CS 6240: Assignment 3

Xiao Wang

GitHub: https://github.ccs.neu.edu/xiaowang/CS6240-MapReduce

1. Design

Preprocesser:

```
parse(line):
    // removes self-loops as they affect PageRank results
    returns [pageName, [link]]
map(line):
    String[] s = parse(line)
    emit(s[0], s[1])
reduce(page, links):
    // count #pages and #dangling to counters for Iterator
    // key: page, value: score|links
    emit(page, "0|" + links)
```

Iterator:

```
// same as the code "MapReduce Code V2" in module 6 slides
// extra codes are used to parse the value format
```

Ranker:

```
map(page, score):
    emit(score, page)
partition(score, page):
    return 0
compare(w1, w2):
    return -w1.compareTo(w2)
reduce(score, page):
    while (count < 100):
        emit(page, score)
        ++count
```

2. Performance

time (ms)	5 workers	10 workers
Preprocesser	2326079	1090294
Iterator	1639505	895843
Ranker	70301	42238

#records	5 workers		10 workers	
	Mapper->Reducer	Reducer->HDFS	Mapper->Reducer	Reducer->HDFS
1	65292104	3175035	65292104	3175035
2	65377046	3175035	65376559	3175035
3	65377046	3175035	65376559	3175035
4	65377046	3175035	65376559	3175035
5	65377046	3175035	65376559	3175035
6	65377046	3175035	65376559	3175035
7	65377046	3175035	65376559	3175035
8	65377046	3175035	65376559	3175035
9	65377046	3175035	65376559	3175035
10	65377046	3175035	65376559	3175035

Q: 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.
A: By comparing the time spent with 5 workers and 10 workers, we can see that Preprocessor and Iterator phases showed a good speedup while Ranker showed fairly poor speedup. The reason is mostly likely to be that Preprocessor and Iterator phases always utilize all worker power to process data while during Ranker phase we always partition all pages to one single reducer in order to get the top 100 results, thus we only utilize one machine despite the total number of workers.

Q: 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?
A: (Please refer to output file for detailed ranking)
Yes, the results seem reasonable as the top pages are mostly countries, years and some very common terms. And the year 2006 when this dataset saved appears on the top of the years which