

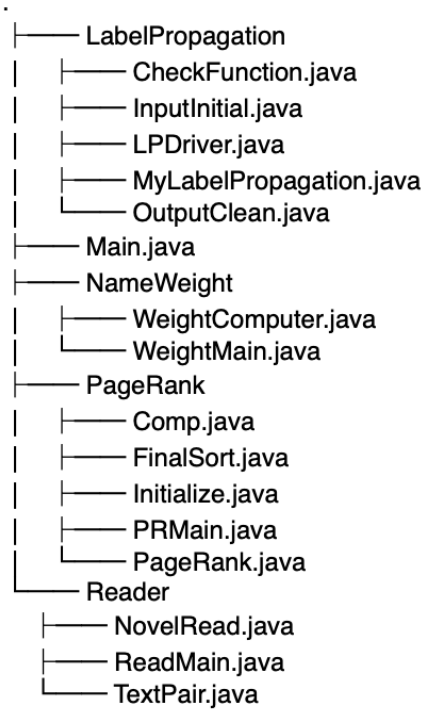
大数据课程设计实验报告

181850055 黄嘉怡 Task1 2 3

181860127 姚逸斐 Task4

181870290 周心瑜 Task5

项目代码结构



`Reader`：读取文本，提取名字，进行人物同现关系统计。

`NameWeight`：计算归一化的关系权重。

`PageRank`：根据人物关系权重计算PageRank值。

`LabelPropagation`：根据人物关系权重进行标签传播。

`Main.java`：整合函数。

任务1 数据预处理

任务简介

从原始的哈利波特系列小说文本中，抽取与人物互动相关数据，并屏蔽与人物姓名无关的文本内容。

设计思路

本次任务完成内容比较简单，为了提高效率，与任务2在一次mapreduce中完成。设计思路见任务2。

任务2 特征抽取：人物同现统计

任务简介

根据提取任务1提取的人名列表，统计在原文中姓名对的同现次数。

设计思路

- Setup
读取用户自定义文本，形成自定义字典DicLibrary
- Map

```
Input key : OFFSET
Input value : 一行小说内容

Output key : <name1,name2>
Output value : 1
```

调用 `DicAnalysis.parse(String)` 函数得到分词结果；

对分词结果进行一次循环，提取并存储符合用户自定义词典中的内容；对同一人物的重名处理也在此进行，判断姓名重复时，将提取的内容替换为同一个；

对任意两个不同姓名对之间进行统计并传出。

- Combine
由于一段文本中有联系紧密的人物关系，很有可能出现多次姓名对，因此为了提高分析效率加入Combiner，合并重复姓名对的次数。
- Reduce

```
Input key : <name1,name2>
Input value : times1,times2,...

Output key : name1,name2,times
Output value : NULL
```

累加value中的次数，合并后传出。

程序运行截图

1	丁沃斯, 上弗莱格利, 1
2	丁沃斯, 奥特里-圣卡奇波尔, 1
3	丁沃斯, 戈德里克, 1
4	丁沃斯, 戈德里克·格兰芬多, 1
5	丁沃斯, 比尔, 1
6	丁沃斯, 罗恩, 1
7	丁沃斯, 芙蓉, 1
8	丁沃斯, 鲍曼·赖特, 1
9	上弗莱格利, 丁沃斯, 1
10	上弗莱格利, 奥特里-圣卡奇波尔, 1
11	上弗莱格利, 戈德里克, 1
12	上弗莱格利, 戈德里克·格兰芬多, 1
13	上弗莱格利, 鲍曼·赖特, 1
14	丹尼斯·克里维, 1
15	丹尼斯, 哈利, 2
16	丹尼斯, 尼克, 1
17	丹尼斯, 科林, 2
18	丹尼斯, 罗恩, 1
19	丹尼斯·克里维, 乔治, 1
20	丹尼斯·克里维, 凯蒂·贝尔, 1
21	丹尼斯·克里维, 卢娜·洛夫古德, 1
22	丹尼斯·克里维, 厄尼·麦克米兰, 1
23	丹尼斯·克里维, 哈利, 8
24	丹尼斯·克里维, 安东尼·戈德斯坦, 1

