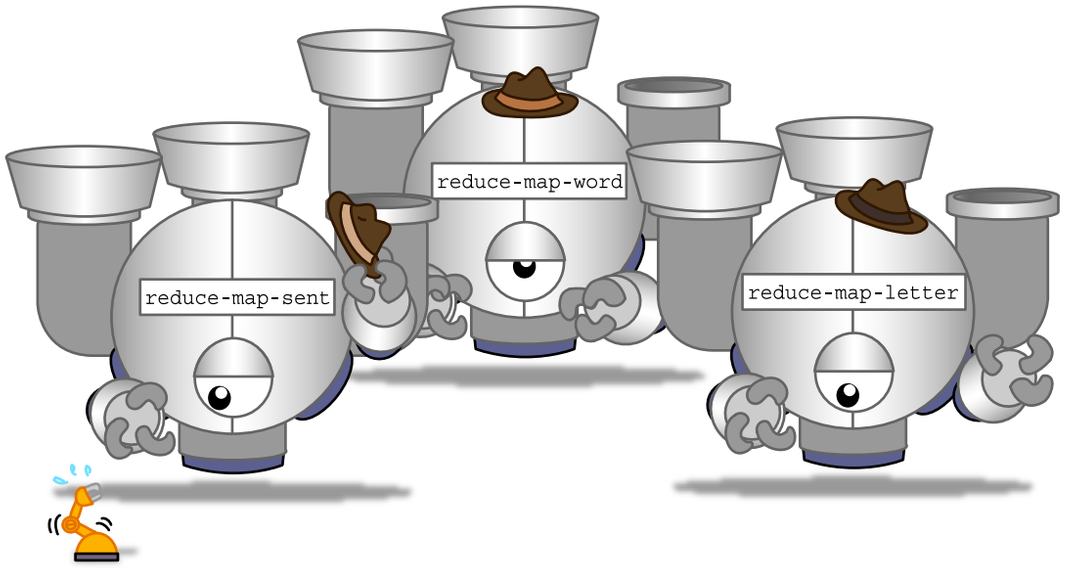
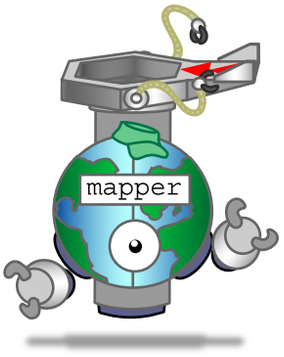
MapReduce

Parallelism and Functional Programming

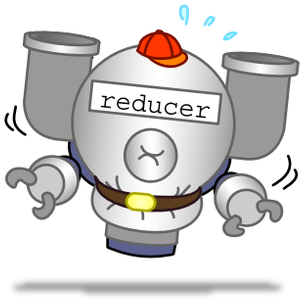


MapReduce is a system that makes writing parallel code easier for programmers.

In CS3, we provide three functions that differ only in how they interpret data. Each takes three arguments:



The first of the arguments is a mapper, a function that takes a sentence, word or letter (depending on whether it was passed to `reduce-map-sent`, `reduce-map-word`, or `reduce-map-letter` respectively) and outputs some value. There are no constraints on what this value can be. (A mapper is the kind of function used as an argument to `map`.)



Next is the reducer, which collects and combines the values returned from the mappers, into one value.
(A reducer is the kind of function used as an argument to `reduce`.)

Finally, there's the vast body of data to be processed, specified by a filename, but encoded as a list. It's either a list of sentences, a list of words or a list of letters (depending on whether it was passed to `reduce-map-sent`, `reduce-map-word`, or `reduce-map-letter` respectively). If given a directory instead of a single file, MapReduce treats the input as a file composed of the concatenation of the individual files in the directory.

Katrina Yim

CS Illustrated .berkeley.edu