Assignment 3 - CS 6240

- Sriharsha Srinivasa Karthik Kaipa, Sec 01

Running the code:

The source code for this assignment is available in the folder named "assignment3". The "src" folder contains the final code. Make sure the inputs are in the folder "input" alongside the "src" folder. Run instructions for the code are as follows:

- For cleaning the project, building the jar and then running it from the command line, please use the following rules:
  ```
  :~$ make pagerank
  ```
- For cleaning the project, use the following rule
  ```
  make clean
  ```
- For building the jar, use the following rule
  ```
  make jar
  ```

Pseudo Code:

Pre-processing:

```
Used code provided by professor. No changes.
```

PageRank Mapper:

Mapper Task:
```
map (line) {
    rank = rank + 0.85 * DELTA / COUNT;
    emit(page, "!"); // realpage
    if adjacencyList is empty {
        DELTA += (rank * 100.0);
    }
    for all lint in adjacencyList {
        emit(link, {rank, adjacencyList.length});
    }
    COUNT += 1;
    emit(page, {|, adjacencyList});
}
```

PageRank Reducer:

Reducer Task:
```
reduce (page, list = ["!", {r1, n1}, … , {|, adL}]) {
    for item in list {
        if item is "!" {
            realPage = true;
            continue;
        }
        if item is {|, adL} {
            adjList = adL;
            continue;
        }
        sumPageRank += ri / ni;
    }
    if not realPage
        return;
    rank = 0.85 * sumPageRank + 0.15 / COUNT;
    emit(page, {rank, adjList});
}
```

TopK Mapper:

Mapper Task:
```
setup () {
    HashMap hs = new HashMap();
}
map (line) {
    rank = rank + 0.85 * DELTA / COUNT;
    hs.put(page, rank);
}
cleanup () {
    hs = sortByValue(hs);
    count = 0;
    for each {page, rank} in hs {
        emit(null, {page, rank});
        count++;
        if count >= k
            break;
    }
}
```

TopK Reducer:

<u>Reducer Task:</u>

```
reduce (null, list = [{p1, r1}, {p2, r2}...]) {
    HashMap hs = new HashMap();
    for each {page, rank} in list {
        hs.put(page, rank);
    }
    hs = sortByValue(hs);
    count = 0;
    for each {page, rank} in hs {
        emit(null, {page, rank});
        count++;
        if count >= k
            break;
    }
}
```

Data transfer for 1 master 10 core run (all iterations):

<u>Iteration 1:</u>
Mappers -> Reducers:
Map output materialized bytes=1410334471 (~1.41 GB)
Reducers -> HDFS:
S3: Number of bytes written=1398201399 (~1.398 GB)

<u>Iteration 2:</u>
Mappers -> Reducers:
Map output materialized bytes=2245224243 (~2.245 GB)
Reducers -> HDFS:
S3: Number of bytes written=1398676946 (~1.399 GB)

<u>Iteration 3:</u>
Mappers -> Reducers:
Map output materialized bytes=2253622193 (~2.254 GB)
Reducers -> HDFS:
S3: Number of bytes written=1398604011 (~1.399 GB)

<u>Iteration 4:</u>
Mappers -> Reducers:
Map output materialized bytes=2257397797 (~2.257 GB)
Reducers -> HDFS:
S3: Number of bytes written=1398553684 (~1.398 GB)

<u>Iteration 5:</u>
Mappers -> Reducers:
Map output materialized bytes=2262446742 (~2.262 GB)

Reducers -> HDFS:
S3: Number of bytes written=1398519492 (~1.398 GB)
Iteration 6:
Mappers -> Reducers:
Map output materialized bytes=2265690722 (~2.266 GB)
Reducers -> HDFS:
S3: Number of bytes written=1398486868 (~1.398 GB)
Iteration 7:
Mappers -> Reducers:
Map output materialized bytes=2265424162 (~2.265 GB)
Reducers -> HDFS:
S3: Number of bytes written=1398445510 (~1.398 GB)
Iteration 8:
Mappers -> Reducers:
Map output materialized bytes=2260843034 (~2.261 GB)
Reducers -> HDFS:
S3: Number of bytes written=1398426886 (~1.398 GB)
Iteration 9:
Mappers -> Reducers:
Map output materialized bytes=2251275329 (~2.251 GB)
Reducers -> HDFS:
S3: Number of bytes written=1398411854 (~1.398 GB)
Iteration 10:
Mappers -> Reducers:
Map output materialized bytes=2250241015 (~2.25 GB)
Reducers -> HDFS:
S3: Number of bytes written=1398375929 (~1.398 GB)

The amount of data transferred from Mapper to Reducer in the first run is significantly lower in comparison to the rest of the runs. This could be because in the very first run, all the page ranks for all the pages is set to the string "NaN" while in later iterations, it varies from repeated calculation.