任务3 特征处理：人物关系图构建与特征归一化

任务简介

根据任务2统计的人物共现次数，生成归一化权重后的人物关系图的临接表表示。

在人物关系图中，人物是顶点，人物之间的互动关系是边，人物之间的互动关系通过人物的共现关系体现。

设计思路

- Map

```
Input key : OFFSET
Inout value : name1,name2,times

Output key : name1
Output value : name2,times
```

获取姓名对和对应一个互动关系的出现次数。

- Reduce

```
Input key : name
Input value : [name1,times],[name2,times]...

Output key : name name1,weight1;name2,weight2;...
Output value : NULL
```

统计人物对应的互动关系总数；

重新计算所有互动关系占总数的比值并插入字符串。

程序运行截图

```
1 丁沃斯 上弗莱格利,0.125;戈德里克·格兰芬多,0.125;罗恩,0.125;比尔,0.125;戈德里克,0.125;鲍曼·赖特,0.125;芙蓉,0.125;奥特里—圣卡奇波尔,0.125;
2 上弗莱格利 奥特里—圣卡奇波尔,0.2;戈德里克,0.2;戈德里克·格兰芬多,0.2;鲍曼·赖特,0.2;丁沃斯,0.2;
3 丹尼斯 克里维,0.14285714285714285;科林,0.2857142857142857;尼克,0.14285714285714285;哈利,0.2857142857142857;罗恩,0.14285714285714285;
4 丹尼斯·克里维 凯蒂·贝尔,0.021739130434782608;卢娜·洛夫古德,0.021739130434782608;厄尼·麦克米兰,0.021739130434782608;哈利,0.17391304347826086;
5 丹尼斯·毕肖普 汤姆,0.3333333333333333;汤姆·里德尔,0.3333333333333333;艾米·本森,0.3333333333333333;
6 丽塔·斯基特 卡卡洛夫,0.013392857142857142;乌姆里奇,0.004464285714285714;乔治,0.004464285714285714;亚瑟,0.004464285714285714;伯莎·乔金斯,0.0
7 丽塔斯基 哈利,1.0;
8 乌姆里奇 艾克莫,0.00507185122569738;丽塔·斯基特,8.453085376162299E-4;乔治,0.02282333051563821;伦考恩,8.453085376162299E-4;傲罗,0.001690617
9 乔丹 莱特林,1.0;
10 乔治 斯克林杰,9.779951100244498E-4;丹尼斯·克里维,4.889975550122249E-4;丽塔·斯基特,4.889975550122249E-4;乌姆里奇,0.013202933985330073;乔治韦斯
11 乔治·韦斯莱 邓布利多,0.046511627906976744;凯蒂·贝尔,0.06976744186046512;卢平,0.023255813953488372;哈利,0.11627906976744186;安吉丽娜,0.02325
12 乔治韦斯莱 克鲁克山,0.020202020202020204;凯蒂,0.010101010101010102;博尔,0.030303030303030304;哈利,0.10101010101010101;奇洛,0.010101010101
13 乔艾·詹肯斯 哈利,1.0;
14 乔雷德 乔治,1.0;
15 书斯莱 伍德,0.02564102564102564;傲罗,0.02564102564102564;卢修斯马尔福,0.02564102564102564;哈利,0.2564102564102564;马尔福,0.025641025641025
16 亚瑟 比尔,0.005494505494505495;丽塔·斯基特,0.005494505494505495;乔治,0.005494505494505495;亚瑟·韦斯莱,0.01098901098901099;佩恩,0.005494505
17 亚瑟·韦斯莱 卢修斯,0.0967741935483871;亚瑟,0.03225806451612903;哈利,0.11290322580645161;哈利·波特,0.04838709677419355;唐克斯,0.01612903225
18 伊万斯 斯内普,0.3;哈利,0.1;詹姆,0.6;
19 伊万诺夫 沃尔科夫,0.3333333333333333;瑞安,0.3333333333333333;沃卡诺夫,0.3333333333333333;
20 伊凡·迪隆斯 巴希达,0.5;艾妮·斯米克,0.5;
21 伊戈尔 卡卡洛夫,0.3333333333333333;邓布利多,0.5;霍格,0.16666666666666666;
22 伊戈尔·卡卡洛夫 雷古勒斯,0.125;克劳奇,0.25;哈利,0.25;小天狼星,0.125;邓布利多,0.125;阿兹卡班,0.125;
23 伊格诺图斯 佩弗利尔,0.125;哈利,0.125;安提俄克·卡德摩斯,0.125;戈德里克,0.25;谢诺菲留斯,0.125;邓布利多,0.25;
24 伊格诺图斯·佩弗利尔 哈利,0.1111111111111111;小天狼星,0.1111111111111111;戈德里克,0.3333333333333333;谢诺菲留斯,0.1111111111111111;赫敏,0.1111111111111111;
```

