Class 3 Homework

Realtime and Big Data Analytics

Homework

Class 3

1. **Optional reading**: “The PageRank citation ranking: Bringing order to the Web”, by Page, Lawrence; Brin, Sergey; Motwani, Rajeev and Winograd, Terry (1999) - http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

2. **Programming assignment using MapReduce: PageRank Problem - Please complete this homework <u>independently</u>.**
   PageRank is an algorithm used by the Google web search engine to rank websites in their search engine results. It is named after Larry Page, although many people think it means "rank of pages".

   You can see a brief introduction at: http://en.wikipedia.org/wiki/PageRank. This is a long page to read, but you don't need to understand every detail here. The important part is the algorithm: http://en.wikipedia.org/wiki/PageRank#Algorithm. This is all you need to know about PageRank for this homework assignment. You can skip the matrix and algebraic parts. Focus on how to get PR(A) from PR(B), PR(C) and PR(D).

   Basically, PageRank is trying to do this: Distribute the page's own PR value to all of the linked pages iteratively, and finally get a stable state, which presents the theoretical PR values of all pages. As described on the wiki, you can transfer PR/outlinks of a page to all linked pages, and you can also add a damping factor. Let's ignore the damping factor, focus on the formula:
   PR(A) = PR(B)/2 + PR(C) + PR(D)/3

   Your task is to implement a simplified PageRank with MapReduce.

   Assume that we have the following small input file of page relationships:

```
A C J 0.166667
B D E J 0.166667
C A B 0.166667
D A B C E J 0.166667
E J 0.166667
J B C 0.166667
```

   The first line - `A C J 0.166667` - is interpreted as follows:

   “`A`” means "Page A“.

   "`C J`" means "Page A" has outlinks to "Page C" and "Page J“.

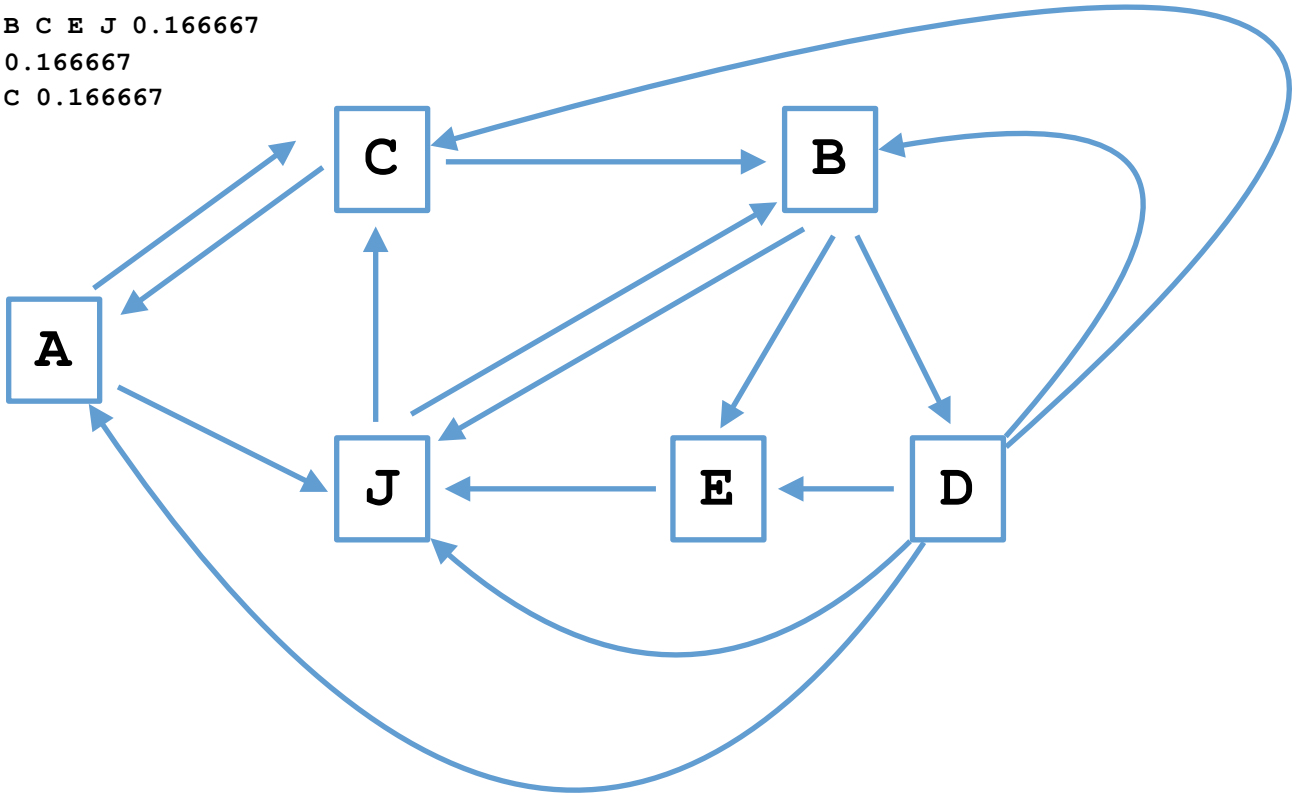
   "`0.166667`" is the initial PR value of Page A.

