

## 2017 云计算实践作业选题

本学期云计算课程实践作业的目标是，让选修该课程的同学掌握基于 Spark、HDFS 和 MongoDB 的本地高效分布式数据处理和存储环境的搭建、使用技术。

实验中训练的具体技术有：网络数据爬取；MongoDB 数据存储和读取；HDFS 数据存储和读取；Spark Streaming、Spark GraphX 和 Spark MLlib 的程序编写等。

实践作业按照使用的 Spark 技术分为三个阶段：Spark Streaming、Spark GraphX 和 Spark MLlib，分组完成，每个小组不超过 6 名同学，作业完成情况以课堂报告的形式进行检查。

### 一、进度安排

项目	第一阶段		第二阶段	
	工作内容	报告时间	工作内容	报告时间
Streaming	确定需求并完成输入数据准备	2017-10-25	编写 Streaming 程序并展示结果	2017-10-30
GraphX	展示图	2017-11-01	图计算并展示结果	2017-11-08
MLlib	聚类算法应用	2017-11-13	分类算法应用	2017-11-22

### 二、选题内容

选题规则：一共 6 个题目，每个题目最多 4 个组选，根据发帖的顺序来确定每组选定的题目。

**重要说明：**每个题目中都有“确定需求”的步骤，这一步骤中我们给出的只是示例需求，希望大家能够基于定义好的数据自主的提出一些有趣的、有意义的、创新的需求，并通过 Streaming 计算、图计算和机器学习的库实现这些需求。最后评分会根据大家的提出的需求的有趣程度给出一定的奖励分。

#### 题目 1. 京东笔记本电脑评论数据

#### 题目 2. 京东手机评论数据

#### 题目 3. 京东微单相机评论数据

以上三个题目除了数据类型、数据处理目标不同之外，其它操作步骤大致相似。

1. 数据准备：从京东网站爬取对应类型商品的不同品牌产品的评论。每一个评论的数据信息至少包括如下内容：

商品类型+品牌+型号+用户信息+评论内容+评分+时间+赞数+评论数（如下图）



2. 数据存储：将爬取到的用户评论信息存入 MongoDB 数据库集群中

3. Spark Streaming:

3.1. 确定需求：即确定本程序的计算目标。

例如，统计某个品牌商品随着时间发展用户对其评价的变化曲线；统计用户对京东网站的评价变化曲线；统计用户对某个商品关注焦点的变化曲线；用户对产品优点的关注变化等等。

3.2. 输入数据准备：从 MongoDB 中将对实现上述需求有用的信息整理成日志形式并模拟数据流。

例如，统计对商品关注焦点的变化时，需要商品类型、评论内容和时间信息，则把这些数据按照某种格式整理为一行，存到一个日志文件中，然后将这个文件存入 HDFS。在存数据到日志中的时候需要模拟为数据流的形式，例如每秒钟写入 5M 数据。

3.3. 编写 Spark Streaming 程序：按照统计目标编写程序，并将结果存入 MongoDB 中。

3.4. 选择结果展示方式：在保存结果数据的同时，需要对结果数据进行可视化展示

例如，评分的变化可以用曲线图展示；关注焦点可以用热词出现频率的柱状图表示等。

#### 4. Spark GraphX:

4.1. 确定需求：确定使用 GraphX 在这些评论数据上将要做的事情

例如，想要计算随机的两个用户之间的最小距离，则以用户和商品之间的关系、商品和品牌之间的关系、品牌和商品类型之间的关系等，构造一个用户、商品为节点的关系图；或者以所有评论信息为输入，构造一个汉字与汉字之间的关系图，如果两个汉字连续地在一句话中出现，则表示两个字之间存在一条边，这样可以计算出哪些字更容易在一起使用，可以扩展为中文分词的工作等。这一步需要定义好顶点和边的概念。

4.2. 构造图：首先准备边对应的日志数据，每个边至少要包括两个顶点、顶点间关系类型等信息。

4.3. 展示图：对构造好的图进行可视化展示

4.4. 按照需求进行图计算：使用 GraphX 提供的一些图计算 API 编写满足定义好的需求的程序。

例如，计算随机两个用户之间的最短路径、计算两个字之间的最短路径、计算通常与一个给定的字一起使用的其它字的集合、计算与给定的字有类似出入度和使用场景的其它字的集合等。该步骤也需要提供对结果的可视化展示。

