Harsh Shukla
CS6240-Section-01 Assignment-3
PageRank Using Hadoop

Pseudo Codes:
1. Pre-Processing:
The Pseudo Code for PreProcessing:

```
map(object key,Text value)
        Check for the name pattern using pattern.matcher
        Parse the Text values to get the adjacency List
        Emit(name,adjacencyList)
        For(record record : adjacency list)
                Emit(record,null)
reduce(key pageName, value al[1,2,3])
        Check for dangling nodes
        If (adjacencyList.size>0)or (!danglingFlag && count>1)
        Increment pageCounter ++1;
        Else(if adjacencyLisy == Null)
        Emit(pageName, al).
```

2. PageRank:
The pseudo code for PageRank:

```
Map(key,value);
  int pageCounter ;
  boolean iterator;
  double increments;
  double foo = 0;
  emit(key,value)
  if (iterator== true)
       pageRankForm = 1.0/(pageCounter);
  else(pageRankForm = (value.pageRank+ 0.85 *(increment/pageCounter));
  //dangling node adjustment
  for(String name : value.adjacencyList)
  emit(e,(pagerank/adjacencyList.size)
  if { (adjacencyList== Null)
       emit(noValue,PAgerank)
       }

reduce(key Node, value al[1,2,3])
        if (node is in al)
        for (p :list)
        update each participant p  participant to participant+v;

        New Node{
        DanglingNode==false;
        For (v:al)
        If(entry == adjacencyList)
        n.adjacencyList = v.adjacencyList;
        else participants == participants+entry;
        }
        node1.pageRank == participants;
        emit(node,node1)
```

3. Top-100 Pages.

Mapper(key,value)

      Initialize topknodes;

      Update the top K records

            If topKnodes.size>100

            Remove the last element and update record with the new one;

            Else update the records;

cleanup(){

      (for node: localtopRecords)

      emit(currentVal, topKNodes.size();)

// public void reduce(NullWritable key, Iterable<Text> values,

      Context context) throws IOException, InterruptedException

reducer (currentVal,[node1,node2....)   ;

      initialize topRecords;

      for(node :inputList)

      if (node is in toprecords)

            add the node to top records;//this removes the least ranked page from the list.

      For (node : globalrecords)

            Emit(Null,x);

Report the amount of data transferred from Mappers to Reducers, and from Reducers to HDFS, in each iteration of the PageRank computation. Does it change over time? If so, briefly discuss why or why not?

6 Machines

Iteration	Data transferred from Mappers to Reducers	Data transferred Reducers to HDFS
1	19068781	1446058
2	19277579	1446058
3	19277579	1446058
4	19277579	1446058
5	19277579	1446058
6	19277579	1446058
7	19277579	1446058
8	19277579	1446058
9	19277579	1446058
10	19277579	1446058
11	19277579	1446058

11 Machines

Iteration	Data transferred from Mappers to Reducers	Data transferred Reducers to HDFS
1	17678663	1341659
2	19068926	1446061
3	19277730	1446061
4	19277730	1446061
5	19277730	1446061
6	19277730	1446061
7	19277730	1446061
8	19277730	1446061
9	19277730	1446061
10	19277730	1446061
11	19277730	1446061

The amount of data transferred from Mappers to Reducers in the 1_{st} iteration for 11 machines is
lesser than the other iterations. This is because of the change in the page rank and inlink contributions
passed along for every page . The records remain consistent after 1st iteration in 6 machines while for 11 machines the records
transfer remains consistent throughout.

Performance Comparison:

6Machines:
Preprocessing Time: `926481 milliseconds`
PageRank Time: `610592 milliseconds`
Top100 records :   `38132 milliseconds`

11 Machines:
Preprocessing Time: `497161 milliseconds`
PageRank Time: `528264 milliseconds`
Top100 records :   `37953 milliseconds`

Speedups:
Preprocessing: 1.86 (indicates good parallelism)
PageRank: 1.15 (indicates the advantage of parallelism
Top100 records: 1.004(increasing the number of machine does not have much advantage
here ,this might be beacause of presense of single reducer)

Top 100 Pages:

Local:

0.005453741115272236,United_States_09d4
0.004104536615524893,Wikimedia_Commons_7b57
0.003404509212509808,Country
0.0022948577996755944,England
0.0022717376079793 54,Europe
0.0022681508052909 08,United_Kingdom_5ad7
0.0022629779913135284,Water
0.002171982659197757,Germany
0.0021688915379444646,France
0.00213865618 1574242,Earth
0.0021307553864373533,Animal
0.0020324385973417907,City
0.0018104100982807851,Week
0.0016861675135594327,Asia
0.001662584120043279,Sunday
0.0016376791725900256,Monday
0.0016211380742672439,Wednesday
0.0016067369109369653,Wiktionary
0.0016012250605147568,Money
0.0015812233480585113,Friday
0.0015751561805757337,Plant
0.001563585832571562,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
0.001324716409963496,Day
0.0013067866792937732,Japan
0.0013012872224230385,Canada
0.001259706767201987,People
0.0012330199856291776,Human
0.001199215924337052,Wikimedia_Foundation_83d9
0.00118555669038321,China
0.0011835802084830167,Australia
0.0011806317748449646,Energy
0.0011492294294854126,Sun
0.0011470167114563676,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
8.66361357102545E-4,Africa
8.29628865930656E-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
7.415970353280629E-4,Netherlands
7.318881731430144E-4,Building
7.186343919935827E-4,Turkey
7.165891228724882E-4,Plural
7.149167207311541E-4,God
7.088099495257791E-4,Information
6.947877783286079E-4,Centuries
6.942208027589603E-4,Chemical_element
6.890653944902166E-4,Portugal
6.78460591726409E-4,Denmark
6.704297038277953E-4,Austria
6.689640326250872E-4,Cyprus
6.662490730894218E-4,Capital_city
6.643165056288547E-4,Ocean
6.562123785555036E-4,North_America_e7c4
6.551499039158885E-4,Inhabitant
6.544220159449451E-4,Moon
6.531322745787359E-4,Species
6.518098214751361E-4,Disease
6.508095822929614E-4,Biology
6.489588261018193E-4,Book

