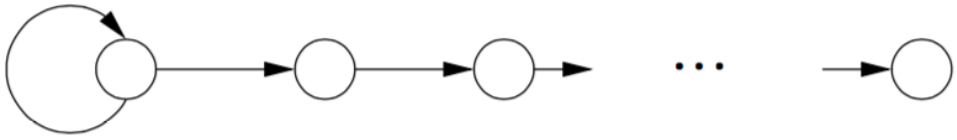
CSC 555: Assignment 6
Sarah Cummings

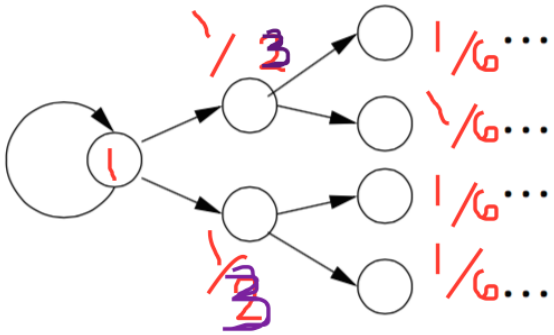
1) **a) Exercise 5.1.6:** Suppose we recursively eliminate dead ends from the graph, solve the remaining graph, and estimate the page rank for the dead end pages as descried in 5.1.4. Suppose the graph is a chain of dead ends, headed by a node with self loop. What would be the page rank associated with each of the nodes?



Call our initial node with self loop A, with the following nodes in a chain of dead ends called B, C, D… respectively. Removing all the dead ends, A has a page rank of **1** since M= [1] and v= [1]. Now solving for B, there is a 50 percent of the traffic from A will self loop back to A and 50 percent will go to B. Thus the page rank of B is 1* 1/2= **1/2**

Similarly, all of the other dead end nodes past B would also have a page rank of 1/2 if solving recursively.

b) Repeat 5.1.6 for a tree of dead ends.



The page rank for our initial node is **1**.
The page rank for our second level nodes are 1/3 and **1/3**
The page rank for the four nodes in the third level of the tree are each 1/3* 1/2= **1/6**
The next level of nodes (of which there are 8 nodes) would each have a rank of 1/6* 1/2= **1/12.**
To find the page rank of any successive node, compute 1/3 * (1/2)^(n-2) where n is the level of the node

2) Exercise 9.3.1: Given the utility matrix below, compute the following:

	a	b	c	d	e	f	g	h
A	4	5		5	1		3	2
B		3	4	3	1	2	1	
C	2		1	3		4	5	3

a) Treating the matrix as Boolean, compute the Jaccard distance between each pair of users

A and B: i= 4 u=8 **jd= 4/8= 1/2**
A and C: i= 4 u=8 **jd= 4/8= 1/2**
B and C: i= 4 u=8 **jd=4/8= 1/2**

e) Normalize the matrix by subtracting the average rating of the user for each non-blank entry:
average for A: 3.333 average for B: 2.333 average for c: 3

new matrix:

	a	b	c	d	e	f	g	h
A:	0.666,	1.666,		, 1.666,	-2.333,		, -0.333,	-1.333
B:		, 0.666,	1.666,	0.666,	-1.333,	-0.333,	-1.333,	
C:	-1,		, -2,	0,		1,	2,	0,