#### 5. Spark MLlib:

5.1. 提出一个能够应用聚类算法的应用需求并实现该应用，包括从 MongoDB 数据库中整理数据集；在数据集上应用聚类算法；对聚类结果进行展示。

5.2. 提出一个能够应用分类算法的应用需求并实现该应用，包括从 MongoDB 数据库中整理训练集、测试集；使用一种分类算法进行训练；展示测试结果。

### 题目 4. “雪球”网股票新帖内容和用户所持股票信息

1. 数据准备：首先要把所有股票的代码抓取下来（通过 TuShare 网站），根据股票代码构造该股票在雪球网中的链接，通过该链接抓取该股票的新帖。

附 1：TuShare 网站：<http://www.waditu.cn/>，其中使用的“股票列表”的功能（返回值不止图中展示内容，需要阅读具体页面：<http://www.waditu.cn/fundamental.html#id2>）：

调用方法：

```
import tushare as ts

ts.get_stock_basics()
```

结果显示：

	name	industry	area	pe	outstanding	totals	totalAssets
code							
600606	金丰投资	房产服务	上海	0.00	51832.01	51832.01	744930.44
002285	世联行	房产服务	深圳	71.04	76352.17	76377.60	411595.28
000861	海印股份	房产服务	广东	126.20	83775.50	118413.84	730716.56
000526	银润投资	房产服务	福建	2421.16	9619.50	9619.50	20065.32
000056	深国商	房产服务	深圳	0.00	14305.55	26508.14	787195.94
600895	张江高科	园区开发	上海	171.60	154868.95	154868.95	1771040.38
600736	苏州高新	园区开发	江苏	48.68	105788.15	105788.15	2125485.75
600663	陆家嘴	园区开发	上海	47.63	135808.41	186768.41	4562074.50
600658	电子城	园区开发	北京	19.39	58009.73	58009.73	431300.19
600648	外高桥	园区开发	上海	65.36	81022.34	113534.90	2508100.75
600639	浦东金桥	园区开发	上海	57.28	65664.88	92882.50	1241577.00
600604	市北高新	园区开发	上海	692.87	33352.42	56644.92	329289.50

附 2：雪球网的股票链接形式：<https://xueqiu.com/S/SH600004>，要根据股票代码信息构造出股票对应的链接，链接页面中需要抓取的新帖信息区如下：

全部讨论交易新闻公告研报

热帖新帖

搜索该股票相关信息

**崔小弟:**  
\$白云机场(SH600004)\$ 10月11日12.61买入20000元。10月12日12.91出，盈利433元。  
9分钟前 来自Android客户端 赞(1) | 转发 | 收藏 | 打赏 | 评论(1)

**风无雅:**  
\$白云机场(SH600004)\$  
展开  
18分钟前 来自Android客户端 赞(3) | 转发 | 收藏 | 打赏 | 评论(2)

**L姓小哥:**  
雪球访谈：旅客吞吐量稳进TOP3，白云机场潜力如何？  
我刚下载了“雪球访谈：旅客吞吐量稳进TOP3，白云机场潜力如何？”全文，也推荐给你。网页链接...  
28分钟前 来自雪球 赞 | 转发 | 收藏 | 打赏 | 评论

点击用户后的页面和需要的用户所持股票列表如下：



全部	沪深	港股	美股	高增长	更多
股票	当前价	买入目标价	卖出目标价	备注	
海康威视	34.80	-	-	-	
国投电力	7.57	-	-	-	
深高速	9.70	-	-	-	
保利地产	10.32	-	-	-	

2. 数据存储：将爬取的数据存到 MongoDB 中；股票的新帖要包括如下信息：

股票+用户+帖子内容+来源+时间+赞数+评论数+评论列表（评论包括：用户+时间+内容）

用户所持股票包括如下信息：

用户+股票列表（股票信息为：名称+价格）

3. Spark Streaming:

- 3.1. 定义需求：例如计算用户对某只股票的情感变化曲线等
- 3.2. 输入数据准备：参照前三个题目对应的“输入数据准备”工作
- 3.3. 编写程序并保存结果到 MongoDB
- 3.4. 展示结果

