Ajay Vardhan
CS6240 section 2
Assignment 3

Pre-processing:

The pre-processing was based on the instructions given in the homework. A SAX Parser was used to parse the data and create the adjacency list. Here are the steps followed by the parser:

1. Read the data one page at a time
2. Extract the page name
3. Skip to the div with the ID “bodyContent”
4. Iterate through the anchor tags and extract the href values
5. Extract the page name from the href and discard any unwanted data as instructed in the homework
6. Remove the duplicate links
7. Create an adjacency list from these links for the page name and output it to a file

When checked with a random bunch of pages, the output was accurate to what was needed.

Page Rank:

The page rank algorithm was implemented based on the pseudo code given in Module 6: Graph Algorithms, with an addition of handling dangling nodes using Solution 2 in the same module to compute the dangling nodes. Here are the steps followed to calculate the page rank:

1. Find all the dangling nodes and the total number of nodes in the pre-processing phase when parsing the data
2. Use that information to calculate the initial page rank for all nodes during the first iteration
3. The delta value was assumed to be 0 for the first iteration
4. During the first iteration, page ranks of all the nodes were calculated, including the delta value for the next iteration. The algorithm given in the Module 6: Graph Algorithms was used here
5. The output for this iteration is used as the input for the next iteration and the delta value calculated here is passed on to the next iteration
6. This is repeated for 10 iteration. During the final iteration, the final output with all the converged page ranks were emitted

The algorithm for calculated page rank was chosen based on the efficiency and amount of data transferred each time. By using this algorithm, we iterate and emit only the required amount of data and all machines are used in a reasonably balanced way. Using a separate job before each page rank job to calculate the dangling nodes was a lot of work for the machines and very inefficient, so the delta value was calculated in the previous iteration for each job and passed on to the next iteration. I did not find any difference in the efficiency between this method and order inversion so I went ahead with this algorithm. The only additional work here is updating a global variable which is of negligible cost when compared with the overall job.

Top K:

TopK algorithm followed the steps given in Module 5: Basic Algorithms. The local winners for each map task were stored in a TreeMap without exceeding a total size of 100. All these local winners are sent to a single reducer by using a null key, and the top 100 from the reducer is also found using a local TreeMap aggregator and they are emitted out.

This was the most efficient method amongst all the methods that I had in mind. Secondary sort would add the unnecessary work of sorting the whole data set before sending it, manually sorting it would also be extra work. Using any other data structure would not give us the sorted output right away. Using TreeMap and limited it's size was the most efficient way. The local winners in each map would guarantee to have the overall winners as well. So this algorithm was chosen.

Amount of data transferred:

6 machines:

Iteration	Mapper to Reducer (no. of records)	Reducer to HDFS (no. of records)
1	55823171	3179022
2	55823653	3179022
3	55823653	3179022
4	55823653	3179022
5	55823653	3179022
6	55823653	3179022
7	55823653	3179022
8	55823653	3179022
9	55823653	3179022
10	55823653	3179022

11 machines:

Iteration	Mapper to Reducer (no. of records)	Reducer to HDFS (no. of records)
1	55823171	3179033
2	55823664	3179033
3	55823664	3179033
4	55823664	3179033
5	55823664	3179033
6	55823664	3179033
7	55823664	3179033
8	55823664	3179033
9	55823664	3179033
10	55823664	3179033

Since we calculate the detla only during our first iteration, there is a slight variation of data being transferred from mapper to reducer, after the delta value is calculated, the amount of data transfer remains the same for all iteration. Since we are processing the same data every time and emitting the page ranks for the same set of pages, we don't see any difference in the data transfer.

Running time:

6 Machines:

	Pre-Processing	Page rank (10 iterations)	Top 100 pages
Running time	40:43	25:44	00:56

11 Machines:

	Pre-Processing	Page rank (10 iterations)	Top 100 pages
Running time	19:13	15:38	00:41

The run times were as I expected them to be. The program complete fairly sooner when ran with more machines since the parallelism is more. The same amount of data, when processed with more machines, completes sooner. If we compare the speedup for each step, the pre-processing had the maximum improvement. Since this is the heaviest part of the program, this had the most impact. We can also verify from the syslog that 6 machines just lauched 9 reduce tasks whereas 11 machines launched 20 reduce tasks. The shuffled maps shuffled were also 954 for 6 machines 2014 for 11 machines. These gave the opportunity for more parallelism which resulted in more speedup. This seemed as a fair result as more machines lead to more speedup for a large chunk of data.