EMR:

3.779273385661079E-4,United_States_09d4
2.1961360800193348E-4,2006
1.688477500608259E-4,Biography
1.5456376961334254E-4,United_Kingdom_5ad7
1.2485366074427835E-4,2005
1.1450014030300159E-4,Canada
1.1292553024262114E-4,England
1.0897098086136885E-4,Geographic_coordinate_system
9.594016128910456E-5,Record_label
9.304143265511149E-5,2004
9.075078764675839E-5,Australia
8.42830412960839E-5,France
8.302796159860154E-5,India
8.172930243879152E-5,Germany
7.975179060673083E-5,Internet_Movie_Database_7ea7
7.751374259275342E-5,Music_genre
7.407072766866945E-5,2003
7.06850755673081E-5,Japan
6.377509111102402E-5,Population_density
5.958039393049909E-5,2001
5.8961268992009324E-5,2002
5.514722897190904E-5,Italy
5.5091661878479973E-5,2000
5.438697412504916E-5,Football_(soccer)
5.404710260311193E-5,Album
5.2890511259998522E-5,Television
5.231024765552962E-5,Europe
5.151037243763329E-5,Record_producer
5.053960440393631E-5,World_War_II_d045
5.053185823109073E-5,London
4.9349848132371616E-5,Politician
4.87012008162015E-5,Studio_album
4.797072221644048E-5,1999
4.6725578038166585E-5,Scientific_classification
4.603691763383647E-5,Census
4.6023391941112281E-5,Race_(United_States_Census)_a07d
4.474900507200227E-5,Personal_name
4.4706243031841995E-5,New_Zealand_2311
4.409327408395738E-5,Wiktionary
4.406900055081635E-5,Film
4.385906461533705E-5,English_language
4.2481932111112243E-5,1998
4.118367514466655E-5,Sweden
4.10482142265168E-5,Norway
4.0968898739437186E-5,Per_capita_income
4.092539571159437E-5,Spain
4.070715338915725E-5,Square_mile
4.067216594853491E-5,Russia
4.012588530452524E-5,Public_domain
3.9543924668000145E-5,1997
3.952421138042314E-5,Marriage
3.9112603850044585E-5,Actor
3.868061881835592E-5,Km²

3.845061656847023E-5,1996
3.836016718112002E-5,United_States_Census_Bureau_2c85
3.825028409379724E-5,New_York_City_1428
3.757191821062533E-5,Poverty_line
3.754071295803072E-5,California
3.5889958971462963E-5,Scotland
3.5103943569658265E-5,New_York_3da4
3.465131032156353E-5,1995
3.4524545780853336E-5,White_(U.S._Census)_c45a
3.362720658342716E-5,1994
3.319515482725735E-5,Animal
3.2053863273759436E-5,Brazil
3.148721074127789E-5,Population
3.1269617106195546E-5,1993
3.104738700076478E-5,Netherlands
3.102656420858723E-5,Band_(music)
3.029371420435805E-5,School
2.9926529267682484E-5,Novel
2.9789050970001273E-5,1990
2.9727261804129897E-5,1991
2.9709650626626932E-5,Ireland
2.9459101440418176E-5,1992
2.8823012218077537E-5,Hispanic_(U.S._Census)_1387
2.8820419600042637E-5,Latino_(U.S._Census)_5f0e
2.8644767092584703E-5,Corporation
2.8007302088875648E-5,USA_f75d
2.7981163203521344E-5,Texas
2.7770016899061564E-5,Writer
2.741234736837408E-5,Rock_music
2.7371831228224326E-5,South_Africa_1287
2.7265770924200255E-5,1989
2.7003508867332595E-5,Wikimedia_Commons_7b57
2.692939410669798E-5,Finland
2.6059808518128593E-5,Mexico
2.605578902635106E-5,Company_(law)
2.596836061742034E-5,Ontario
2.5939862814122375E-5,Poland
2.5887945327467267E-5,China
2.5795969646111874E-5,All_Music_Guide_0e49
2.5672241577993552E-5,Native_American_(U.S._Census)_1a7a
2.5621147189738595E-5,Website
2.5529410089230642E-5,1986
2.5503284465668106E-5,1985
2.5379844030955565E-5,1980
2.528539407115206E-5,1988
2.512568409285069E-5,Poet
2.4745355503573036E-5,1987