Class Number: SEC 01(BOS-2-TR) (CRN: 13036)

HW Number: 3 **Name: Anoop Pai**

Design Discussion:

The pre-processing approach that I incorporated made use of the given Bz2WikiParser.java. This approach ensures that it only preserves those links within 'bodyContent' div tag and ignores all others links. It also effectively strips off any file path prefixes and .html suffix leaving only the page name. This approach also successfully discarded any pagenames containing ~(tilde) character in them.

The PageRank program implemented the pseudo-code outlined in the modules. This approach divides the problem into two phases. In the first phase, a page sends out fractions of its current page rank along its outgoing edges. In the second phase, each page sums up the PageRank contributions along all its incoming edges. The only addition to my implemented program from the pseudo-code described in the modules is the handling of dangling nodes by including the value of delta in the formula. There were several ways to compute delta for dangling nodes in the modules. The approach that I followed was by maintaining a global counter which follows the principle of 'Merge computation of Delta into previous reduce phase'. In this approach delta is computed at the end of iteration i, so that it could be used for iteration i+1. This counter is incremented in the mapper phase every time a dangling node is encountered.

In order to compute TopKResults, I followed the approach given in the modules. This approach scans the input list, line by line, parsing the line to get page rank and page name values and inserting the pagerank value as key into a TreeMap data structure. As the property of tree ensures that the entries are sorted by its key, this approach leverages this by entering one record into it at a time, and discarding the first entry if a new entry is added and the size of the TreeMap exceeds 100 entries. Therefore, every mapper sends out local TopK results to the reducer. The Reducer reads through every local Top K result from the mappers and produces a global Top K result by essentially performing similar functionality from that of the mapper. The output is a TopK list of records in descending order of the pagerank values.

Data Transferred:

Iteration Value	Mappers to Reducers (bytes)	Reducers to HDFS (bytes)
Iteration1	1538158778	1491590570
Iteration2	1756292810	1491400195
Iteration3	1756547658	1491327310
Iteration4	1756416871	1491331343
Iteration5	1756520558	1491318139
Iteration6	1756635265	1491322638
Iteration7	1756611218	1491312139
Iteration8	1756755435	1491311869
Iteration9	1756667163	1491311584
Iteration10	1756725608	1491309324

The difference for both, amount of data transfer between mappers to reducers and reducers to HDFS is very negligible across different iterations. The change in value is roughly the same over 10 iterations. Although, as noticeable from the table, there is a significant difference of data transfer between mappers to reducers between Iteration1 and 2 than between other iterations. This could be explained by the fact that more outlinks can be discovered after iteration one. In the case of the first iteration, the outlinks for each node is read from the adjacency list generated from the pre-processing job. After Iteration 1, the file that is read by the mapper at Iteration 2, depends on the output of the reducer in Iteration 1. As new outlinks are discovered, they are added to the node structure, and therefore results in more data sent from mapper to reducer in Iteration 2.

Performance Comparison:

	6 m4.large machines	**11 m4.large machines**
Pre-processing time	13 mins 53 secs	8 mins 41 secs
Ten iterations of Pagerank	27 mins 17secs	14 mins 39 secs
TopK Results	1 min 2 secs	42 secs

As evident from the table above, better speedup can be observed in the pre-processing phase and PageRank job phase. This difference is significant in the case of pre-processing and pagerank jobs as there is a lot of data involved (more traffic between mappers and reducers, also traffic from reducers to HDFS) which could be leveraged by using more machines. The difference in the case of TopK results is minimal as in both cases we are dealing with emitting top100 records for the input data set. Having 5 more machines doesn't drastically improve the speedup as we're concerned with 100 records in both cases, and having several more machines doesn't improve the running time in a great way. As observed from the syslog files,

```
Number of bytes written=3238 for 6 machines
Number of bytes written=3239 for 11 machines
```

As this value is almost the same for both 6 machines vs 11 machines, this results in rather poor speedup.