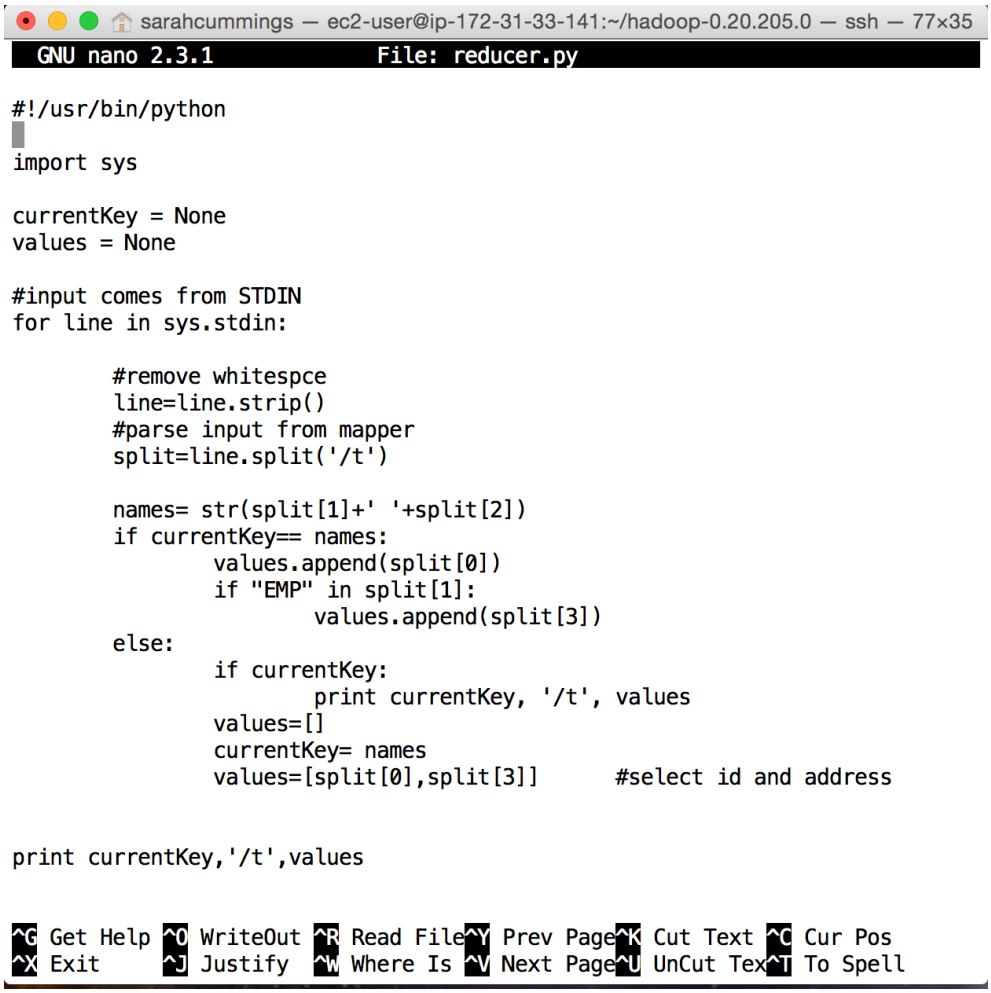
3) Custom map reduce job

cd $HADOOP_HOME
wget http://rasinsrv07.cstcis.cti.depaul.edu/CSC555/employee.txt
wget http://rasinsrv07.cstcis.cti.depaul.edu/CSC555/customer.txt

