CT414 Map Reduce

Rory Murray 17395111

Implementation

Firstly I altered the original MapReduceFiles code so only approach 3 remained in the class as the other two approaches were not necessary. In order to make sure only proper individual words were included in the map I created a new method called clean which essentially checked that each input contained letters only without any additional characters such as fullstops or commas. So each word was checked first prior to being mapped.

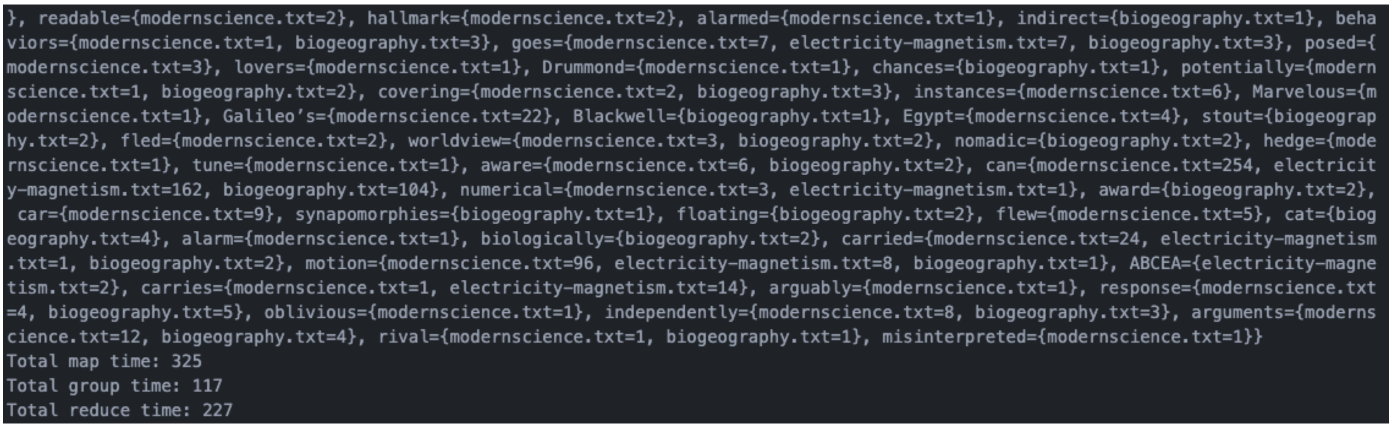
To measure the length of each phase of the program I simply stored the current system time from before and after each phase was run. I then subtracted the difference and printed out the result.

In order to specify the number of threads per line in the map phase I firstly had to alter the readFile method. I changed the method to store each line in an array list rather than append each line to one big string. This would allow me to iterate through each line in the text file and create multiple threads for a specified number of lines. Before this however I created a new string containing a block of lines as the map method is not designed to work with an array-list of strings.

Within the reduce method I followed a similar approach adding each word to an array-list before using a series of for loops to create a specified block of words for each thread. Finally I allowed the number of lines and words per thread to be specified at run time via the command line.

