AYUSH SINGH

CS 6240: Assignment 3

Page Rank

Design Discussion

I followed the 3 step design strategy mentioned in the homework as follows:

Sample Graph: A -> B, C | B -> D | C

Preprocessing: For starters I handled the case where the html file had char '&' which is invalid so I added '&amp;' as per world wide web consortium. Apart from that, for a sample graph here's how we handle parsing from Map to Reduce:

1. **Map** handles two cases based on a line by line basis from sample graph:
   a. Emit all nodes with *maybeDangling* from the list of Page Names for a page
   b. Emit the page itself initializing it with *initPageRank*
2. **Reduce** handles three cases that are handled based on sample graph:
   a. A points to B which is a normal node
   b. A also points to C which is a dangling node
   c. B points to D but D is not present as a node in data i.e. dangling Node
3. **Output format** PageName**\t**AdjacencyList**\t***initPageRank*

PageRank: I used the solution 2 for computing dangling node pagerank share for *i+1* iteration at the end of *i*th iteration phase and making it available to all tasks using hadoop counters. I grasped the concept of PageRank from pseudo code which I found logical but since that did not handling dangling nodes, I started programming map reduce solution for it based off that code and took it from there.

** All elements in a line are guaranteed to be tab separated from parser*

1. **Map** has three global counters (*iterationCount, danglingNodePageRank, initialPageRank*) and two phases with each handling 2 and 3 cases respectively.
   Phase 1 sets initial PageRank based on whether its first iteration or not.
   a. If line has *initPageRank* means its fresh from parser, so set it to *danglingPageRankShare* (0 *initially*) + (1/*totalLinksCount*)
   b. Else add *danglingPageRankShare to* incoming *PageRank*

Phase 2 emits nodes based on certain conditions

a. If this is the 10th iteration, simply emit the key:*pageName value:pageRank*

b. Emit all nodes with key:*pageName* & value:*maybeDangling* from the list of Page Names and emit the page itself initializing it with key:*pageName* & value:*adjacencyList + initPageRank*

c. Lastly emit the *pageName* and its *initPageRank* value and key:*danglingNode* to make sure all danglingNodes reach one reducer to add their share.

2. **Reduce** handles three cases that are handled as follows:

   a. If this is the 10th iteration, simply emit the key:*pageName value:pageRank*

   b. Else if key is *danglingNode* calculate danglingSum and update global counter

   c. Else update the page using formula and update node with its adjList with newPageRank emitting everything in the end.

3. **Output format** PageName**\t**AdjacencyList**\t***initPageRank*

TopK Sort: I based off my sorting from the TopK design pattern mentioned from the book map reduce design patterns because I liked the idea of reducing the number of nodes emitted from mapper itself taking off a huge load from the reducers. On both mapper and reducer I initialized a global TreeMap which only keeps 100 values, in the cleanup phase of Mapper emit all values with a single key and repeat the same in reducer except we would not need the cleanup phase and can directly print the result. Since TreeMap is sorted we automatically get to top 100 values eventually.

Data Transferred in bytes

Iteration	**6 Machines**		**11 Machines**	
#	**Data Transferred**	**to HDFS**	**Data Transferred**	**to HDFS**
1	4118685182	1369947803	4118685182	1369940829
2	4089974794	1369940439	4089920177	1369932598
3	4080629335	1369930967	4080609892	1369922050
4	4091264175	1369952542	4091176536	1369945785
5	4097965550	1369969236	4098002858	1369960882

6	4099643140	1369983251	4099681985	1369976949
7	4102671935	1370018927	4102814350	1370012865
8	4102119710	1370041774	4102238845	1370035358
9	4104556884	1370084079	4104524674	1370077551
10	115111161	115111005	115110611	115110455

The amount of data transferred remains same for both except in the last step and makes sense because in the last iteration all we are doing is simply emitting the incoming pageranks along with their nodes while rest of the time lots of emits take place as discussed in the design discussion problem.

Performance Comparison

Time (*ms)*	**6 Machines**	**11 Machines**
Pre-processing	38m29s	22m7s
PageRank	25m17s	13m56s
Top-100	32s	59s

The first two phases as expected show almost double the speedup but speed-up of Top-100 for 11 machines is half that of 6 machines and after comparing stats from logs is that with 11 machines the numbers of launched map tasks were 19 compared to 9 of 6 machine which means 19*100 records would be emitted from the mapper in total compared to 9*100 of 6 machines which in turn leads to merging of 19 map outputs which eats up lot of time.