The amount of data transferred from Reducer to file remains more or less constant over repeated runs. I don't have an explanation for why this is happening.

Performance Comparison:

6 m4.large machines:
Preprocessing time: 33:26 (mm:s)
PageRank run time: 34:18 (mm:s)
TopK run time: 01:05 (mm:s)
10 m4.large machines:
Preprocessing time: 23:46 (mm:s)
PageRank run time: 27:32 (mm:s)
TopK run time: 00:45 (mm:s)

The expected behavior was that every phase except the TopK run would be almost twice as fast with 11 machines when compared to 6 machines. As can be seen from the run times above, it is the case. The only phase which doesn't show twice the speedup is the TopK run. This is happening because unlike the preprocessing and pagerank phases, this phase only has only one reduce task running, irrespective of the number of core machines dedicated to MapReduce. There is a decent amount of speedup, however, but the speedup is not as significant as that seen in the other phases.

<u>Sample Wiki PageRank:</u>
```
0.010656700141346977     United_States_09d4
0.008220639776057527     Wikimedia_Commons_7b57
0.0064952402353202566     Country
0.004711283065929041     England
0.0045635853871618965     Germany
0.004414657920218037     United_Kingdom_5ad7
0.004380909397817181     Europe
0.004359713264360057     France
0.004308947196780668     Water
0.004172974263868258     Animal
0.004124838763467     City
0.003990314502819405     Earth
0.0032022623206404612     Wiktionary
0.003190687587569432     Asia
0.0031397724967365728     Week
0.0030900864024562403     Money
0.0030672136089680358     Plant
0.0030177594605749116     Computer
0.003006232620767282     Sunday
0.002965622101995127     Monday
0.0029584070412357914     English_language
0.002938916857904719     Italy
0.0029376069177789255     Wednesday
0.002891313724770175     India
0.0028679582384330796     Friday
0.002859088683119939     Government
0.0028359297174876197     Saturday
0.002800200551367329     Thursday
0.0027797327121635543     Tuesday
0.0027158710035411282     Spain
0.002628420081539975     Number
0.002623675284060614     Japan
0.0025578671694649027     Canada
0.002441746192137675     People
0.002405821692748891     Human
0.0023446916812946664     Australia
```

```
0.002326782767279969    China
0.002308180786288308    Day
0.0023064364423891173    Wikimedia_Foundation_83d9
0.002241928372616844    Food
0.002174809526271151    Energy
0.0021670058929277227    Mathematics
0.0021543794871223734    Science
0.002136262375946121    Television
0.0021041390580040977    Sun
0.0020360651170750724    Capital_(city)
0.0020252283756537004    Music
0.0020069998433729277    Russia
0.0019582379703325797    State
0.0019081671466966736    Greece
0.0019024539949288237    index
0.0019005915011195112    Scotland
0.0018930928409946729    Year
0.0018916134932925557    Language
0.0018364339229566622    Metal
0.0018346998701719212    2004
0.0017946652579968441    Wikipedia
0.0017803392350589503    London
0.0017796319188465941    Greek_language
0.0017453826941465605    Religion
0.0017393334979107248    Sound
0.0016949247662825155    Africa
0.0016665522626982648    Planet
0.0016394780227830278    20th_century
0.001634741537357645    Poland
0.0016095159378284234    19th_century
0.0016014789732554573    Law
0.0015636273637329805    Geography
0.0015584184765731562    Liquid
0.0015508915158835817    World
0.001530950799436273    Scientist
0.00150874870954171     Society
0.001492023174735682    Inhabitant
0.0014871899132532646    Latin
0.0014829147468236852    Netherlands
0.0014804876270619798    Sweden
0.0014783865575128813    War
0.0014709528105714936    History
0.0014363723901016283    Light
0.0014363259811335820    Atom
0.0014225398458921927    Building
0.001417104455261456    Culture
```

```
0.0014159428457179684    God
0.0013843277505049129    Centuries
0.0013723761053671705    Information
0.001371173133459838     Capital_city
0.0013708669728049654    Turkey
0.0013551765855783241    Plural
0.001334247735034764     Portugal
0.0013353346242485215    Chemical_element
0.0013169168937867152    University
0.0013154875173614937    Denmark
0.0013129632257306593    Book
0.001312082444868015     Austria
0.0013042374491607416    Species
0.0012939934363504517    Disease
0.0012786131678996253    North_America_e7c4
0.0012645015761637707    Biology
0.0012585104653527607    Ocean
0.0012529807724905777    U.S._state_5a68
```

