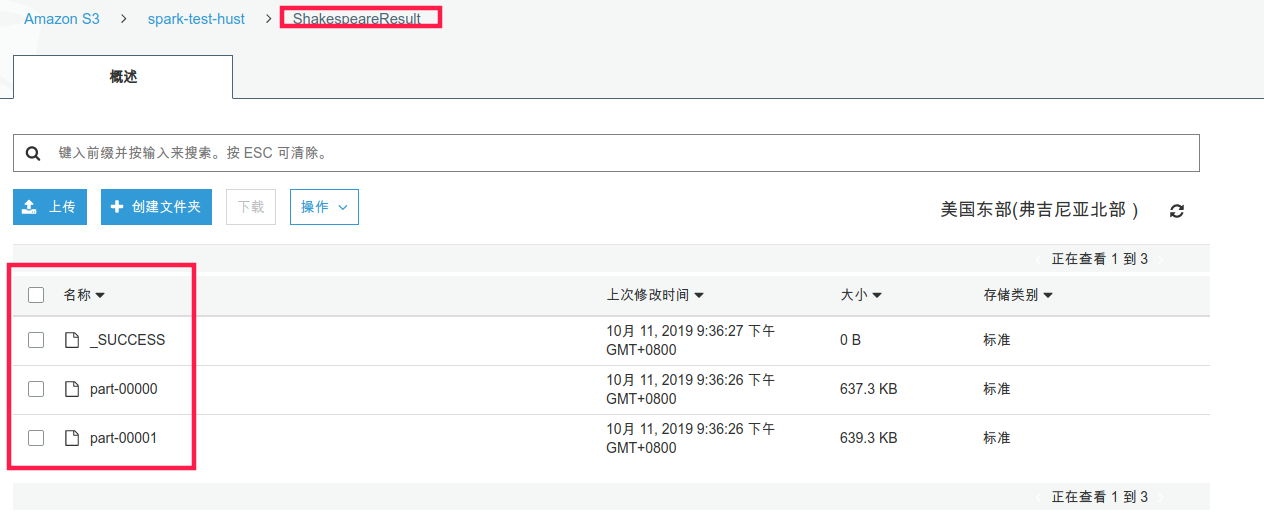


图6-3运行结果（1）

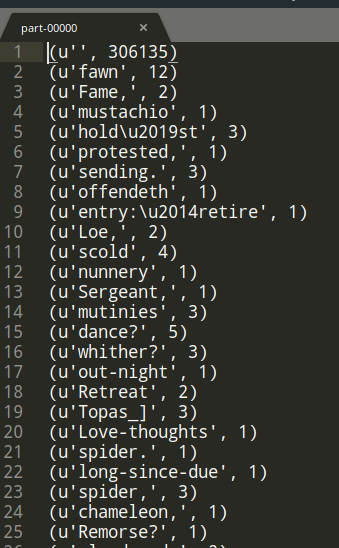


图6-4 运行结果（2）

## 校徽测试结果

整个流程如下：爬数据(数据清理, 标准化) -> 通过autoencoder中间层拿到图片的高维向量表示 -> kmeans 分出20个 cluster(K设置为20, 因为有14个国家/城市) -> 得到属于每一个cluster的图片 -> 对同一个cluster之中的图片使用社区发现算法，算出更小的"社区" -> 分析结果。