Top 100 Pages

Sorted from highest ranking to lowest
The answers seems reasonable looking at the keywords in the sample dataset while in the mega dataset since the dataset belongs to year 2006 I'm pretty happy with the result in 10 iterations placing 2006 on the top with most pagerank.

SIMPLE WIKIPEDIA DATA

United_States_09d4	0.005355777601255395
Country	0.004027824046355395
Wikimedia_Commons_7b57	0.0036022714985541565
Week	0.003084465720751175
Earth	0.0026752943563853387
Water	0.002546000793679085
Europe	0.002523803647437281
Sunday	0.002435140370849413
Monday	0.002384176065598382
Wednesday	0.0023542561216078825
United_Kingdom_5ad7	0.002304059494456286
Friday	0.0022884877073897055
Saturday	0.0022622064161520465
Thursday	0.0022291401408207814
Tuesday	0.0022154232994574677
Day	0.002214672571305371
index	0.0021428657009387743
Asia	0.002048696137474343
Animal	0.002019930329063209
France	0.0018433959710020043
City	0.0017698940247303772
Money	0.0017096034585455699
Government	0.0016776941838347683
Number	0.0016539807641294132
Energy	0.0015556498502864366
Sun	0.0015435842672796775
English_language	0.0015398752446945734
Plant	0.0015367815772914688
England	0.0014954793083519719
India	0.001474593356160982
Germany	0.001465669214320054

```
Italy    0.0013886001687866954
Wiktionary    0.0013628464893483602
Wikimedia_Foundation_83d9    0.0013460253833042647
Computer    0.0013324712512722398
People    0.001305977143147495
Planet    0.0012922282160995502
Science    0.0012772149801701903
Canada    0.0012494104235240542
Human    0.0012150250625922526
State    0.0011455735684420825
China    0.0011447345133376349
Year    0.0011414494449842518
Spain    0.001112032104007919
Wikipedia    0.0010729174818036337
Japan    0.001068713408744579
Mathematics    0.0010629869193989033
Food    0.001059291804649356
Australia    0.0010474679016676523
Geography    0.0010385557642966254
Russia    0.001035641036663431
Greek_language    0.00103330695704128
Capital_(city)    0.001022697843835778
Atom    9.90285448638182E-4
Society    9.627595314902565E-4
Liquid    9.468816031824316E-4
Language    9.435983218665094E-4
Moon    9.263060285517775E-4
Africa    9.220092243932372E-4
Metal    9.11448578054992E-4
World    9.024211689218325E-4
Sound    8.925511641359888E-4
Cyprus    8.884938768969182E-4
Light    8.812600713629492E-4
```

```
Culture    8.795620015871022E-4
Greece    8.762906211813577E-4
History    8.696008691701632E-4
Law    8.670235864565729E-4
Turkey    8.55910398497589E-4
Scientist    8.523872220858339E-4
Plural    8.483304168102193E-4
Religion    8.34546008939885E-4
Scotland    8.321642740565298E-4
Circle    8.105646859580774E-4
Gas    8.017461004974286E-4
2004    7.950566438796031E-4
Ocean    7.741850211831742E-4
20th_century    7.732364121600164E-4
Poland    7.643205099705985E-4
Solid    7.636188505689122E-4
Information    7.627704616486011E-4
Sweden    7.593342065596659E-4
Television    7.582907444708231E-4
Nation    7.496152107354E-4
War    7.429907486244271E-4
Trade    7.40708869926796E-4
Denmark    7.354747520959662E-4
Building    7.32203871729194E-4
19th_century    7.320051566582628E-4
Continent    7.316903496399993E-4
Portugal    7.29182994494692E-4
Electricity    7.136784745130115E-4
Chemical_element    7.08076888666907E-4
Austria    6.882539890516241E-4
Image    6.815065310045475E-4
Republic_of_Ireland_10e7    6.780880414452558E-4
Music    6.748331871234114E-4
```

```
Belgium    6.669926820299989E-4
Time    6.638254426235764E-4
God    6.552185960502633E-4
```

