Course: CS6200
Homework 3
Name: Rushabh Shah

Github Repository: https://github.ccs.neu.edu/rushabh0812/HW3

Pseudo-Code for Pre-Processing:

```
method map(Key k, value v):
        name =  Parse v to obtain page-name
        adjacencyList = Parse v to obtain adjacencyList
        emit(name, adjacencyList)

        for each record r in adjacencyList:
                emit(r, null);
end

method reduce(pageName p, adjacencyList[ l1, l2, l3, …]):
        pageCountCounter = newGlobalCounter
        if( isInLinksPresent() or isOutLinksresent()):
                pageCountCounter =  pageCountCounter +1

        if(adjacencyList is null):
                emit(p, new adjacencyList[])
                return
        emit(p, adjacencyList)
```

Pseudo-Code for Page Rank:

```
method map(Key k, value v):
    alpha = Initialize alpha component of page rank
    pageCount =  Extract pageCount from counters
    delta = get delta from counters
    emit(k,v)

    newPage = v.pageRank + (1-alpha)*delta/pageCount
    for each entry e in v.adjacencyList:
        emit(e, newPageRank/sizeOf(adjacencyList))

    if(v.adjacencyList is empty):
        emit(dummy, newPageRank)

method reduce(Node n,  List[dummy, c1, c2, c3]):
    if(n is dummy):
        for each contribution 'c' in list:
            totalContributions += c
        update delta global counter with totalContributions
        return

    Node n1 = Initialize new node
    for each entry e in List:
        if(e in adjacencyList):
            n1.adjacencyList = e
        else:
            totalContributions += e

    n1.pageRank = totalContributions
    emit(n,n1)
```

Pseudo Code for TopK records:

```
Class Mapper{
localTopK

setup(){
        initialize localTopK
}

map(… , x){
if (x is in localTopK)
        // Adding x also evicts the now (k+1)-st record from localTopK
        localTopK.add(x)
}

cleanup(){
        for each x in localTopK
        emit(dummy, x)
}

reduce(dummy,  [x1, x2, x3….]){
        initialize globalTopK

        for each record x in input list
        if (x in gloablTopK)
        // Adding x also evicts the now (k+1)-st record from globalTopK
        globalTopK.add(x)

        for each x in localTopK
        emit(NULL, x)

}
```

Performance Comparison

Time For	6 m4.large machines	11 m4.large machines
Pre-Processing	1562206	801178
10 iterations of Page Rank	1184316	894631
Top-100 pages	52778	51820

Report for both configurations the amount of data transferred from Mappers to Reducers, and from Reducers to HDFS, separately for each iteration of the PageRank computation. Does it change with the cluster size? Does it change over time?

6m4.large syslog screenshot:

```
        Total megabyte-milliseconds taken by all reduce task
Map-Reduce Framework
        Map input records=2292111
        Map output records=68942930
        Map output bytes=3601269553
        Map output materialized bytes=1121059069
        Input split bytes=2628
        Combine input records=85621388
        Combine output records=31648992
        Reduce input groups=2292112
        Reduce shuffle bytes=1121059069
        Reduce input records=14970534
        Reduce output records=2292111
        Spilled Records=46619526
        Shuffled Maps =162
        Failed Shuffles=0
        Merged Map outputs=162
        GC time elapsed (ms)=8743
        CPU time spent (ms)=584790
        Physical memory (bytes) snapshot=16345899008
        Virtual memory (bytes) snapshot=100520697856
        Total committed heap usage (bytes)=15251013632
```

11m4.large syslog screenshot:

```
        Total megabyte-milliseconds taken by all reduce tasks
Map-Reduce Framework
        Map input records=7012253
        Map output records=68931961
        Map output bytes=3613989271
        Map output materialized bytes=1271392380
        Input split bytes=11554
        Combine input records=0
        Combine output records=0
        Reduce input groups=3150469
        Reduce shuffle bytes=1271392380
        Reduce input records=68931961
        Reduce output records=1635970
        Spilled Records=143046037
        Shuffled Maps =2014
        Failed Shuffles=0
        Merged Map outputs=2014
        GC time elapsed (ms)=122021
        CPU time spent (ms)=10696450
        Physical memory (bytes) snapshot=95290146816
        Virtual memory (bytes) snapshot=436883947520
        Total committed heap usage (bytes)=83450396672
```

Metric	**6m4.large**	**11 m4.large**
Map Input Records	2292111	7012253
Map Output Records	68942930	68931961
Reduce Input Records	14970534	68931961
Reduce Output Records	2292111	1635970
HDFS: Number of bytes read	2628	11554
HDFS: Number of bytes written	46619526	143046037

Which of the computation phases showed a good speedup? If a phase seems to show fairly poor speedup, briefly discuss possible reasons—make sure you provide concrete evidence, e.g., numbers from the log file or analytical arguments based on the algorithm's properties.