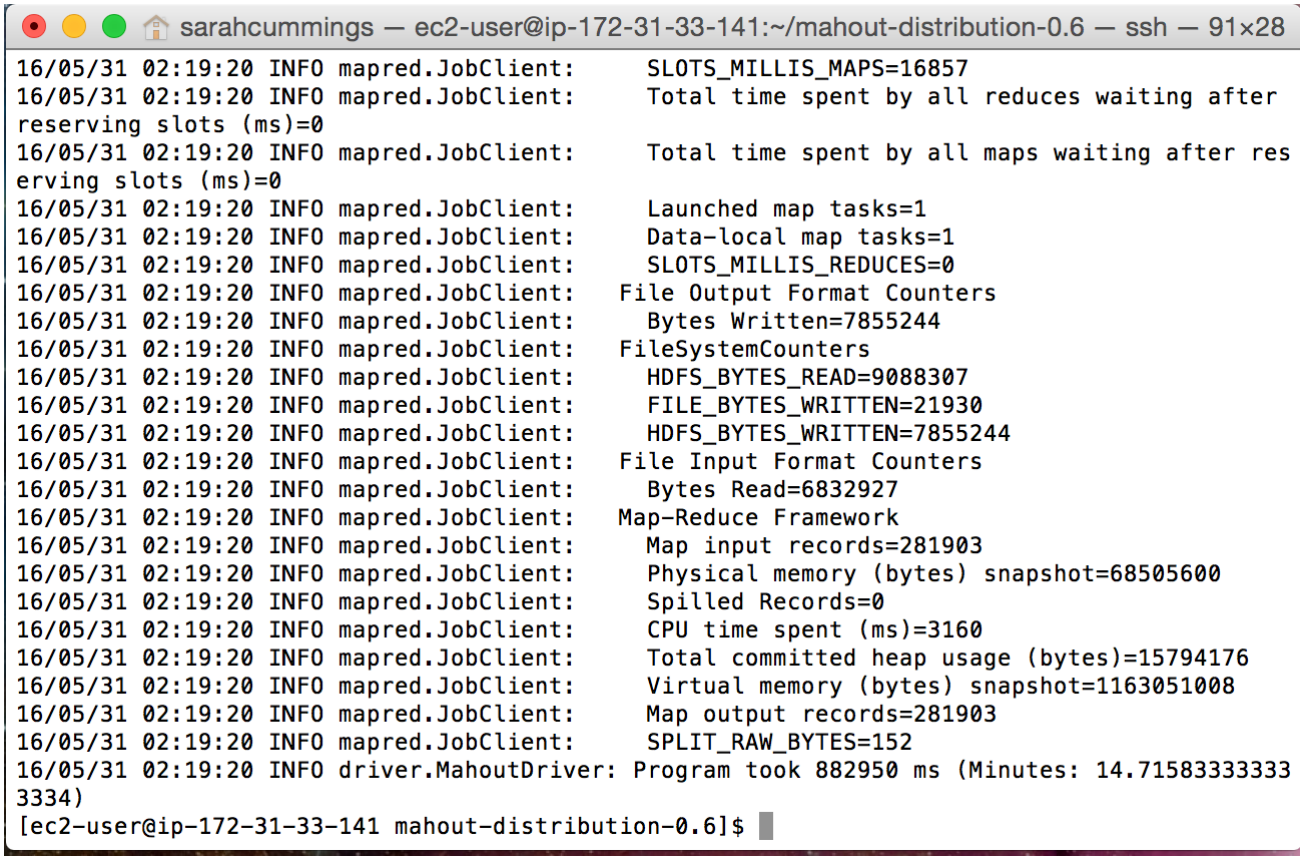
bin/hadoop fs -mkdir joinDir
bin/hadoop fs -put employee.txt customer.txt joinDir/

nano mapper.py
nano reducer.py

bin/hadoop jar contrib/streaming/hadoop-streaming-0.20.205.0.jar -file mapper.py -mapper mapper.py -file reducer.py -reducer reducer.py -input joinDir -output joinResult2


sarahcummings — ec2-user@ip-172-31-33-141:~/hadoop-0.20.205.0 — ssh — 77×35
GNU nano 2.3.1 File: reducer.py
#!/usr/bin/python
import sys
currentKey = None
values = None
#input comes from STDIN
for line in sys.stdin:
#remove whitespce
line=line.strip()
#parse input from mapper
split=line.split('/t')
names= str(split[1]+' '+split[2])
if currentKey== names:
values.append(split[0])
if "EMP" in split[1]:
values.append(split[3])
else:
if currentKey:
print currentKey, '/t', values
values=[]
currentKey= names
values=[split[0],split[3]] #select id and address
print currentKey,'/t',values
^G Get Help ^O WriteOut ^R Read File^Y Prev Page^K Cut Text ^C Cur Pos
^X Exit ^J Justify ^W Where Is ^V Next Page^U UnCut Tex^T To Spell