任务4 数据分析：基于人物关系图的PageRank计算

任务简介

对于任务3输出的归一化权重后的任务关系图，进行数据分析，计算PageRank值（后文简称为PR值），并对人物的PR值进行全局排序，从而定量地分析出哈利波特系列小说的“主角”们是哪些。

设计思路

本任务的解决方法可以划分为3个阶段，分别是初始化各人物的PR值，迭代计算各人物的PR值，排序最终的PR值并输出。

初始化各人物的PR值

任务3的输出格式为 `name name,weight;name,weight;...`，不便于计算PR值，第一次初始化所有人物PR值为1。

- Map思路
Map接收任务3的输入，发送的key为name，value为空格分隔符之后的list。
- Reduce思路
Reduce将作为key的人物PR值设为1，发送的key依旧为原来的name，发送的value在原先value头部加入1和分隔符“#”。

迭代计算各人物的PR值

- Map思路

遍历每条的value，计算每条作为key的人物对list中各人物的PR值的贡献值，也就是当前key人物的PR值乘以list中对应的weight，发送的key为list中的对应人物，value为原先作为key的人物对其的贡献值格式为 `name*#contribution`，同时也需要发送 `name name,weight;name,weight...` 此格式的信息便于后续迭代。

- Reduce思路

如果接收到 `name*#contribution` 的信息，则累加起来得到sum，并在最后引入阻尼系数（默认为0.85）计算：

$$1 - 0.85 + 0.85 * \text{sum}$$

如果接收到 `name name,weight;name,weight...`，则获取与该人物相关联的list。

最后发送key为该人物，value为`newPR+#list`，格式为 `name newPR#name,weight;name,weight...`

排序最终的PR值并输出

- Map思路

负责将 `name PR#name,weight;name,weight...` 转换成 `name PR`，发送的key为PR，value为name。

- Reduce思路

接收的key为PR，value为相同的PR值的人物list，发送的key为PR，value为人物。

最后需要对key也就是PR进行全局排序，在这里自定义Comparator使之为降序排序。

程序运行截图

```

njucs@njucs-VirtualBox:~/hadoop_installs/hadoop-2.7.1$ bin/hadoop dfs -cat /testmodi/FinalRank/part-r-00000
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

24.38163248871616      哈利
9.645302784424835     罗恩
7.00213692079063      赫敏
5.72144710232315      海格
4.997872759010512     格兰芬
4.857721064717463     斯内普
4.386461233636157     达力
3.9458131468296584     纳威
3.7818696317291187     马尔福
3.7767323309339615     邓布利多
3.411976747585627     莱特林
2.5336363441957563     霍格
2.4349754068512732     弗农
2.343312343800337     奇洛
2.171214792304675     德思礼
1.7838235411824193     佩妮
1.5790586346651336     弗雷德
1.3477029173458681     费尔奇
1.2566645232931157     宾斯
1.1241227663339441     伍德
1.043943916683736     弗立维
1.0      巴沙特
1.0      巴希达
0.9731275394662837     赫奇帕奇
0.8962455465997481     哈利波特
0.8772692541949383     乔治
0.871667738353442     墨瑞克
0.871667738353442     尤里克
0.8677412015887723     拉文克劳
0.8090087612981932     高尔
0.807177183050645     查理
0.7523349023354657     约翰逊
0.7469456850385204     皮尔
0.7359689113220993     弗林特
0.7220788064486794     佩蒂尔
0.6613278433702885     洛丽丝
0.6149119819299329     迪安
0.6006736003062257     乔治·韦斯莱
0.5904379245952499     费伦泽