The speedup for different phases are as follows:

1. Pre-processing: time on 6 machines/11machines = 1562206/801178= 1.94
2. Page Rank: time on 6 machines/11 machines = 1184316/894631 = 1.323
3. Top100: time on 6machines/11 machines = 52778/51820 = 1.018

Top 100 Wikipedia Pages:

Simple Dataset

0.0054537411152722 36,United_States_09d4
0.0041045366155248 93,Wikimedia_Commons_7b57
0.0034045092125098 08,Country
0.0022948577996755944,England
0.0022717376079793 54,Europe
0.0022681508052909 08,United_Kingdom_5ad7
0.0022629779913135284,Water
0.0021719826591977 57,Germany
0.0021688915379444646,France
0.0021386561815742 42,Earth
0.0021307553864373533,Animal
0.0020324385973417907,City
0.0018104100982807851,Week
0.0016861675135594327,Asia
0.0016625841200432 79,Sunday
0.0016376791725900256,Monday
0.0016211380742672439,Wednesday
0.0016067369109369653,Wiktionary
0.0016012250605147568,Money
0.0015812332480585113,Friday
0.0015751561805757337,Plant
0.0015635858325715 62,Saturday
0.0015432469741492245,Thursday
0.0015322597127829711,Tuesday
0.0015226660733272503,Computer
0.0015206333183474597,English_language
0.0014886861646351845,Government
0.0014854208861482282,Italy
0.0014803779262863906,India
0.0013989641624862293,Number
0.0013433209043226485,Spain
0.0013247164099634 96,Day
0.0013067866792937732,Japan
0.0013012872224230385,Canada
0.0012597067672019 87,People
0.0012330199856291776,Human
0.0011992159243370 52,Wikimedia_Foundation_83d9

0.0011855569038321,China
0.0011835802084830167,Australia
0.0011806317748449646,Energy
0.0011492294294854126,Sun
0.001147016711456367 6,index
0.0011432737157235494,Food
0.0011309142327184433,Science
0.001110240554157543,Mathematics
0.0010524948126738657,Television
0.0010341124010405758,Capital_(city)
0.0010272504631897096,Russia
0.0010185780992265463,State
9.962916269389546E-4,Year
9.890574283683957E-4,Music
9.631823330129022E-4,Greece
9.585194243681958E-4,Language
9.565535753398897E-4,Scotland
9.427108993083259E-4,Metal
9.347069737086805E-4,Wikipedia
9.288391777659076E-4,Greek_language
9.197337782593858E-4,Planet
9.084557885467258E-4,2004
8.908705039989364E-4,Sound
8.876963803266156E-4,Religion
8.754673511143459E-4,London
8.663613571025 45E-4,Africa
8.296288659306 56E-4,Geography
8.275228053720926E-4,Law
8.260124904412899E-4,20th_century
8.239106344310348E-4,Liquid
8.095556887868571E-4,19th_century
8.089858293821437E-4,World
7.995000625663615E-4,Society
7.985908636151757E-4,Scientist
7.780969934950783E-4,Atom
7.660343358245363E-4,Latin
7.660326403598595E-4,History
7.599655695752877E-4,Light
7.555621797389816E-4,Sweden
7.551883781897179E-4,Poland
7.545850991653409E-4,War

7.438619807426295E-4,Culture
7.4159703532806 29E-4,Netherlands
7.3188817314301 44E-4,Building
7.1863439199358 27E-4,Turkey
7.1658912287248 82E-4,Plural
7.1491672073115 41E-4,God
7.0880994952577 91E-4,Information
6.9478777832860 79E-4,Centuries
6.9422080275896 03E-4,Chemical_element
6.8906539449021 66E-4,Portugal
6.7846059172640 9E-4,Denmark
6.7042970382779 53E-4,Austria
6.6896403262508 72E-4,Cyprus
6.6624907308942 18E-4,Capital_city
6.6431650562885 47E-4,Ocean
6.5621237855550 36E-4,North_America_e7c4
6.5514990391588 85E-4,Inhabitant
6.5442201594494 51E-4,Moon
6.5313227457873 59E-4,Species
6.5180982147513 61E-4,Disease
6.5080958229296 14E-4,Biology
6.4895882610181 93E-4,Book

