实验三 PageRank 算法实现

一、实验目的

PageRank 网页排名的算法，曾是 Google 发家致富的法宝。用于衡量特定网页相对于搜索引擎索引中的其他网页而言的重要程度。通过对 PageRank 的编程在 Hadoop 和 Spark 上的实现，熟练掌握 MapReduce 程序与 Spark 程序在集群上的提交与执行过程，加深对 MapReduce 与 Spark 的理解。

二、实验平台

1）操作系统：Linux（实验室版本为 Ubuntu17.04，集群环境为 centos6.5）；

2）Hadoop 版本：2.9.0；

3）JDK 版本：1.8；

4）Java IDE：Eclipse 3.8。

6）Spark 版本：实验室版本为 2.1.0，集群环境为 2.3.0；

三、实验内容

1）在本地编写程序和调试

在本地 eclipse 上分别使用 MapReduce 和 Spark 实现 PageRank 算法,Spark 程序可采用 Java、Python、Scala 等语言进行编程。数据集中每一行内容的格式:网页+\t+该网页链接到的网页的集合(相互之间用英文逗号分开)。例如，下图为其中一行的数据，因为一行显示不出来所以使用多行显示。

```
Golden_mean_(philosophy)        Cyprus,The_Atlas_Society,Log_Cabin_Republicans,s
chizophrenia,alcoholic_beverage_control_state,Little_Iliad,Walter_Damrosch,Abdom
inalSurgery,Federal_Reserve,March_5,The_Red_Balloon,sardine,Presidents'_Day_(Uni
ted_States),Ford's_Theater,Barbary_pirate,chromosome,Mike_Todd,RTÉ,Single-lens_r
eflex_camera,Particulate,Hammadid,Spain,Alabama_River,AtlasShrugged,Mike_D._Roge
rs,cattle,April_3,An_Anarchist_FAQ,International_Bureau_of_Weights_and_Measures,
voice,WOMBLES,Huntsville_Channel_Cats,Academy Awards/Best Actor,diffusion_(anthr
opology),archeology,Harry_S._Truman,education,Art_Deco,AnnaKournikova,daira,Doug
las_Fairbanks,sword,smelt,Democritus,Libby_Prison,Djelfa_Province,70_mm_film,sec
ond-wave_feminism,Northern_Flicker,November_10,American_International_School_-_A
bu_Dhabi,gene_duplication,Mythology_(book),Christian_anarchism,Abbeville,Simon_C
ameron,Aeschines,Hebrew_alphabet,Assist,Union_blockade,Tissemsilt_Province,liber
alism,AnarchoCapitalists,pseudoarchaeology,Lalla_Fatma_N'Soumer,Kansas-Nebraska_
Act,library,Will_(law),Allan Dwan,Randolph_Bourne,Ä,Penthesilea,BAFTA_Award_for_
Best_Direction,Confederación_Nacional_del_Trabajo,As_of_2006,Abecedarian,PPG_16,
David_Hume,Lady_for_a_Day,Talladega_Superspeedway,Kant,1967,Battlefield_Earth_(n
ovel),Robin_Hood_(1922_film),Battle_of_Stillman's_Run,Randolph_Bourne,North_Alab
ama,Cut_(archaeology),bank,Midwest_U.S.,Golden_Globes,International_Phonetic_Alp
habet,vineyard,archaeological_culture,autosome,April_19,strike_action,free_rider
,Algeria,archaeological_culture
```

要求能够利用 PageRank 算法的思想计算出每个网页的 PR 值（迭代 10 次即可），在伪分布式环境下完成程序的编写和测试。

2）在集群上提交作业并执行

集群的服务器地址为 10.102.0.197，用户名和密码为自己的学号，用户主目录为/home/用户名，hdfs 目录为/user/用户名。集群上的数据集存放目录为 hdfs://master:9000/Experiment_3/.

提交到 Hadoop 的具体步骤如下：

(1)使用 scp PageRank_Hadoop.jar 用户名@10.102.0.197:/home/用户名 命令将本地程序提交到 Hadoop 集群，通过 ssh 用户名@10.102.0.197 命令远程登录到 Hadoop 集群进行操作；

(2)使用 hadoop jar PageRank_Hadoop.jar 命令在集群上运行 Hadoop 作业，在程序中指定输出目录为自己 hdfs 目录下的 Experiment_3_Hadoop；

提交到 Spark 的具体步骤如下：

(1)使用 scp PageRank_Spark.jar 用户名@10.102.0.197:/home/用户名 命令将本地程序提交到集群，通过 ssh 用户名@10.102.0.197 命令远程登录到 Hadoop 集群进行操作；

(2)进入 spark 的 bin 目录下，使用./spark-submit --class "包名.主类名" PageRank_Spark.jar 命令运行 Spark 作业。Spark 程序中将结果保存到集群 Experiment_3_Spark 目录下

(3)在浏览器中打开 http://10.102.0.197:50070，可以查看集群的基本信息以及 hdfs 目录；在浏览器中打开 http://10.102.0.197:8088，可以查看集群上作业的基本执行情况。

四、实验要求

实验结果为连续迭代 10 次后的输出结果

两个程序输出的结果格式均为(PR 值,网页)，其中 PR 值保留 10 位小数，部分结果如下所示：

```
(recovered_factory,1.3190784132)
(AbboT,1.3097577719)
(Abbess,1.2965970002)
(Asteropaeus,1.2749416435)
(Best_Adapted_Screenplay,1.2726999836)
(For_Your_Consideration,1.2719216982)
(Martin_Scorsese,1.2695332683)
(Whittaker_Chambers,1.2605422944)
(capitalism,1.2596672758)
(Maurice_Champagne,1.2547214604)
```

使用hdfs dfs -get命令将自己的运行结果与hdfs上/Experiment_3_result目录的标准结果保存到本地，使用 diff 命令判断程序执行结果差异，结果正确无误，方可提交。

五、实验报告

计算机科学与技术学院 大数据管理与分析 课程实报告

实验题目：		学号：201500000000
日期：2018.3.20	班级：2015 级 1 班/菁英班	姓名：张三
Email：zhangsan@qq.com		
实验目的：		
实验软件和硬件环境：		
实验原理和方法：		
实验步骤：（不要求罗列完整源代码）		
结论分析与体会：		

就实验过程中遇到和出现的问题，你是如何解决和处理的，自拟 1－3 道问答题：