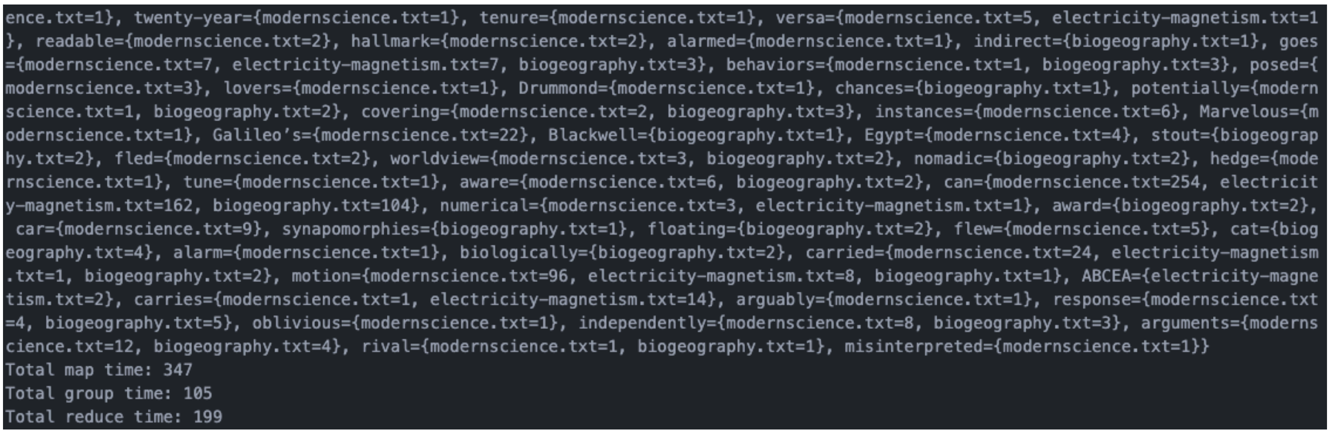
Run Time

Command : java MapReduceFiles electricity-magnetism.txt biogeography.txt modern science.txt 1000 100



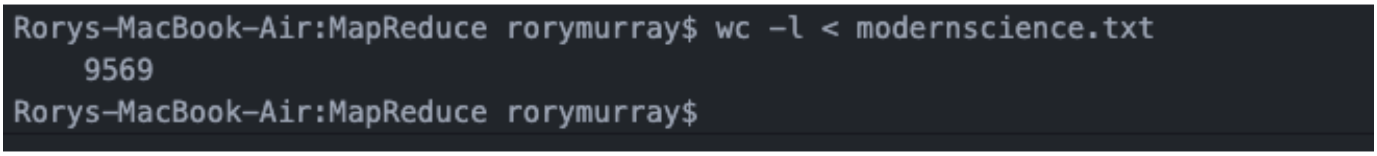
```
}, readable={modernscience.txt=2}, hallmark={modernscience.txt=2}, alarmed={modernscience.txt=1}, indirect={biogeography.txt=1}, beha
viors={modernscience.txt=1, biogeography.txt=3}, goes={modernscience.txt=7, electricity–magnetism.txt=7, biogeography.txt=3}, posed={
modernscience.txt=3}, lovers={modernscience.txt=1}, Drummond={modernscience.txt=1}, chances={biogeography.txt=1}, potentially={modern
science.txt=1, biogeography.txt=2}, covering={modernscience.txt=2, biogeography.txt=3}, instances={modernscience.txt=6}, Marvelous={m
odernscience.txt=1}, Galileo’s={modernscience.txt=22}, Blackwell={biogeography.txt=1}, Egypt={modernscience.txt=4}, stout={biogeograp
hy.txt=2}, fled={modernscience.txt=2}, worldview={modernscience.txt=3, biogeography.txt=2}, nomadic={biogeography.txt=2}, hedge={mode
rnscience.txt=1}, tune={modernscience.txt=1}, aware={modernscience.txt=6, biogeography.txt=2}, can={modernscience.txt=254, electricit
y–magnetism.txt=162, biogeography.txt=104}, numerical={modernscience.txt=3, electricity–magnetism.txt=1}, award={biogeography.txt=2},
 car={modernscience.txt=9}, synapomorphies={biogeography.txt=1}, floating={biogeography.txt=2}, flew={modernscience.txt=5}, cat={biog
eography.txt=4}, alarm={modernscience.txt=1}, biologically={biogeography.txt=2}, carried={modernscience.txt=24, electricity–magnetism
.txt=1, biogeography.txt=2}, motion={modernscience.txt=96, electricity–magnetism.txt=8, biogeography.txt=1}, ABCEA={electricity–magne
tism.txt=2}, carries={modernscience.txt=1, electricity–magnetism.txt=14}, arguably={modernscience.txt=1}, response={modernscience.txt
=4, biogeography.txt=5}, oblivious={modernscience.txt=1}, independently={modernscience.txt=8, biogeography.txt=3}, arguments={moderns
cience.txt=12, biogeography.txt=4}, rival={modernscience.txt=1, biogeography.txt=1}, misinterpreted={modernscience.txt=1}}
Total map time: 325
Total group time: 117
Total reduce time: 227
```