Full Data set

0.00229961763254 30014,United_States_09d4
0.00207067574628 88286,2006
0.00109625469698 93415,United_Kingdom_5ad7
9.5034070648680 11E-4,2005
7.2379704011689 99E-4,Biography
7.1264583907500 62E-4,Canada
7.0965854862208 68E-4,England
7.0835865813224 75E-4,France
6.6066208150528 57E-4,2004
6.0463460526431 52E-4,Germany
5.8346746132448 76E-4,Australia
5.6128675836717 42E-4,Geographic_coordinate_system
5.3194043012879 87E-4,2003
5.1464680665487 6E-4,India
5.0481255829450 26E-4,Japan

4.33279513939790 37E-4,Italy
4.26645765558558 84E-4,2001
4.21568270729594 7E-4,2002
4.13489547405617 6E-4,Internet_Movie_Database_7ea7
4.04452910163842 2E-4,Europe
3.99259106965594 5E-4,2000
3.86515649153754 06E-4,World_War_II_d045
3.73901785776255 7E-4,London
3.56199557377535 53E-4,English_language
3.53726975448877 93E-4,Population_density
3.53258021417646 64E-4,1999
3.53175891866077 4E-4,Spain
3.46027900200547 84E-4,Record_label
3.33400792178800 45E-4,Russia
3.30548732877100 1E-4,Race_(United_States_Census)_a07d
3.27363801211693 6E-4,Wiktionary
3.12564769217292 4E-4,Wikimedia_Commons_7b57
3.05039171134665 15E-4,1998
2.91759642679380 03E-4,1997
2.91082979921704 3E-4,Music_genre
2.88512677558281 7E-4,New_York_City_1428
2.87619016673120 2E-4,Scotland
2.73639731768505E-4,1996
2.68919999748903 7E-4,Sweden
2.684842086577E-4,Football_(soccer)
2.66845914856958 94E-4,Television
2.58858700447629 04E-4,Square_mile
2.57817301881391 7E-4,1995
2.56519634812461 5E-4,Census
2.55085637071077 37E-4,California
2.53999728691777 63E-4,China
2.50081584768632 2E-4,Netherlands
2.46856517850542 1E-4,New_Zealand_2311
2.46333208568941 4E-4,1994
2.35241520833838 53E-4,1991
2.32569133861287 47E-4,1993
2.31547764491342 16E-4,1990
2.29860500277395 92E-4,Public_domain
2.29831567916729 52E-4,New_York_3da4
2.22995286781781 19E-4,1992
2.20667904922505 17E-4,United_States_Census_Bureau_2c85

2.190506978619827E-4,Film
2.1646293103969893E-4,Ireland
2.16282183811990 82E-4,Norway
2.1580939558354207E-4,Actor
2.1512975985160752E-4,Scientific_classification
2.099371604986171E-4,Population
2.0952298720396085E-4,1989
2.0769744461113 85E-4,January_1
2.0523016494857353E-4,Latin
2.0516951536134 35E-4,1980
2.0213895815303896E-4,Brazil
2.0182693943882 49E-4,Mexico
2.0095768651471615E-4,Marriage
1.995367176470662E-4,1986
1.966298132821705E-4,French_language
1.9465051955724517E-4,1979
1.9432337003785546E-4,1985
1.937740471605 6828E-4,1982
1.937191174994288E-4,1981
1.9249523714410635E-4,1974
1.9227148925034625E-4,Poland
1.9221462292518752E-4,Politician
1.9011369775103333E-4,South_Africa_1287
1.9009030541808163E-4,Switzerland
1.89768029544612E-4,1984
1.896208630127703E-4,1983
1.8954205245400683E-4,1987
1.8926253265456644E-4,Per_capita_income
1.8769605494478594E-4,1970
1.852752611772E-4,1988
1.8527158953722436E-4,1976
1.8509163911032442E-4,Album
1.8297151685262863E-4,Record_producer
1.8297085586158377E-4,1975
1.810133251335694E-4,1969
1.8055320354770323E-4,Paris
1.798253208501527E-4,Greece
1.7978309989739278E-4,Km²
1.797518484831652E-4,1945
1.7953735757391068E-4,1972
1.7842203913378776E-4,Soviet_Union_ad1f

1.77715620042849 96E-4,1977
1.76727598706007 38E-4,1978
1.74959962235086 56E-4,1973