We also see a good speedup in the other phases but since the data is relatively smaller, the difference is not as drastic as pre-processing. The page rank calculation also had a increase in reduce tasks (6 machines – 9, 11 machines – 20) and shuffle maps (6 machines – 162, 11 machines – 361).

So over all, the programs behaved as I expected them to behave based on the number of machines used.

Top 100 pages:

The output seemed reasonable since all of them are popular pages and it makes sense that these will have the most pagerank in the given dataset. The pagerank also seem to have converged to a reasonable number from the initial values. The distribution is seems accurate. There weren't any surprises in the output with both the inputs.

Full dataset:

United_States_09d4 : 0.0004963795733764
Biography : 0.0003012077285819
United_Kingdom_5ad7 : 0.0001985218456853
2006 : 0.0001967050596331
Geographic_coordinate_system : 0.0001810134002697
England : 0.0001657455332593
Canada : 0.000156532209805
2005 : 0.000148691532904
Record_label : 0.0001255303950844
Australia : 0.0001206709517121
Music_genre : 0.0001206665271824
2004 : 0.0001189692747664
France : 0.0001181584262688
India : 0.0001158430694202
Internet_Movie_Database_7ea7 : 0.0001137600727995
Germany : 0.0001101072930085
2003 : 0.0001002594762911
Japan : 0.0000969705936161
Population_density : 0.0000921705108175

2001 : 0.0000833100969702
Politician : 0.0000800081273703
Europe : 0.0000797775273589
2002 : 0.0000797273248454
Football_(soccer) : 0.0000763385241577
2000 : 0.0000760270201774
Scientific_classification : 0.0000749549376807
Record_producer : 0.0000742734630802
Studio_album : 0.0000738294728309
Census : 0.0000713308463664
Personal_name : 0.0000694697831563
Album : 0.0000679204432178
World_War_II_d045 : 0.0000677468137214
London : 0.0000675260101411
1999 : 0.0000658372671755
Television : 0.0000649603054051
Italy : 0.0000632999590828
1998 : 0.0000592929796894
Actor : 0.0000585816246104
Marriage : 0.0000580508978683
Public_domain : 0.0000579728465948
Square_mile : 0.0000575710550211
Km² : 0.000055854339381
Per_capita_income : 0.0000556790255762
1997 : 0.0000553872588404
United_States_Census_Bureau_2c85 : 0.0000551081834618
Poverty_line : 0.0000549797143183
Spain : 0.0000544812425546
Scotland : 0.0000534982454683
California : 0.0000534542376655
1996 : 0.0000528715241651
English_language : 0.0000526639750111
Wiktionary : 0.0000523641380725
Film : 0.0000521763279065
Animal : 0.0000516117092817
Population : 0.0000512440517563
White_(U.S._Census)_c45a : 0.0000494439668716
Sweden : 0.0000492691052152
New_York_City_1428 : 0.0000486016428744
1995 : 0.0000484713954455
School : 0.0000484318111996
Writer : 0.0000471650510085
Russia : 0.0000471089555599
New_York_3da4 : 0.0000469866923321
1994 : 0.0000463754422042
China : 0.0000463458068601
New_Zealand_2311 : 0.0000461922528967

Norway : 0.0000454794676142
1993 : 0.0000446881341361
1992 : 0.0000424952748481
1990 : 0.0000420617489019
Poet : 0.0000420105128676
1991 : 0.0000418775003426
Brazil : 0.0000416188783121
Corporation : 0.0000415694286005
Latino_(U.S._Census)_5f0e : 0.0000410181664038
Hispanic_(U.S._Census)_1387 : 0.0000410077938645
USA_f75d : 0.0000408666192044
Ireland : 0.0000408092119751
Website : 0.0000400008334025
Poland : 0.0000397801463561
Binomial_nomenclature : 0.0000391930967244
Netherlands : 0.000038916591021
1989 : 0.0000386666743122
Race_(United_States_Census)_a07d : 0.0000382288887407
1980 : 0.0000377687530177
Company_(law) : 0.0000375334683054
Building : 0.0000369472587079
Band_(music) : 0.0000368474153688
1982 : 0.0000365102120758
All_Music_Guide_0e49 : 0.0000363928255755
1986 : 0.0000363921898214
1983 : 0.0000361912995949
Native_American_(U.S._Census)_1a7a : 0.000036034877043
1981 : 0.0000360060560732
1985 : 0.0000359958806123
1984 : 0.0000359827000107
1987 : 0.0000358487948299
1988 : 0.0000355106395725
Rock_music : 0.0000351555081881
1979 : 0.0000351214038387