```

任务 5 数据分析：在人物关系图上的标签传播

任务介绍

标签传播(Label Propagation)是一种半监督的图分析算法，他能为图上的顶点打标签，进行图顶点的聚类分析，从而在一张类似社交网络图中完成社区发现(Community Detection)。

设计思路

在设计算法时，标签传播任务可以分为三个子任务：初始化信息处理、标签传播算法计算出最终信息、处理人物最终的标签输出最终格式。

初始化信息处理

- Map思路

Map将任务3的输入划分为每一条边的单独信息并给邻居结点附上初始化标签信息。Map的输出key为本结点的name，value为邻居节点的信息，可以记为 `<neighbor_info>`，格式如下：

```
<name>,<label>,<weight>;
```

其中name为人物的名称，每一个人物的label初始化值为自身的name，weight为任务三中输入的权重。

- Reduce思路

Reduce把所有邻居结点（与本结点人物有关系的人物）信息进行整合，把无向图的信息以以下格式保存：

```
<label>&<neighbor_info><neighbor_info>...&<max_weight>
```

其中label为本结点的初始化标签，即自身的name；neighbor_info为Map中的输出的value；初始化的最大权重表示的含义为此结点标签接管本结点的权重，初始值为0。

标签传播算法

因为异步标签传播的算法对于处理标签信息的步骤有较强的依赖关系，所以我们采用同步算法设计Hadoop框架下的具体实现。同时，还需要考虑到同步算法在一定的图中会产生振荡而无法收敛，故引入 `MAX_EPOCH_NUM` 限制最大迭代次数。其中，当epoch为t时，停止迭代的条件为：

$$\forall i, weight_t^{(i)} \leq weight_{t-1}^{(i)}$$

其中， $weight_t^{(i)}$ 表示第t次迭代，人物i标签所占的权重。

- Map思路

计算当前人物关系中所有标签的权重记录在HashMap中。找到最大支配的标签和标签权重，将人物信息中label与max_weight两项值更新，将新的无向图信息输出给当前结点。key为当前人物name，value为更新后的无向图信息。

同时，将自身结点更新的标签信息输出给所有邻居结点，便于在reduce过程中更新邻居结点标签信息。key为所有邻居结点的name，value格式如下：

```
<name>/<new_label>
```