（1）本地集群测试结果：12min

（2）亚马逊云测试结果：2.3min



图6-5 运行时间

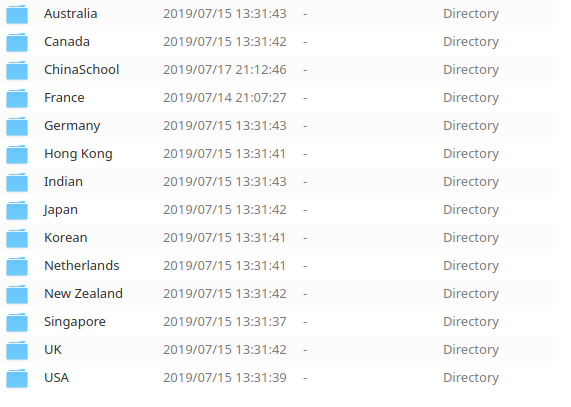


图6-6 爬到的数据（1）



图 6-7 爬到的数据（2）

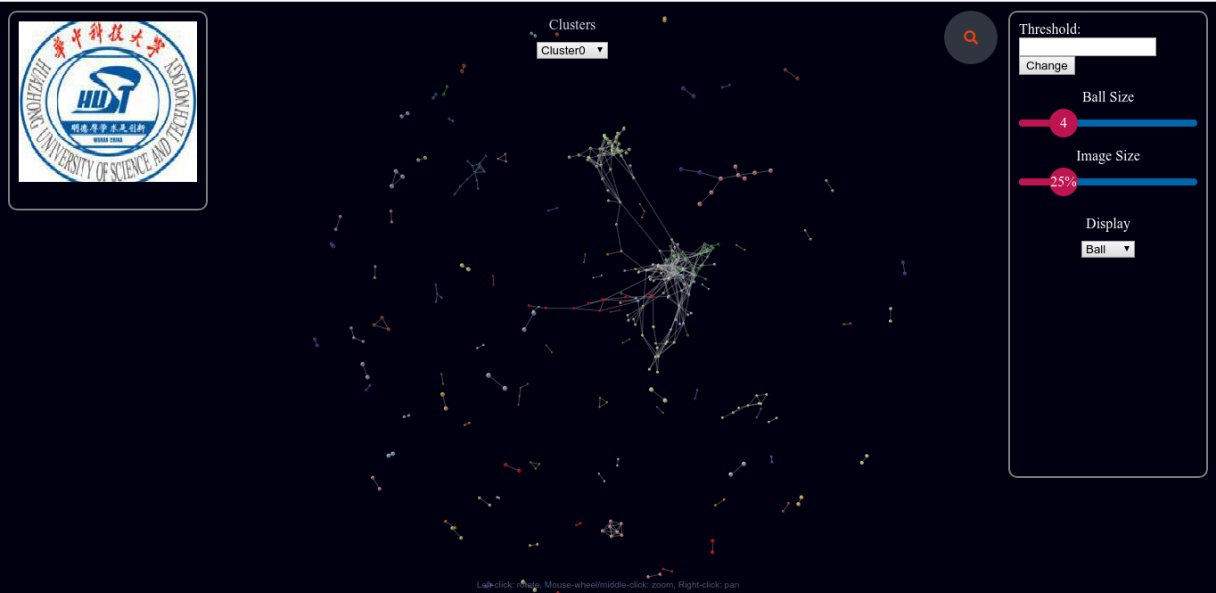


图6-8 可视化界面（1）



图6-9 可视化界面（2）

## 典型样例

1. 在Kmean中的Cluster中某一个国家的占比非常大，也就是某些国家具有典型性的特征

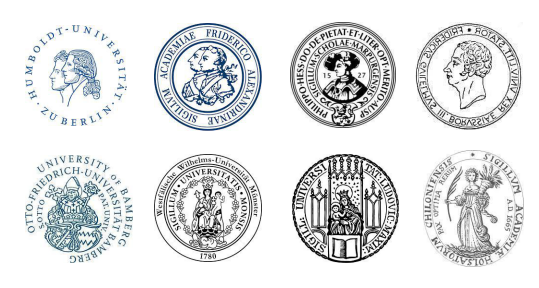


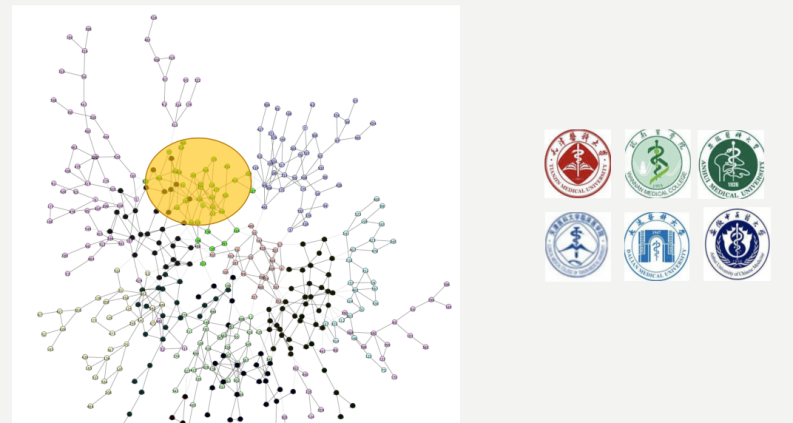
图6-10 印度与德国的典型校徽



图6-11 美国与日本的典型校徽

如印度由于主神是太阳神，并且境内恒河被尊为圣河，所以许多印度大学的校徽都以太阳与河流作为主要构图元素。又如美国许多学校的校徽中都含有“火炬”象征着自由，而“书本”象征着知识，就如同自由女神像一般、英国校徽大多都有盾牌以及皇冠等元素，而新加坡由于是“狮城”，所以几所著名大学之中都有狮子作为主要元素等等。可见许多国家还是有一些由于宗教历史等各种原因而累积下来的校徽主体构图要素。

1. 某些同一类型的学校在社区发现得到的小Cluster中靠的很近



图中的靠的很近的6张校徽之中均具有手杖以及蛇形图案作为主要构图要素，由于在一些神话传说当中蛇杖为巫医所持有之物，象征着医术与治疗，而这些校徽所对应的学校无一例外也都是医学院，符合我们的预期。

# 七.小组分工与心得体会

## 小组分工

**小组内负责部分:**

1. 数据爬取以及预处理
2. Spark亚马逊云配置
3. 参与Kmean代码编写
4. 莎士比亚WordCount代码编写
5. 本地社区发现(Community Detection)代码编写
6. 前端可视化以及对应后台编写