Simple Dataset:

United_States_09d4 : 0.0010003884949175
Wikimedia_Commons_7b57 : 0.0008400444266429
England : 0.0007040868171239
Germany : 0.0006426416935204
France : 0.0004798383492342
City : 0.0004138498523064
Inhabitant : 0.0004076637959362
Wiktionary : 0.0003646176245935
Country : 0.0003543602532404

Animal : 0.0003497121922316
Japan : 0.0003356708472961
United_Kingdom_5ad7 : 0.0003351307225881
Computer : 0.000333962186426
Water : 0.0003093583570966
Europe : 0.0003046168373541
India : 0.0003040704711936
Spain : 0.0002862751020975
Australia : 0.0002858460526661
English_language : 0.0002821516939026
Italy : 0.0002820428932741
Canada : 0.0002760210802776
Television : 0.0002733396362203
Plant : 0.0002647377781386
Earth : 0.000264361326668
London : 0.000241804561945
Money : 0.0002415552986099
China : 0.0002398613512482
Greece : 0.0002358295243245
Music : 0.0002350365942618
Scotland : 0.000234830931415
Food : 0.0002321834442976
Football_(soccer) : 0.0002308334468301
Capital_(city) : 0.0002248335958143
Human : 0.0002224587795103
Metal : 0.0002218690514029
Capital_city : 0.0002161496867612
Mathematics : 0.0002134476779881
Movie : 0.0002127455974942
Netherlands : 0.0002116566237034
Government : 0.0002080389750704
Russia : 0.0002048664428054
Brazil : 0.0002044031752767
U.S._state_5a68 : 0.0002043228133021
Number : 0.0002037510022386
Greek_mythology : 0.0002035912073174
Book : 0.0002035837607729
People : 0.0002034718625752
2005 : 0.0002008819915561
Poland : 0.0001985161664598
2004 : 0.0001976206914645
Language : 0.0001975864646322
2006 : 0.0001945903408034
Religion : 0.0001899476004632
Year : 0.0001890388262255
Actor : 0.000187064591437
God : 0.0001840618030375

Asia : 0.0001840504435217
California : 0.0001825962727973
Sweden : 0.0001823023522473
Science : 0.0001807172881675
University : 0.0001799836234908
19th_century : 0.0001790154809193
Fruit : 0.0001749876335488
Car : 0.0001711875989002
Chemical_element : 0.0001684642359408
Africa : 0.0001684361911754
Disease : 0.0001660813138709
Film : 0.0001659795923117
Internet : 0.0001653987455037
World_War_II_d045 : 0.0001651838915599
Species : 0.0001646381721874
Latin : 0.0001638989953097
Company : 0.000162517003906
River : 0.0001596566912962
North_America_e7c4 : 0.0001590850537459
Fish : 0.0001585263020904
20th_century : 0.0001571584250426
Liquid : 0.0001549390952479
1970s : 0.000154833245238
Island : 0.0001544823696856
Centuries : 0.0001540831578346
Greek_language : 0.0001539532754094
Internet_Movie_Database_7ea7 : 0.0001528517618542
Video_game : 0.0001519498045559
Sport : 0.0001515700981988
War : 0.0001504317264891
1960s : 0.0001475172482981
Mammal : 0.0001471607626914
Christianity : 0.0001471008082695
German_language : 0.0001467105285373
Law : 0.000146658334246
Prefecture : 0.000145747812797
Sun : 0.0001451004709616
County : 0.0001445944509396
Singer : 0.0001442421610144
State : 0.000143081130508
Tree : 0.00014278679846
Austria : 0.0001427544656804
Chad : 0.0001421239311374
Child : 0.0001414240428774