Full Wiki PageRank:
```
0.020328173945099632     United_States_09d4
0.01676857403516821      2006
0.009379956797138371     United_Kingdom_5ad7
0.007976707136631527     2005
0.007816284494847574     Biography
0.006403586799651899     Canada
0.006328172721341767     England
0.005875132531300701     France
0.005623137574084618     2004
0.005465648295730867     Geographic_coordinate_system
0.005193700055033483     Australia
0.005163551443629481     Germany
0.004632736551948406     India
0.00454805730413696      2003
0.00442459242921374      Japan
0.00387304648328624      Internet_Movie_Database_7ea7
0.0037118203429510376    2001
0.0036294043486943027    2002
0.003520935210772373     Italy
0.0034763517281045696    Record_label
0.003460132535920062     2000
0.0034325887639080664    Europe
0.0032467589437966725    Population_density
0.0031977663123053999    World_War_II_d045
0.0031443132743244552    London
0.003031902837474414     1999
```

```
0.0029645902411086045    Music_genre
0.002293111443324066     Spain
0.0027924565089871667    Race_(United_States_Census)_a07d
0.002767867474258461     English_language
0.0027088960240600193    Football_(soccer)
0.0026971219995995544    Russia
0.0026310676898812437    Wiktionary
0.002629982966837057     1998
0.002490265453607359     1997
0.002480644449033358     Scotland
0.00244392496105335      Television
0.002399543540459641     Census
0.0023872836566535315    New_York_City_1428
0.0023658860548404536    Sweden
0.0023461114805196906    Wikimedia_Commons_7b57
0.0023417826815513455    1996
0.00229899361301265      Square_mile
0.0022564758736645365    California
0.0021948342750785434    1995
0.0021585897616115589    New_Zealand_2311
0.0021335908634593271    Scientific_classification
0.0021148590561000764    China
0.0020939112947327577    Actor
0.0020917731177998266    1994
0.002049047181290158     Netherlands
0.0020474821402696455    Public_domain
0.002041315387841829     Politician
0.0020032764602457664    Film
0.0020006233317782663    New_York_3da4
0.0019885924252686584    United_States_Census_Bureau_2c85
0.0019836226086137375    1993
0.0019644005879764937    1991
0.001947383042797961     1990
0.0019397407028829239    Norway
0.0019302329005678675    Album
0.0018940257913438161    1992
0.0018881123586562952    Marriage
0.0018878081770618937    Population
0.0018662397823301633    Ireland
0.0018657206747686049    Poland
0.0018608058759947799    Record_producer
0.0018358869318855484    Personal_name
0.0017978074335498295    Per_capita_income
0.0017889129420747539    Brazil
0.0017607738686730645    1989
0.0017199458511551401    Mexico
```

```
0.0017064644912262938     1980
0.0016905682817106625      Studio_album
0.001663916076440879      1986
0.0016612547348724238      Poverty_line
0.0016330279890918664      1985
0.001626331837988042      1982
0.0016174863282596252      1981
0.0016120809203950246      Animal
0.0016092398801357975      1979
0.0016079671463468603      South_Africa_1287
0.001603195334692069      1987
0.0016001705386681042      1983
0.0015993641704791716      1984
0.0015740804855251243      1974
0.00157222768911212     Switzerland
0.0015681051082448592      1988
0.001553505183337771      January_1
0.0015281069107824495      Latin
0.0015218300236394188      1970
0.0015158575159968366      1976
0.0014971983401575616      1975
0.0014851192435254832      Paris
0.0014725613132124326      French_language
0.0014718499399261108      1978
0.001470968169533231     Greece
0.0014685829069004885      1977
0.0014648182391194186      1972
0.0014613959487998417      1969
```

The ranking seems reasonable for the full Wiki page dump. The same is more or less applicable for sample data as well, except for the page named "index", which shouldn't be there.