**报告编写负责部分：**

1. 问题挑战之 亚马逊平台搭建
2. 问题挑战之 数据获取转化以及PySpark代码编写
3. 应用测试的结果数据之 典型样例

## 个人感想

由于这次项目实际上是由之前的一个小项目移植而来，所以总的来说难度不是特别大，尤其是数据收集以及预处理的部分，之前基本上已经完成，但是对于Spark这个大数据处理平台的学习还是一个全新的过程。期间免不了遇到各种难以解决的各种问题，但是在不断遇到问题并尝试解决问题的过程之中还是学到了很多东西，如对数据流的基本理解，Spark等各种框架想要解决的问题。

除此之外，我们在将Spark大数据处理平台应用到已有的数据之后也的确发现了许多有趣的关联和发现（参见 典型样例 一节），更是深刻体会到了大数据所潜藏的无限可能。

而在前端可视化的制作当中我也回想起了课上老师多次提到的对数据可视化的看法，有了数据还不够，还需要合适的方式把数据所揭示的现象和关联以某种合适的方式展现出来才是完整的大数据分析，虽然说前端的确做的有点丑，但是还是可以借其一窥不同校徽之间的关联，并看到隶属于同一个Cluster的校徽所具有的一些共性，得到不少有趣的发现。

同时，在对于国家层面的分析以及小的集群的分析的不同也非常具有典型性，一个国家所具有的共性只有在Kmeans初步划分出来的大Cluster中才容易展现出来，而诸如“蛇杖与医学院”之类的关系就需要通过在Kmeans的大Cluster中进行社区发现算法划分出小的“社区”才容易发现,毕竟不可能有非常多的学校都具有十分相似的校徽,这就表现出了大数据处理的另一个方面:从不同的角度和方法分析同一份数据常常可以得出许多不同或者说多样的结果和发现。

总的来说整个项目的完成的经历了不少困难，不仅仅是代码的编写，还有本地平台的搭建，前端的编写等等，但是这些困难都在最终结果形成之后变得值得，的确是一次非常棒的经历，而且原始项目的初心也就是简单的满足自己的兴趣和好奇心，没想到在这次的大数据导论课上刚好可以借助大数据处理平台进行进一步的拓展和延伸，也是挺惊喜的。