Comparing Word Count

Problem
Do Alice and Bob agree on their word count output?

Goals
- Identify differences
- Fast
Alice chose to run 3 reducers, resulting in 3 parts
Bob chose to run 2 reducers, resulting in 2 parts
Alice and Bob agree on their outputs
Files are sorted
Now Bob counted 2 more of "why"
MapReduce

SourceMapper

Alice’s Output

a 20
hi 2
i 13
the 31
why 12

Bob’s Output

a 20
hi 2
i 13
the 31
why 14

CompareReducer

a 20
a 20
i 13
i 13
hi 2
hi 2
why 12
why 14
the 31
the 31

why 2
MapReduce

SourceMapper

Alice's Output

Bob's Output

CompareReducer

a 20
hi 2
i 13
the 31
why 12
a 20
hi 2
why 14
hi 2
i 13
the 31
why 12
the 31
why 14
the 31
why 2

Kenneth Heafield (Google Inc)
Reducer Joins

Implementation

- Input files named alice00000, bob00000, etc.
- Mapper determines if it is reading Alice or Bob by the file name
- Map input key is *word* and value is *count*.
  - Alice Output key *word* and value (Alice, *count*).
  - Bob Output key *word* and value (Bob, *count*).
- Reducer sees key *word* and values (Alice, *count*) and (Bob, *count*)
  - But not necessarily in order

Issues

- Copies data to Mapper and again to Reducer.