4) b) time for $MAHOUT_HOME/bin/mahout org.apache.mahout.graph.linkanalysis.PageRankJob --vertices /data/Stanford/web-Stanford_uniquenodes.txt --edges /data/Stanford/web-Stanford.txt --numIterations 20 --output /data/Stanford/PageRank --tempDir temp1



Time: 14.715

Node:	Rank:
119738	5.320979202065959E-7
158282	6.580824496897322E-7
280000	1.1413961373557471E-6

4c) Bayes Classification for news-room 20

`time for:`

time bin/mahout trainclassifier -i 20news-bydate/bayes-train-input -o 20news-bydate/bayes-model-output -type bayes -ng 1 -source hdfs

7.21 minutes as seen below

```
sarahcummings — ec2-user@ip-172-31-33-141:~/mahout-distribution-0.6 — ssh — 114×32
16/06/01 01:09:57 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=73171
16/06/01 01:09:57 INFO mapred.JobClient:     HDFS_BYTES_WRITTEN=932
16/06/01 01:09:57 INFO mapred.JobClient:   Map-Reduce Framework
16/06/01 01:09:57 INFO mapred.JobClient:     Map output materialized bytes=763
16/06/01 01:09:57 INFO mapred.JobClient:     Map input records=310363
16/06/01 01:09:57 INFO mapred.JobClient:     Reduce shuffle bytes=763
16/06/01 01:09:57 INFO mapred.JobClient:     Spilled Records=42
16/06/01 01:09:57 INFO mapred.JobClient:     Map output bytes=10617979
16/06/01 01:09:57 INFO mapred.JobClient:     Total committed heap usage (bytes)=336404480
16/06/01 01:09:57 INFO mapred.JobClient:     CPU time spent (ms)=6180
16/06/01 01:09:57 INFO mapred.JobClient:     Map input bytes=15718807
16/06/01 01:09:57 INFO mapred.JobClient:     SPLIT_RAW_BYTES=388
16/06/01 01:09:57 INFO mapred.JobClient:     Combine input records=310363
16/06/01 01:09:57 INFO mapred.JobClient:     Reduce input records=21
16/06/01 01:09:57 INFO mapred.JobClient:     Reduce input groups=20
16/06/01 01:09:57 INFO mapred.JobClient:     Combine output records=21
16/06/01 01:09:57 INFO mapred.JobClient:     Physical memory (bytes) snapshot=492883968
16/06/01 01:09:57 INFO mapred.JobClient:     Reduce output records=20
16/06/01 01:09:57 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=3491328000
16/06/01 01:09:57 INFO mapred.JobClient:     Map output records=310363
16/06/01 01:09:57 INFO common.HadoopUtil: Deleting 20news-bydate/bayes-model-output/trainer-docCount
16/06/01 01:09:57 INFO common.HadoopUtil: Deleting 20news-bydate/bayes-model-output/trainer-termDocCount
16/06/01 01:09:57 INFO common.HadoopUtil: Deleting 20news-bydate/bayes-model-output/trainer-featureCount
16/06/01 01:09:57 INFO common.HadoopUtil: Deleting 20news-bydate/bayes-model-output/trainer-wordFreq
16/06/01 01:09:57 INFO common.HadoopUtil: Deleting 20news-bydate/bayes-model-output/trainer-tfIdf/trainer-vocabCou
nt
16/06/01 01:09:57 INFO driver.MahoutDriver: Program took 432650 ms (Minutes: 7.210833333333333)

real    7m15.668s
user    0m5.008s
sys     0m0.728s
[ec2-user@ip-172-31-33-141 mahout-distribution-0.6]$
```

time for : time bin/mahout testclassifier -m 20news-bydate/bayes-model-output -d 20news-bydate/bayes-test-input -type bayes -ng 1 -source hdfs -method mapreduce