4. Spark GraphX:

- 4.1. 定义需求：例如以用户和股票为顶点构造一个用户、股票之间关系的图结构
- 4.2. 构造图
- 4.3. 展示图
- 4.4. 利用图计算实现定义的需求并展示结果

5. Spark MLlib:

- 5.1. 提出一个能够应用聚类算法的应用需求并实现该应用，包括从 MongoDB 数据库中整理数据集；在数据集上应用聚类算法；对聚类结果进行展示。
- 5.2. 提出一个能够应用分类算法的应用需求并实现该应用，包括从 MongoDB 数据库中整理训练集、测试集；使用一种分类算法进行训练；展示测试结果。

## 题目 5. 阿里商品销售数据集

1. 数据准备:

1.1. <https://tianchi.aliyun.com/datalab/dataSet.htm?spm=5176.100073.888.29.48031a90VhajRS&id=22> 上给出了淘宝的用户购买记录数据文件 (IJCAI17\_data.zip, 需要先用淘宝账户登陆才能下载), 以及数据文件的数据格式

1.2. user\_pay.txt 文件就是的用户购买记录, 里面的记录时间上是乱序的, 需要先对记录按时间进行升序排序

2. 将数据按照原来内容存到 mongodb 里

3. Spark streaming:

输入：按时间先后顺序，每秒的店铺购买记录

处理：统计各个店铺的每分钟的销售单数以及总体的销售单数。

结果展示：以分钟为单位的店铺销售单数曲线，以及总体销售单数曲线

PS: streaming 的时间间隔不一定要设置为 1s, 可根据实际的处理能力进行调整, 比如如果能在 1s 之内处理完所有数据就设定为 1s, 如果记录太多一般不能在 1s 之内处理完可适当延长。

#### 4. Spark graphx:

第一阶段:

1. 构造图: 每一次的购买记录都可以构造一条边, 顶点是用户和店铺, 店铺节点还可以有店铺的一些属性(shop\_info.txt 提供了店铺的属性), 对于重复的购买, 我们可以增加这条边的权重。

2. 展示图: 对图进行可视化展示, 由于数据太多可选取一个小样本进行展示即可

第二阶段:

结合 graphx 提供的一些操作来完成一些主题, 比如说给用户推荐店铺或给店铺推荐优质用户

#### 5. Spark MLlib:

5.1. 提出一个能够应用聚类算法的应用需求并实现该应用, 包括从 MongoDB 数据库中整理数据集; 在数据集上应用聚类算法; 对聚类结果进行展示。

5.2. 提出一个能够应用分类算法的应用需求并实现该应用, 包括从 MongoDB 数据库中整理训练集、测试集; 使用一种分类算法进行训练; 展示测试结果。

### 题目 6. 豆瓣小组和用户数据

#### 1. 数据准备:

1.1. 可以从小组界面入手 (<http://www.douban.com/group/explore>)



爬取红框标出的所有标签下的小组信息, 每一个标签页面的右侧都有很多小组, 爬取所有的小组, 点进小组的详情页面



上海租房&找室友合租@房东直租 加入小组

创建于2015-06-26 组长：百步穿杨

【走哪住哪】  
听说  
去罗马有许多条路  
我不知道你会走哪条  
于是  
我在每条路上找好了房子  
只为你累了能够有处安歇

- 【备注】
- 1、凡已经租出去的房子，标题前请注明 [已出租]，可提高大家的找房效率！
  - 2、不反对品牌公寓方在这里发帖，但是如果【发帖过多（超过五贴）或刷帖者】、又或是被用户举报的品牌公寓方，一律永久封禁处理！
  - 3、这里不欢迎中介人员在此发帖，欢迎个人房东在此发布。
  - 4、转租、短租、合租、长租、床位、沙发客，都可以在此发布求租或是出租，在发帖时请注明类型，方便大家！

小组标签 上海租房 上海合租 走哪住哪 找室友 合租

推荐

友情小组



上海短租&沙发客 (212)



上海合租@找室友合租 (4004)