其中name为当前人物名称，new_label为当前人物新的标签。

- Reduce的思路

通过特殊字符 `/` 来区分接受的消息为标签更新消息还是无向图边信息。更新标签后输出key为当前人物name，value为更新过后的无向图边信息。

最终的标签输出

- Map思路

将最终的无向图信息转变为最终输出格式，格式如下：

```
<name> <label>
```

输出key为name，value为label。

- Reduce思路

将接收到key直接输出，value为list但只有一个元素即为当前人物的最终label。

平台运行截图

MapReduce运行记录：

	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress	Tr
8921	nju_st37	MyLabelPropagation	MAPREDUCE	root.2021team37	Mon Jul 26 00:52:18 +0800 2021	Mon Jul 26 00:52:34 +0800 2021	FINISHED	SUCCEEDED		Hi
8920	nju_st37	LPInitialize	MAPREDUCE	root.2021team37	Mon Jul 26 00:52:01 +0800 2021	Mon Jul 26 00:52:16 +0800 2021	FINISHED	SUCCEEDED		Hi
8919	nju_st37	PageRank Final Sort	MAPREDUCE	root.2021team37	Mon Jul 26 00:51:40 +0800 2021	Mon Jul 26 00:51:57 +0800 2021	FINISHED	SUCCEEDED		Hi
8918	nju_st37	PRIter,out:/user/nju_st37/FinalLab/Task4//Data10	MAPREDUCE	root.2021team37	Mon Jul 26 00:51:20 +0800 2021	Mon Jul 26 00:51:36 +0800 2021	FINISHED	SUCCEEDED		Hi
8917	nju_st37	PRIter,out:/user/nju_st37/FinalLab/Task4//Data9	MAPREDUCE	root.2021team37	Mon Jul 26 00:50:59 +0800 2021	Mon Jul 26 00:51:16 +0800 2021	FINISHED	SUCCEEDED		Hi
8916	nju_st37	PRIter,out:/user/nju_st37/FinalLab/Task4//Data8	MAPREDUCE	root.2021team37	Mon Jul 26 00:50:39 +0800 2021	Mon Jul 26 00:50:54 +0800 2021	FINISHED	SUCCEEDED		Hi
8915	nju_st37	PRIter,out:/user/nju_st37/FinalLab/Task4//Data7	MAPREDUCE	root.2021team37	Mon Jul 26 00:50:18 +0800 2021	Mon Jul 26 00:50:33 +0800 2021	FINISHED	SUCCEEDED		Hi
8914	nju_st37	PRIter,out:/user/nju_st37/FinalLab/Task4//Data6	MAPREDUCE	root.2021team37	Mon Jul 26 00:50:14	Mon Jul 26 00:50:14	FINISHED	SUCCEEDED		Hi
8913	nju_st37	PRIter,out:/user/nju_st37/FinalLab/Task4//Data5	MAPREDUCE	root.2021team37	Mon Jul 26 23:03:53 +0800 2021	Mon Jul 26 23:04:12 +0800 2021	FINISHED	SUCCEEDED		Hi
89148	nju_st37	PRIter,out:/user/nju_st37/FinalLab/Task4/Data2	MAPREDUCE	root.2021team37	Mon Jul 26 23:03:32 +0800 2021	Mon Jul 26 23:03:50 +0800 2021	FINISHED	SUCCEEDED		Hi
89147	nju_st37	PRIter,out:/user/nju_st37/FinalLab/Task4/Data1	MAPREDUCE	root.2021team37	Mon Jul 26 23:03:12 +0800 2021	Mon Jul 26 23:03:29 +0800 2021	FINISHED	SUCCEEDED		Hi
89146	nju_st37	Initialize for PageRank	MAPREDUCE	root.2021team37	Mon Jul 26 23:02:52 +0800 2021	Mon Jul 26 23:03:09 +0800 2021	FINISHED	SUCCEEDED		Hi
89145	nju_st37	WeightJob	MAPREDUCE	root.2021team37	Mon Jul 26 23:02:31 +0800 2021	Mon Jul 26 23:02:49 +0800 2021	FINISHED	SUCCEEDED		Hi
89144	nju_st37	ReaderJob	MAPREDUCE	root.2021team37	Mon Jul 26 23:02:08 +0800 2021	Mon Jul 26 23:02:27 +0800 2021	FINISHED	SUCCEEDED		Hi
89143	nju_st37	WeightJob	MAPREDUCE	root.2021team37	Mon Jul 26 23:01:11 +0800 2021	Mon Jul 26 23:01:28 +0800 2021	FINISHED	SUCCEEDED		Hi
89142	nju_st37	ReaderJob	MAPREDUCE	root.2021team37	Mon Jul 26 23:00:49 +0800 2021	Mon Jul 26 23:01:09 +0800 2021	FINISHED	SUCCEEDED		Hi

ReaderJob运行记录：



Application application_1626070675586_9142

Cluster

About

Nodes

Node Labels

Applications

NEW

NEW SAVING

SUBMITTED

ACCEPTED

RUNNING

FINISHED

FAILED

KILLED

Scheduler

Tools

Kill Application

Application Overview

User: nju_st37

Name: ReaderJob

Application Type: MAPREDUCE

Application Tags:

YarnApplicationState: FINISHED

Queue: root.2021team37

FinalStatus Reported by AM: SUCCEEDED

Started: Mon Jul 26 23:00:49 +0800 2021

Elapsed: 19sec

Tracking URL: History

Diagnostics:

Application Metrics

Total Resource Preempted: <memory:0, vCores:0>

Total Number of Non-AM Containers Preempted: 0

Total Number of AM Containers Preempted: 0

Resource Preempted from Current Attempt: <memory:0, vCores:0>

Number of Non-AM Containers Preempted from Current Attempt: 0

Aggregate Resource Allocation: 558316 MB-seconds, 82 vcore-seconds

Show 20 entries

Search:

Attempt ID	Started	Node	Logs	Blacklisted Nodes
appattempt_1626070675586_9142_000001	Mon Jul 26 23:00:49 +0800 2021	http://slave016:8042	Logs	N/A

Showing 1 to 1 of 1 entries

First Previous 1 Next Last

WeightJob运行记录：



Logged in as: dr.who

Application application_1626070675586_8907

Cluster

About

Nodes

Node Labels

Applications

NEW

NEW SAVING

SUBMITTED

ACCEPTED

RUNNING

FINISHED

FAILED

KILLED

Scheduler

Tools

Kill Application

Application Overview

User: nju_st37

Name: WeightJob

Application Type: MAPREDUCE

Application Tags:

YarnApplicationState: FINISHED

Queue: root.2021team37

FinalStatus Reported by AM: SUCCEEDED

Started: Mon Jul 26 00:47:43 +0800 2021

Elapsed: 11sec

Tracking URL: History

Diagnostics:

Application Metrics

Total Resource Preempted: <memory:0, vCores:0>

Total Number of Non-AM Containers Preempted: 0

Total Number of AM Containers Preempted: 0

Resource Preempted from Current Attempt: <memory:0, vCores:0>

Number of Non-AM Containers Preempted from Current Attempt: 0

Aggregate Resource Allocation: 198615 MB-seconds, 34 vcore-seconds

Show 20 entries

Search:

Attempt ID	Started	Node	Logs	Blacklisted Nodes
appattempt_1626070675586_8907_000001	Mon Jul 26 00:47:43 +0800 2021	http://slave020:8042	Logs	N/A

PageRank Final Sort运行截图：



Logged in as: dr.who

Application

application_1626070675586_8919

Cluster

About
Nodes
Node Labels
Applications
NEW
NEW_SAVING
SUBMITTED
ACCEPTED
RUNNING
FINISHED
FAILED
KILLED
Scheduler

Tools

Kill Application

Application Overview

User: nju_st37
Name: PageRank Final Sort
Application Type: MAPREDUCE
Application Tags:
YarnApplicationState: FINISHED
Queue: root.2021team37
FinalStatus Reported by AM: SUCCEEDED
Started: Mon Jul 26 00:51:40 +0800 2021
Elapsed: 16sec
Tracking URL: History
Diagnostics:

Application Metrics

Total Resource Preempted: <memory:0, vCores:0>
Total Number of Non-AM Containers Preempted: 0
Total Number of AM Containers Preempted: 0
Resource Preempted from Current Attempt: <memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt: 0
Aggregate Resource Allocation: 116093 MB-seconds, 31 vcore-seconds

Show 20 entries

Search:

Attempt ID	Started	Node	Logs	Blacklisted Nodes
appatempt_1626070675586_8919_000001	Mon Jul 26 00:51:40 +0800	http://slave011:8042	Logs	N/A

MyLabelPropagation运行截图:



Logged in as: dr.who

Application

application_1626070675586_8921

Cluster

About
Nodes
Node Labels
Applications
NEW
NEW_SAVING
SUBMITTED
ACCEPTED
RUNNING
FINISHED
FAILED
KILLED
Scheduler

Tools

Kill Application

Application Overview

User: nju_st37
Name: MyLabelPropagation
Application Type: MAPREDUCE
Application Tags:
YarnApplicationState: FINISHED
Queue: root.2021team37
FinalStatus Reported by AM: SUCCEEDED
Started: Mon Jul 26 00:52:18 +0800 2021
Elapsed: 16sec
Tracking URL: History
Diagnostics:

Application Metrics

Total Resource Preempted: <memory:0, vCores:0>
Total Number of Non-AM Containers Preempted: 0
Total Number of AM Containers Preempted: 0
Resource Preempted from Current Attempt: <memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt: 0
Aggregate Resource Allocation: 114292 MB-seconds, 30 vcore-seconds

Show 20 entries

Search:

Attempt ID	Started	Node	Logs	Blacklisted Nodes
appatempt_1626070675586_8921_000001	Mon Jul 26 00:52:18 +0800	http://slave013:8042	Logs	N/A