Top-100 Wikipedia Pages:

Simple data set:

```
United_States_09d4:  0.005458351352856293
Wikimedia_Commons_7b57:  0.004365927492811352
Country:  0.0035433989688555606
England:  0.0024407253491515795
Germany:  0.0023809651253248244
Europe:  0.002376637248554111
United_Kingdom_5ad7:  0.0023679249857932274
Water:  0.0023371693279392106
France:  0.0023083966644495187
Animal:  0.002247760926708638
Earth:  0.0022097515921970915
City:  0.0020789290382824607
Week:  0.0018251695235291133
Asia:  0.0017575203048763915
Wiktionary:  0.0017081886805031
Sunday:  0.0016989365027587884
Monday:  0.0016729927772312484
Money:  0.0016679759859938876
Plant:  0.0016583072118842553
Wednesday:  0.001657531881654813
Friday:  0.0016177020974040028
Computer:  0.0016105243727129 71
Saturday:  0.0015994526544411134
English_language:  0.0015952802116620318
Thursday:  0.0015788453603585554
Italy:  0.0015752137766022205
Tuesday:  0.0015675384087696304
India:  0.0015460420315318438
Government:  0.0015114025519524339
Number:  0.0014541077076003323
Spain:  0.0014199002592347367
Japan:  0.0013816755155036744
Day:  0.001336594584541612
```

People: 0.001314821905928574
Canada: 0.0013052590569882653
Human: 0.0013016762493325501
Wikimedia_Foundation_83d9: 0.001258468032999788
China: 0.0012449984539525997
Energy: 0.001219837304181776
Australia: 0.0012182500526570304
Food: 0.0012037604875280911
Science: 0.001179393463618545
Sun: 0.0011767638589870167
Mathematics: 0.0011709791915695347
index: 0.0011491499722276948
Television: 0.001117151880936274
Russia: 0.0010797140725622724
Music: 0.0010545668804916735
Year: 0.0010333828771668966
State: 0.0010300302581225974
Greece: 0.0010204693318833001
Language: 0.0010189503562045607
Capital_(city): 0.0010158517836664663
Scotland: 0.0010095119197330198
Metal: 9.901845957727999E-4
Wikipedia: 9.820742423295E-4
2004: 9.69800974033915E-4
Greek_language: 9.691041095275052E-4
London: 9.416007566927337E-4
Planet: 9.414039576193629E-4
Sound: 9.408390439254908E-4
Religion: 9.362983323975534E-4
Africa: 9.19530032174076E-4
Poland: 8.78267425375345E-4
20th_century: 8.691324817957274E-4
Geography: 8.626849781417545E-4
Law: 8.606866706063362E-4
Liquid: 8.564718543578831E-4
19th_century: 8.48754270783759E-4
World: 8.443068589640385E-4
Scientist: 8.337257210208316E-4
Society: 8.296480233459732E-4
Atom: 8.030029246576257E-4
History: 8.011747465506352E-4
Latin: 8.0029123897156E-4
Sweden: 7.981434947455812E-4
Light: 7.912789822080848E-4
War: 7.876680635409976E-4
Netherlands: 7.861933365863827E-4
Culture: 7.755963053692902E-4
God: 7.590794898430814E-4
Building: 7.554068306302891E-4
Turkey: 7.536018919083421E-4
Inhabitant: 7.464825122008327E-4
Plural: 7.44981035186368E-4
Information: 7.426612302161724E-4
Centuries: 7.327148962651551E-4
Portugal: 7.261256490071358E-4
Chemical_element: 7.257394886785997E-4
Denmark: 7.115370055190203E-4

Austria: 7.083242607692868E-4
Cyprus: 6.961411976910087E-4
Species: 6.950367104677497E-4
Book: 6.930899429413559E-4
Disease: 6.909513443532878E-4
University: 6.90131919762503E-4
Ocean: 6.892400615551847E-4
Biology: 6.863661383322514E-4
Capital_city: 6.862677397560723E-4
North_America_e7c4: 6.70863994035089E-4

Full data set:

United_States_09d4: 0.0010536895282195387
2006: 9.747476951720468E-4
United_Kingdom_5ad7: 5.386421744725589E-4
2005: 4.53936872728234E-4
Biography: 3.8307073123498213E-4
France: 3.4731055861474624E-4
England: 3.4522020339047225E-4
Canada: 3.300336014644149E-4
2004: 3.200369813314975E-4
Germany: 3.016317482282253E-4
Australia: 2.75352661479081E-4
India: 2.582408106633031E-4
2003: 2.5336661323952544E-4
Japan: 2.4251834984754107E-4
Geographic_coordinate_system: 2.1971322631715056E-4
Italy: 2.1501253865354768E-4
Internet_Movie_Database_7ea7: 2.1009493901645758E-4
2002: 2.0698284352624108E-4
2001: 2.0396955863792498E-4
Europe: 2.0163366733447037E-4
London: 1.9073469428723837E-4
World_War_II_d045: 1.8631261302942543E-4
2000: 1.840206098889639E-4
Record_label: 1.804332244787803E-4
English_language: 1.7653755067307316E-4
1999: 1.7315274624915368E-4
Spain: 1.7223300633700673E-4
Wiktionary: 1.699566547162186E-4
Russia: 1.636285383898102E-4
Music_genre: 1.518268303417644E-4
Wikimedia_Commons_7b57: 1.5127854775750732E-4
1998: 1.4994607752877065E-4
Football_(soccer): 1.433393973245959E-4
1997: 1.425673015010558E-4
Scotland: 1.3844263705762715E-4
Television: 1.346843600907152E-4
Sweden: 1.3439114215389354E-4
1996: 1.326985860307239E-4
New_York_City_1428: 1.3043538291118635E-4
1995: 1.2626375279739394E-4
China: 1.2480920383123938E-4
Netherlands: 1.2185520783497464E-4
1994: 1.2048442316747488E-4
New_Zealand_2311: 1.1863534773419481E-4

1991: 1.1494277239527325E-4
Public_domain: 1.1458177359541244E-4
Scientific_classification: 1.1404566478786986E-4
1993: 1.1403425802299908E-4
1990: 1.1186712284315937E-4
California: 1.1166500002639193E-4
Film: 1.1141904475000256E-4
Actor: 1.102721460813951E-4
1992: 1.0930719865493796E-4
Poland: 1.0763173489896846E-4
Norway: 1.063714689074772E-4
Population_density: 1.0602261687629597E-4
Ireland: 1.043890690441676E-4
1989: 1.0244401080987395E-4
Latin: 1.0235478846150315E-4
Brazil: 1.0058998935144309E-4
1980: 9.896944731565784E-5
January_1: 9.869113045162882E-5
Album: 9.730053882874919E-5
1986: 9.724069943605147E-5
New_York_3da4: 9.666738370030173E-5
Politician: 9.608201706592814E-5
Mexico: 9.594582657497062E-5
French_language: 9.576727283366687E-5
Record_producer: 9.549546679944901E-5
1985: 9.501850408449496E-5
1982: 9.459418586651086E-5
1979: 9.438177023233195E-5
1981: 9.411168592931413E-5
Paris: 9.408584670626314E-5
1984: 9.281392776893489E-5
1983: 9.246873067424363E-5
1987: 9.246021500839021E-5
1974: 9.237626654883038E-5
South_Africa_1287: 9.178604180887902E-5
Switzerland: 9.079577361505942E-5
Personal_name: 9.042285852931988E-5
1988: 9.040235316441634E-5
1970: 8.957904778988453E-5
1976: 8.950652498674797E-5
1975: 8.873279803260922E-5
Animal: 8.832745342105989E-5
Soviet_Union_ad1f: 8.781725161912768E-5
Greece: 8.770261741221979E-5
1945: 8.717669622760933E-5
1969: 8.691958893279049E-5
1972: 8.66156544038222E-5
1977: 8.630722603877652E-5
1978: 8.600470338267009E-5
Portugal: 8.517152849655552E-5
Austria: 8.439110840993373E-5
1973: 8.435518811491949E-5
Studio_album: 8.421852228247513E-5
Iran: 8.357296772018084E-5
Denmark: 8.347051581970005E-5
1971: 8.278672109717845E-5

Yes, the observed results do follow my intuition because it is reasonable to assume that a page like United_States_09d4 would have many incoming links. Also, the pages linking to United_States_09d4 have a high Page Rank in itself and therefore leading to high Page rank of United_States_09d4.