Name: Atindra Mardikar
Class: Nat Tuck (Tue/Fri 1:35-3:15pm)
HW-03 Report

Page Rank Source Code:

In the submitted HW folder there is a folder called source code inside which the complete package folder (mapredpagerank) for the java programs is present.

** To test update the arguments in the makefile for input.

Program Design:

Pre-processing:

For the preprocessing, I used the code provided (Bz2WikiParser.java) with a small change for the "&" to parse the input file and get it in the desired format.
A standalone Map only program was built which read the input and emitted the result in the desired form:

PageID: AdjacencyList

Pseudo code:

```
map(Object,Text){
        read input file;
        parse and filter out relevant links and strip URLs to only the page name;
        emit(pageID, adjacencyList);
}
```

PageRank:

PageRank uses the Pseudo code given in the module with an extra InputMapper to parse the output of the parser and emit pageID and Node.
The output of the InputMapper goes to the PageRankReducer and then it is fed to PageRankMapper and so on for 10 iterations.

InputMapper → PageRankReducer0 → PageRankMapper1 → PageRankReducer1 → ……. PageRankMapper10 → PageRankReducer10 → FinalOutput

Each PageRankMapper and PageRankReducer follows the Pseudo code provide in the learning modules.

Pseudo Code: (from the learning modules)

PageRankMapper

```
map(pageId n, PageNode N){

        // Emit the structure
        emit(PageID n, PageNode N);

        // Compute the contributions to send along the outgoing links
        p= N.pagerank/sizeof(n.adjacencyList);
        for all pageId m in N.adjacencyList
                emit(PageId m, p);
}
```

PageRankReducer

```
reduce(pageId n, PageNode [p1,p2,p3…]){
        s=0;
        PageNode M=null;
        for all p in (p1,p2,p3…){
                if(p.IsNode()){
                        M=p; //recover graph structure
                        if(p.danglingNode){
                                //update the global counter by adding the PR.
                                globalCounter.increment(p.pagerank);
                        }
                }
                else{
                        //pagerank contribution from inlink
                        s+=p.pagerank;
                }
        }

        // Retrieve V and totalDanglingPr from previous iteration.
        V=getCounter();
        dpr=getCounter();

        M.pagerank= (0.15/V)+ 0.85*(dpr/V+s);
        emit(PageId m, PageNode M);
}
```

Top-K:

For top K, I used a mapper to flip the key and value from the last iteration. So now my rank is the key. I used a keyComparator to sort key in descending instead of ascending. Also set the number of reduce tasks to 1 so that all my key,value pair go to one reducer and just print the top K.

Pseudo Code:

```
map(pageId n, PageNode N){

        // Emit rank as key
        emit(doubleWritable(N.pagerank), pageId n);
}

reduce(pagerank p, PageId [n1,n2..]){

        // Use key comparator to sort in descending order and then just emit

        count=0;
        for ( PageId n in [n1,n2…]{
                if(count<100){
                        emit(pageId n, pagerank p); //write to final output
                        count++;
                }
        }
}
```

From Input to first iteration the number of records for mapper to records in reducer are increased because for all the nodes in the adjacency list we emit from the mapper and there may be a few nodes which are in the adjacencylist but does not have its own pageId in the input file.

```
        Map input records=18326
        Map output records=437324
        Map output bytes=14453777
        Map output materialized bytes=15336102
        Input split bytes=156
        Combine input records=0
        Combine output records=0
        Reduce input groups=18803
        Reduce shuffle bytes=15336102
        Reduce input records=437324
        Reduce output records=18803
```

The number of records in reducer from this iteration is all fed into the next mapper and then the numbers of records are consistent for each mapper and reducer for every iteration.

```
Map input records=18803
Map output records=437801
Map output bytes=14467413
Map output materialized bytes=15350692
Input split bytes=150
Combine input records=0
Combine output records=0
Reduce input groups=18803
Reduce shuffle bytes=15350692
Reduce input records=437801
Reduce output records=18803
```

Just the number of reduce output records in the final iteration is 100 as we just output the top 100 page rank values.

```
Map input records=18803
Map output records=18803
Map output bytes=41447
Map output materialized bytes=452088
Input split bytes=151
Combine input records=0
Combine output records=0
Reduce input groups=13284
Reduce shuffle bytes=452088
Reduce input records=18803
Reduce output records=100
```

Performance Comparison:

** All times are approx. and calculated from the syslog file
** 2 syslog were generated for both the config and are included in the HW folder

6 m4.large machines (1 master and 5 workers):

Pre-processing: 50 mins
Ten iterations of pagerank: 42 mins
Top 100: 1 mins 30 secs

11 m4.large machines (1 master and 10 workers):

Pre-processing: 23 mins
Ten iterations of pagerank: 31 mins
Top 100: 1 mins

Technically as we double the machines the time required should be halved. The best speedup is for the pre-processing step, which is exactly half as expected. On the other hand the other steps does not show convincing speedup. For the top 100 the reason could be the logic I have used where in I just use one reducer. So irrespective of the machines it is going to use one reducer and hence there is not much difference in the time.