BIG WIKIPEDIA DATA

```
2006    0.0017561678956932896
United_States_09d4    0.0015629256171461271
United_Kingdom_5ad7    8.128548074567297E-4
2005    7.564963469890166E-4
France    5.925301936896899E-4
2004    5.099211551020347E-4
England    4.6864921101578874E-4
Germany    4.673017525794225E-4
Canada    4.5387147485567546E-4
2003    4.0821507598863153E-4
Australia    3.823167537352121E-4
Italy    3.7942937213960887E-4
Japan    3.712414683098072E-4
English_language    3.441478626653587E-4
India    3.306205198168228E-4
World_War_II_d045    3.257223788921249E-4
Europe    3.2413335109059176E-4
Wikimedia_Commons_7b57    3.190315197028015E-4
2002    3.152210480279782E-4
2001    3.067781066910124E-4
Russia    2.976467741991308E-4
London    2.939697118177294E-4
Spain    2.9388671993097197E-4
Wiktionary    2.9087192154469244E-4
2000    2.900563417746145E-4
1999    2.6604268563140676E-4
Geographic_coordinate_system    2.605359080494864E-4
Race_(United_States_Census)_a07d    2.516181276495843E-4
```

```
New_York_City_1428    2.349706752953641E-4
1998    2.2683950071502245E-4
index    2.2537299431849066E-4
1997    2.2229350413132716E-4
Internet_Movie_Database_7ea7    2.1809382069533858E-4
January_1    2.1776223196021512E-4
Latin    2.1405555021091627E-4
Sexagenary_cycle    2.1395666889813969E-4
Netherlands    2.1235611032947364E-4
Population_density    2.1084636247127543E-4
Scotland    2.098679197724891E-4
China    2.0940714138364917E-4
1996    2.067576559341469E-4
French_language    2.0358887219561834E-4
1995    1.9784243755527093E-4
1991    1.9256456201430992E-4
1994    1.9057852019116069E-4
Gregorian_calendar    1.9018501335305366E-4
Biography    1.8987844110374437E-4
Sweden    1.877519843059108E-4
1990    1.85735493794695E-4
Soviet_Union_ad1f    1.8442914940787583E-4
1993    1.7870174784252857E-4
1992    1.743009157291231E-4
1945    1.7355067815388685E-4
Egypt    1.7352763673565312E-4
New_Zealand_2311    1.7283377472129526E-4
1980    1.7093646700823602E-4
California    1.6978941783386627E-4
1989    1.686558190404602E-4
1974    1.669965256442851E-4
Greek_language    1.669815145504117E-4
Square_mile    1.6696225923309852E-4
```

1970	1.6661433343946735E-4
European_Union_e368	1.663306992990674E-4
1979	1.6525851782910593E-4
1986	1.640052552408662E-4
International_Phonetic_Alphabet_96f8	1.6234059557987404E-4
New_York_3da4	1.6218406912468114E-4
1969	1.6165358996418416E-4
1976	1.6109815081919275E-4
1981	1.6020688572732572E-4
1975	1.5964401164388577E-4
Ireland	1.5804692014397817E-4
1982	1.575246188062329E-4
19th_century	1.573149434213807E-4
Public_domain	1.569875990428344E-4
Switzerland	1.5688485519409865E-4
1972	1.5680244312973025E-4
1985	1.5649492178643963E-4
Poland	1.5635162879973753E-4
Television	1.56043197092036E-4
Greece	1.5599289619959325E-4
Portugal	1.554131856898917E-4
People's_Republic_of_China_82bf	1.546906842926289E-4
Paris	1.5466448380756806E-4
Austria	1.5370765775585613E-4
1971	1.5323070188567075E-4
German_language	1.5320009068820422E-4
1973	1.5269889882460117E-4
Mexico	1.524303669140567E-4
1984	1.5184881565541674E-4
1983	1.513047618813404E-4
1977	1.5059998664674107E-4
1968	1.5049201699578754E-4
World_War_I_9429	1.5045643188356077E-4

```
1967    1.4990820328140901E-4
United_Nations_3208    1.4954232897493303E-4
1987    1.4901742052346803E-4
Denmark    1.4872554335581447E-4
South_Africa_1287    1.4799637778879806E-4
Israel    1.46878199854074E-4
```