最近加入



卡尔加里路



白白



星新一



木子程程



Angel



王思梦



笑看风云



蛋蛋

> 浏览所有优秀租客 (6911)

爬取左下角的小组标签数据，右边的红框点进去可以查看这个小组的所有用户

安全 | <https://www.douban.com/group/zounazhuna/members>

管理员



百步穿杨  
(上海)

优秀租客



卡尔加里路  
(上海)



白白



星新一  
(上海)



木子程程  
(上海)



Angel  
(南京)



王思梦  
(上海)



笑看风云



蛋蛋  
(上海)



神秘嘉宾  
(上海)



Isyemily



河畔建筑鼓噪  
(上海)



露露啦啦下花  
(上海)



李女士



圆脸的反击  
(上海)



贪吃猫



夏天的微笑zzj



陈梦莹  
(上海)



心想事成  
(上海)



有你才幸福  
(上海)



枯鱼  
(上海)



最好的自己  
(上海)



转折点  
(上海)



牛油果和马卡龙  
(上海)



人生若只如初见  
(上海)



欢乐  
(上海)



seanleong  
(上海)



Nancy



胜利者  
(上海)



sumy



Monster  
(上海)



天天想你  
(上海)



Mrs 陈  
(上海)



悦  
(上海)



种太阳  
(上海)



出国务工  
(成都)

<前页 1 2 3 4 5 6 7 8 9 ... 197 198 后页> (共6910人)

爬取小组的所有用户数据，点击用户，进入用户的详情页面，可以看到这个用户加入了哪些小组



## 白白的豆瓣小组

小组主页 加入的小组 喜欢 推荐



我们假设用户是依次加入这些小组的，对于每一个小组我们构造一个用户加入小组事件<用户，小组，加入时间>，其中加入时间是我们随机设定的。最后需要对所有的用户加入小组事件按照时间先后进行排序，从而对真实的场景进行了模拟。

2. 将爬取的数据按照上述涉及到的内容全量存到 mongodb 里

### 3. Spark streaming:

输入：每个时间戳的所有用户加入小组事件，按时间升序处理

处理：对每个用户加入小组事件，更新用户画像，用户画像由一系列标签组成，标签可以直接用小租标签，比如说用户加入了第一个小组 A，用户画像为小组 A 的标签，然后用户加入了第二个小组，此时用户画像更新为小组 A 的标签和小组 B 的标签（实际的用户画像刻画挺复杂的，我们这里只是为了模拟一个流计算的场景），将用户

结果展示：只要能展现出用户画像在不断变化即可，比如说可以在处理过程中给用户画像数据截个图，过一段事件后再截个图，可以看出用户画像有变化即可

#### 4. Spark graphx:

第一阶段:

1.构造图: 每一个<用户, 小组>都可以构造一条边, 顶点是用户和小组。

2.展示图: 对图进行可视化展示, 由于数据太多可选取一个小样本进行展示即可

第二阶段: 结合 graphx 提供的一些操作来完成一些主题, 比如说给用户推荐小组; 给用户推荐好友

#### 5. Spark MLlib:

5.1. 提出一个能够应用聚类算法的应用需求并实现该应用, 包括从 MongoDB 数据库中整理数据集; 在数据集上应用聚类算法; 对聚类结果进行展示。

5.2. 提出一个能够应用分类算法的应用需求并实现该应用, 包括从 MongoDB 数据库中整理训练集、测试集; 使用一种分类算法进行训练; 展示测试结果。

### 三、评分规则

以报告内容为评分标准, 报告至少需要包括如下内容:

1. 数据情况说明: 基础分 1 分, 最高分 4 分, 记为 A
2. 程序情况说明: 基础分 1 分, 最高分 3 分, 记为 B
3. 结果情况说明: 基础分 1 分, 最高分 3 分, 记为 C
4. 小组有报告则小组获得基础分 =  $(A+B+C) \times \text{成员数}$
5. 小组报告人获得报告奖励分 = 成员数, 剩余分数在所有成员中平均分配
6. 根据“需求”内容给出奖励分 (3-4 分), 成员均可获得加分
7. 如果小组有一次没有报告, 则本课程成绩无分