Isai Mercado Oliveros

Misaie

October 5, 2015

# Netflix MapReduce Algorithm

When I was preparing for solving the Netflix problem, I read that since mapreduce algorithms are used in really big sets of data that computing needs a lot of time, data scientist try to only have one passing of map reduce instead of two or three. Therefore, I took my notebook, and started designing my algorithm to solved the Netflix problem in one mapreduce as follows.

MAPPER

InputValue = Textline with tokens userID movieID stars

outputKey = movieID stars // groups users saw same movie, and gave same rank

outputValue = userID

COMBINER

inputKey = movieID stars

inputValues = List of userIDs // these are the same users that had one similarity

outputKey = userID userID // alphabetical ordered to avoid duplicates of same pair

outputValues = 1

REDUCER

inputKey = userID userID

inputValue = 1

outputKey = userID userID

outputValue = counter

At the end of the algorithm we have lines of text that have one user next to another user and the number of movies that they ranked the same. If we needed a top ten, we could do a second mapreduce to filter the 10 most similar users.