Top 100 Pagerank:

Simple dataset:

United_States_09d4	0.0036457336514012714
Wikimedia_Commons_7b57	0.002900211896448498
Country	0.0023841930025667655
England	0.0016076354928546597
Europe	0.0015933116155583784
United_Kingdom_5ad7	0.0015799597781994393
Water	0.001571754007258558
Germany	0.0015686775452684073
France	0.001531445614343456
Earth	0.0014988870322325027
Animal	0.0014958840282320051
City	0.0013810933222232935
Week	0.0012601469526201522
Asia	0.00118357535920737
Sunday	0.0011576600309512028
Monday	0.00113947752499526 2
Wiktionary	0.0011320878560800844
Wednesday	0.001128687015999079
Money	0.001115526582271338
Plant	0.0011052005445277239
Friday	0.0011013976632123104
Saturday	0.0010888137917031122
Thursday	0.0010746325652031873
Tuesday	0.0010669479717028826
Computer	0.0010669225101869132
English_language	0.0010656025668084703
Italy	0.0010469039398389218
India	0.0010333585441193405
Government	0.0010132042636543533
_D.C._323f	0.0010018378324730 25
Number	9.855574595061616E-4
Spain	9.406654067074949E-4
Day	9.224994734035039E-4
Japan	9.153854119918876E-4
People	8.796167589891882E-4
Canada	8.710942360287429E-4
Human	8.686985873664649E-4
index	8.529201091173797E-4

Wikimedia_Foundation_83d9	8.40931992223139E-4
China	8.295991280424488E-4
Energy	8.276148061790142E-4
Australia	8.114646908573078E-4
Sun	8.034399065516392E-4
Food	8.011831453471419E-4
Science	7.923885156043215E-4
Mathematics	7.82651778372165E-4
Television	7.35986131568437E-4
Russia	7.208563970496322E-4
Year	6.961548322727732E-4
Los_Angeles	6.924078018382897E-4
_California_b493	6.924078018382893E-4
Music	6.916830562589343E-4
State	6.914331367404398E-4
Greece	6.787279638417431E-4
Capital_(city)	6.78404509085539E-4
Language	6.78362396720809E-4
Scotland	6.70162173307612E-4
Metal	6.623374408318624E-4
Wikipedia	6.560903595997063E-4
Greek_language	6.500613079612854E-4
Planet	6.462953540618662E-4
2004	6.433628257067803E-4
Sound	6.264262609857914E-4
Religion	6.236739564126083E-4
London	6.204100192718527E-4
Africa	6.180356051345437E-4
Poland	5.852750438202857E-4
Geography	5.812612819203004E-4
Liquid	5.777047110862671E-4
20th_century	5.76114398988235E-4
Law	5.747524438640871E-4
World	5.663472987962343E-4
19th_century	5.622349664351726E-4
Scientist	5.59432652398932E-4
Society	5.579988436816589E-4
Atom	5.468071730352902E-4
History	5.372370234330368E-4
Latin	5.357806963518491E-4
Light	5.332097866950045E-4
Sweden	5.331274184389002E-4
War	5.252912885462298E-4
Netherlands	5.219408236606958E-4
Culture	5.207168683643369E-4
Turkey	5.065003152894516E-4
God	5.058889138784565E-4

Building	5.046897480612292E-4
Plural	5.008986206913951E-4
Information	4.96629864002508E-4
Chemical_element	4.885551798428397E-4
Portugal	4.8531122420991565E-4
Inhabitant	4.842023345969617E-4
Centuries	4.840585430152251E-4
Denmark	4.763826217595932E-4
Austria	4.728231511692797E-4
Cyprus	4.7034871240464807E-4
Ocean	4.647943443190837E-4
Moon	4.596621382377667E-4
Species	4.594335938218219E-4
Disease	4.583093198737373E-4
Book	4.577428022501388E-4

Full dataset:

United_States_09d4	0.0010279860187608954
2006	9.479427740076897E-4
United_Kingdom_5ad7	5.247855474966013E-4
2005	4.419089548708686E-4
Biography	3.761496655078716E-4
France	3.3799927065331585E-4
England	3.3679070434677443E-4
Canada	3.2222247361135434E-4
2004	3.116759046277542E-4
Encyclopædia_Britannica_Eleventh_Edition_8e5e	3.0220260529691606E-4
Germany	2.9378745743555406E-4
Australia	2.686776358291673E-4
India	2.5206745773757226E-4
2003	2.4674720733036746E-4
Km²	2.4661766229137807E-4
United_States_Census	2.3668768076431985E-4
Japan	2.3637337743324653E-4
Los_Angeles	2.2837923337397855E-4
_California_b493	2.238113627767767E-4
Geographic_coordinate_system	2.1540507033727615E-4
_D.C._323f	2.144123986492433E-4
United_Kingdom_general_election	2.111360523690077E-4
Italy	2.0914536016384053E-4
Internet_Movie_Database_7ea7	2.0532602212139026E-4
2002	2.0161636097251734E-4
2001	1.988052039172225E-4
Europe	1.963152271906992E-4

London 1.8572680463963033E-4
World_War_II_d045 1.8126126317880833E-4
2000 1.7936250403006461E-4
Record_label 1.7677443024407692E-4
_2004 1.7508354371180494E-4
English_language 1.7148491932864445E-4
University_of_California 1.6906793284030854E-4
1999 1.6865448164254742E-4
Spain 1.6762820495588645E-4
Wiktionary 1.6517587122415647E-4
Russia 1.5909649414938128E-4
Département_in_France_e00c 1.4938219600062194E-4
Music_genre 1.488227568223986E-4
_2005 1.475474804381259E-4
Wikimedia_Commons_7b57 1.4672345338696767E-4
Côte_d'Ivoire_ed5b 1.4663438696880534E-4
1998 1.4607400278309778E-4
Football_(soccer) 1.4028825608996339E-4
1997 1.3884249466926038E-4
Scotland 1.3493324290704564E-4
Television 1.315014006105302E-4
Sweden 1.3105610168534583E-4
_2006 1.298140655341638E-4
1996 1.2924126652725406E-4
New_York_City_1428 1.2697017266894247E-4
U.S._presidential_election 1.2531253500500672E-4
1995 1.2295162022195148E-4
China 1.2150928746740511E-4
_Massachusetts_d688 1.2142239345840578E-4
Netherlands 1.1855500450915757E-4
1994 1.1731052892234571E-4
New_Zealand_2311 1.1567700741738518E-4
_Pennsylvania_7d25 1.1284225055960716E-4
_2003 1.1226827717530464E-4
1991 1.118395703845326E-4
Public_domain 1.118161745484435E-4
Scientific_classification 1.1166246436038564E-4
1993 1.1104753090033526E-4
California 1.0900579468253738E-4
1990 1.0887867506175356E-4
Film 1.0878657272203412E-4
Actor 1.0789363089471828E-4
1992 1.0641885196308314E-4
Poland 1.0493224750001418E-4
Population_density 1.0381711031770996E-4
Norway 1.0377382016162713E-4
San_Francisco 1.0374414760143086E-4

_Illinois_2106	1.0318946355277588E-4
Ireland	1.0174629178576052E-4
_California_b6e2	9.970152710163918E-5
1989	9.969599510924079E-5
Latin	9.930257263459967E-5
Brazil	9.812521277295592E-5
1980	9.627844503787853E-5
January_1	9.569366758303565E-5
Album	9.548010398633883E-5
1986	9.460631912931416E-5
Politician	9.432600157182427E-5
New_York_3da4	9.4313002166003E-5
Record_producer	9.359504404457276E-5
Mexico	9.349478501895666E-5
French_language	9.290341560691247E-5
_DC_48ce	9.251926822035532E-5
1985	9.24675676420264E-5
1982	9.204908956218216E-5
1979	9.178994258763709E-5
_Georgia_4e3e	9.160398235191414E-5
1981	9.156340364535536E-5
Paris	9.15402735098818E-5
St._Louis	9.051148908897998E-5
1984	9.03330979175941E-5
1987	9.000481843155942E-5
1983	9.00018646975797E-5

The results of pagerank are quiet as expected. We have united states at the top which is fine and as we go down we have few other popular countries and days of the week. This is fairly as expected. As we go down there are few popular cities and languages and years. So yeah in an article there are generally outlinks pointing to such things so no wonder they rank higher in the table.