Homework

Class 3

We can depict the network graphically:

```
A C J 0.166667
B D E J 0.166667
C A B 0.166667
D A B C E J 0.166667
E J 0.166667
J B C 0.166667
```



Homework

Class 3

2. Programming assignment using MapReduce: PageRank Problem (continued)

Remember, this is a DIRECTIONAL graph, i.e. links have direction. For instance, "A C J" means A has outlinks to C and J, and "B D E J" means B has outlinks to D, E and J.

After you read in this input, your MR job should parse and process the data, and output the PR value for ONE iteration. This means you only need to use the formula on the data ONCE.

Your output file should look like this, where *PR* is the pagerank value computed by your program (notice that your output is formatted the same way as your input.- that's so you can easily run more iterations … see extra credit on next page):

```
A C J PR
B D E J PR
C A B PR
D A B C E J PR
E J PR
J B C PR
```

Remember you MUST output the Page, the outlinks, and the new PR value as shown above. This will be useful if you want to investigate iterations because since the output format matches the input format, the program need not be modified.

For submission, please pack your code and output files together (no libraries and .class files) and submit to NYU Classes. Include a screenshot as in previous homework.

Homework

Class 3

2. Programming assignment using MapReduce: PageRank Problem (continued)

HINTS

Here are some hints on how to write this MapReduce job…

Your Map jobs should output key-value pairs as follows:
outlink_target source_page, PR/number_of_outlinks

You can also output the original outlinks information in case you need it. For the first input line (`A C J 0.166667`) it would be:

```
A C J
```

For the first line in the input (`A C J 0.166667`), your Map job could output the following, for example:

Key	Value
C	A, PR/2
J	A, PR/2
A	C J

Where the key in the first line is *C* and the value in the first line is this entire string: *A, PR/2*

Therefore, the Reducer step will see the data formatted as follows coming in from the Map step for the key value C:

Key	Value
C	A, PR1
C	J, PR2
C	D, PR3
C	A B

Finally, the Reducer should be able to compute the PR value of C, for example, by just summing each PR value.

Your final output will be formatted just like the input file (the input to the map phase). For example, one of your output lines will be formatted like this: A C J 0.123456

Homework

Class 3

3. PageRank Problem: *Extra Credit*

You may have noticed that our input and output files have the same format. This means that you should be able to write a program to iteratively read input from the previous output, and then figure out the next step PR value.

Model an iterative MapReduce program by calling your MapReduce jobs three times from command line to figure out PageRank of 3 steps. Adjust the input and output directories as needed so that the output produced by the first job becomes the input to the second job, etc.

For submission, please pack your code and output files together (no libraries and .class files) and submit to NYU Classes. Include a screenshot.