Command : java MapReduceFiles electricity-magnetism.txt biogeography.txt modern science.txt 500 100



```
ence.txt=1}, twenty-year={modernscience.txt=1}, tenure={modernscience.txt=1}, versa={modernscience.txt=5, electricity-magnetism.txt=1
}, readable={modernscience.txt=2}, hallmark={modernscience.txt=2}, alarmed={modernscience.txt=1}, indirect={biogeography.txt=1}, goes
={modernscience.txt=7, electricity-magnetism.txt=7, biogeography.txt=3}, behaviors={modernscience.txt=1, biogeography.txt=3}, posed={
modernscience.txt=3}, lovers={modernscience.txt=1}, Drummond={modernscience.txt=1}, chances={biogeography.txt=1}, potentially={modern
science.txt=1, biogeography.txt=2}, covering={modernscience.txt=2, biogeography.txt=3}, instances={modernscience.txt=6}, Marvelous={m
odernscience.txt=1}, Galileo’s={modernscience.txt=22}, Blackwell={biogeography.txt=1}, Egypt={modernscience.txt=4}, stout={biogeograp
hy.txt=2}, fled={modernscience.txt=2}, worldview={modernscience.txt=3, biogeography.txt=2}, nomadic={biogeography.txt=2}, hedge={mode
rnscience.txt=1}, tune={modernscience.txt=1}, aware={modernscience.txt=6, biogeography.txt=2}, can={modernscience.txt=254, electricit
y-magnetism.txt=162, biogeography.txt=104}, numerical={modernscience.txt=3, electricity-magnetism.txt=1}, award={biogeography.txt=2},
 car={modernscience.txt=9}, synapomorphies={biogeography.txt=1}, floating={biogeography.txt=2}, flew={modernscience.txt=5}, cat={biog
eography.txt=4}, alarm={modernscience.txt=1}, biologically={biogeography.txt=2}, carried={modernscience.txt=24, electricity-magnetism
.txt=1, biogeography.txt=2}, motion={modernscience.txt=96, electricity-magnetism.txt=8, biogeography.txt=1}, ABCEA={electricity-magne
tism.txt=2}, carries={modernscience.txt=1, electricity-magnetism.txt=14}, arguably={modernscience.txt=1}, response={modernscience.txt
=4, biogeography.txt=5}, oblivious={modernscience.txt=1}, independently={modernscience.txt=8, biogeography.txt=3}, arguments={moderns
cience.txt=12, biogeography.txt=4}, rival={modernscience.txt=1, biogeography.txt=1}, misinterpreted={modernscience.txt=1}}
Total map time: 347
Total group time: 105
Total reduce time: 199
```

Text File Size



```
Rorys-MacBook-Air:MapReduce rorymurray$ wc -l < modernscience.txt
    9569
Rorys-MacBook-Air:MapReduce rorymurray$
```

Results

Map size Reduce size	Map Time	Group Time	Reduce Time	Total
1000 50	307	99	157	563
1000 100	376	119	115	690
1000 200	310	114	91	**515**
500 50	432	123	134	689
500 100	339	113	229	681
500 200	400	130	224	754
250 50	465	140	139	744
250 100	395	103	196	694
250 200	360	122	95	577
Original	211	88	3426	3725

*All times are in milli seconds

Evaluation

Both programs were tested on three text files with 9569,4222 and 7540 lines. As we can see quite clearly that the adjusted program is significantly faster than original in regards to the total time. The quickest runtime achieved by the program was 515ms with 1000 lines per thread for the mapping phase and 200 words per thread for reduction phase. This is more than 7 times faster than the original program which is a considerable improvement in performance. For all the runs of the adjusted program it is worth noting that the total times are relatively close with them sitting between a range of 515-754. The program seemed to perform better with higher values for mapping phase however performance varied with different reduce values. For example the best run when the map value was 1000 was with a reduce value of 200, however the when the map value was 500 the worst run was with a reduce value of 200.