Map-Reduce Algorithm

Preprocessor Pseudocode:

```
map (…, Line L) {
        Call WikiParser class and parse the .bz file
        extract page, and it's adjacency list
        emit [page, :DummyPageRankValue: [Adjacency List]]
}
```

PageRank Pseudocode:

```
map (…, Line L) {
        Extract page P, pageRank PR, Adjacency List lst
        Compute p = PR/NumOfNodes
        Calculate finalPageRank = p + (1-alpha) + prOfDanglingNodes/NumOfNodes
        Iterate over Adjacency List lst
                Calculate pr = PR/NumOfNodes
                emit [page p, pagerank pr]
        emit [p, lst]
}
reduce (page p, List of (PageRanks && Adjacency List lst)) {
        for each page p, calculate the sum of PageRanks
        Calculate finalPageRank fpr= p + (1-alpha) + sumOfPageRanks * alpha/NumOfNodes
        emit [page, :fpr:lst]
        Increment the pageRankValue of the Dangling Nodes by adding finalPageRank
}
```

TopK Pseudocode:

```
Class Mapper {
        Setup () {
                Map<String, Double> hm = new HashMap<String, Double>(); //Key is page, Value is
                pagerank
        }
        map (…, Line L) {
                Extract page P, pageRank PR, Adjacency List lst
                Compute p = PR/NumOfNodes
                Calculate finalPageRank = p + (1-alpha) + prOfDanglingNodes/NumOfNodes
                Hm.put(page, finalpageRank)
        }
        Cleanup () {
        Sort HM by Value using a Sort Utility
        For each key page,
```

```
                emit [ Null, [page, pageRank]] //First 100
        }
}

reducer ( …, List of(page, PageRanks pr)) {
        Extract page and page rank
        Put in topk HM
        Sort topK HM by value using Sort Utility
        emit [ page, pagerank] //Top 100 map values
}
```

Performance Comparison

Machines	ParseJob(in ms)	PageRankJob(in ms)	TopKJob(in ms)
6 m4.Large	1720723	1995712	591565
11 m4.Large	1313246	1143019	220731

Which of the computation phases showed a good speedup? If a phase seems to show fairly poor speedup, briefly discuss possible reasons—make sure you provide concrete evidence, e.g., numbers from the log file or analytical arguments based on the algorithm's properties.

The Preprocessing step(ParseJob) phase shows a good speed up since it derives the first processing of the Page Rank step by generating the Adjacency List of all the nodes.

Output of 6 m4.Large Execution

Below is the result of the 10 iterations of Page Rank algorithm on 6 m4.Large machines on full dataset. The records transferred from mapper to reducer and reducer to HDFS only changes from 1st iteration to 2nd iteration post which the record transfer remains the same throughout the rest of the 9 iterations.

1st Iteration:
Map input records=3074760
Map output records=77427575
Map output bytes=4223336990
Map output materialized bytes=1783221142
Input split bytes=2862
Combine input records=0
Combine output records=0
Reduce input groups=3364771

Reduce shuffle bytes=1783221142
Reduce input records=77427575
Reduce output records=3069902

2nd Iteration:
Map input records=3069902
Map output records=77395587
Map output bytes=4225192290
Map output materialized bytes=2149176951
Input split bytes=2862
Combine input records=0
Combine output records=0
Reduce input groups=3364724
Reduce shuffle bytes=2149176951
Reduce input records=77395587
Reduce output records=3069902

3rd Iteration:
Map input records=3069902
Map output records=77395587
Map output bytes=4226736637
Map output materialized bytes=2148329923
Input split bytes=2862
Combine input records=0
Combine output records=0
Reduce input groups=3364724
Reduce shuffle bytes=2148329923
Reduce input records=77395587
Reduce output records=3069902

4th Iteration:
Map input records=3069902
Map output records=77395587
Map output bytes=4228082778
Map output materialized bytes=2148182123
Input split bytes=2862
Combine input records=0
Combine output records=0
Reduce input groups=3364724
Reduce shuffle bytes=2148182123
Reduce input records=77395587
Reduce output records=3069902

5th Iteration:

Map input records=3069902
Map output records=77395587
Map output bytes=4228082778
Map output materialized bytes=2148182123
Input split bytes=2862
Combine input records=0
Combine output records=0
Reduce input groups=3364724
Reduce shuffle bytes=2148182123
Reduce input records=77395587
Reduce output records=3069902

6th Iteration:
Map input records=3069902
Map output records=77395587
Map output bytes=4228082778
Map output materialized bytes=2148182123
Input split bytes=2862
Combine input records=0
Combine output records=0
Reduce input groups=3364724
Reduce shuffle bytes=2148182123
Reduce input records=77395587
Reduce output records=3069902

7th Iteration:
Map input records=3069902
Map output records=77395587
Map output bytes=4228082778
Map output materialized bytes=2148182123
Input split bytes=2862
Combine input records=0
Combine output records=0
Reduce input groups=3364724
Reduce shuffle bytes=2148182123
Reduce input records=77395587
Reduce output records=3069902

8th Iteration:
Map input records=3069902
Map output records=77395587
Map output bytes=4228082778
Map output materialized bytes=2148182123

Input split bytes=2862
Combine input records=0
Combine output records=0
Reduce input groups=3364724
Reduce shuffle bytes=2148182123
Reduce input records=77395587
Reduce output records=3069902

9th Iteration:
Map input records=3069902
Map output records=77395587
Map output bytes=4228082778
Map output materialized bytes=2148182123
Input split bytes=2862
Combine input records=0
Combine output records=0
Reduce input groups=3364724
Reduce shuffle bytes=2148182123
Reduce input records=77395587
Reduce output records=3069902

10th Iteration:
Map input records=3069902
Map output records=77395587
Map output bytes=4228082778
Map output materialized bytes=2148182123
Input split bytes=2862
Combine input records=0
Combine output records=0
Reduce input groups=3364724
Reduce shuffle bytes=2148182123
Reduce input records=77395587
Reduce output records=3069902

Report the top-100 Wikipedia pages with the highest PageRanks, along with their rank values and sorted from highest to lowest, for both the simple and full datasets. Do they seem reasonable based on your intuition about important information on Wikipedia?

The top-100 Wikipedia pages result is present in the Output folder of the submission for both simple and full datasets.

According to my intuition, it ranks the pages in order of the important information on Wikipedia website.