3.435 minutes as seen below

```
sarahcummings — ec2-user@ip-172-31-33-141:~/mahout-distribution-0.6 — ssh — 205×42
16/06/01 01:19:19 INFO mapred.JobClient:     Combine output records=230
16/06/01 01:19:19 INFO mapred.JobClient:     Physical memory (bytes) snapshot=5326213120
16/06/01 01:19:19 INFO mapred.JobClient:     Reduce output records=230
16/06/01 01:19:19 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=24388349952
16/06/01 01:19:19 INFO mapred.JobClient:     Map output records=7532
16/06/01 01:19:19 INFO bayes.BayesClassifierDriver: ==================================================
Confusion Matrix
-------------------------------------------------------
a       b       c       d       e       f       g       h       i       j       k       l       m       n       o       p       q       r       s       t       <--Classified as
381     0       0       0       9       1       0       0       0       1       0       2       0       0       1       0       0       3       0       | 398         a     = rec.motorcycles
1       284     0       0       0       0       1       0       6       3       11      0       3       66      0       1       6       0       4       9       | 395         b     = comp.windows.x
2       0       339     2       0       3       5       1       0       0       0       0       1       1       12      1       7       0       2       0       | 376         c     = talk.politics.midea
st
4       0       1       327     0       2       2       0       0       2       1       1       5       0       1       4       12      0       2       0       | 364         d     = talk.politics.guns
7       0       4       32      27      7       7       2       0       12      0       0       0       6       100     9       7       31      0       0       | 251         e     = talk.religion.misc
10      0       0       0       0       359     2       2       0       1       3       0       6       1       0       1       0       0       11      0       | 396         f     = rec.autos
0       0       0       0       0       1       383     9       1       0       0       0       0       0       0       0       0       0       3       0       | 397         g     = rec.sport.baseball
1       0       0       0       0       0       9       382     0       0       0       0       1       1       1       0       2       0       2       0       | 399         h     = rec.sport.hockey
2       0       0       0       0       4       3       0       330     4       4       0       12      5       0       0       2       0       12      7       | 385         i     = comp.sys.mac.hardwa
re
0       3       0       0       0       0       1       0       0       368     0       0       4       10      1       3       2       0       2       0       | 394         j     = sci.space
0       0       0       0       0       3       1       0       27      2       291     0       25      11      0       0       1       0       13      18      | 392         k     = comp.sys.ibm.pc.har
dware
8       0       1       109     0       6       11      4       1       18      0       98      3       1       11      10      27      1       1       0       | 310         l     = talk.politics.misc
6       0       1       0       0       4       2       0       5       2       12      0       321     8       0       4       14      0       8       6       | 393         m     = sci.electronics
0       11      0       0       0       3       6       0       10      6       11      0       13      299     0       2       13      0       7       8       | 389         n     = comp.graphics
2       0       0       0       0       0       4       1       0       3       1       0       1       3       372     6       0       2       1       2       | 398         o     = soc.religion.christ
ian
4       0       0       1       0       2       3       3       0       4       2       0       12      7       6       342     1       0       9       0       | 396         p     = sci.med
0       1       0       1       0       1       4       0       3       0       1       0       4       8       0       2       369     0       1       1       | 396         q     = sci.crypt
10      0       4       10      1       5       6       2       2       6       2       0       1       2       86      15      14      152     0       1       | 319         r     = alt.atheism
4       0       0       0       0       9       1       1       8       1       12      0       6       3       0       2       0       0       341     2       | 390         s     = misc.forsale
8       5       0       0       0       1       6       0       8       5       50      0       2       40      1       0       9       0       3       256     | 394         t     = comp.os.ms-windows.
misc


16/06/01 01:19:19 INFO driver.MahoutDriver: Program took 206155 ms (Minutes: 3.4359333333333333)

real    3m29.125s
user    0m3.676s
sys     0m0.600s
[ec2-user@ip-172-31-33-141 mahout-distribution-0.6]$
```