BigData HW10

Tiago Antunes

May 2021

1 Introduction

The solution is in the *simple_app_python/* folder. The *inputs* folder contains the files generated by the generator script, and the *outputs* folder contains the top-k for each iteration.

The generator script was slightly modified, creating a checkpoints folder and terminating automatically after 5 iterations.

2 Implementation

The changes done to the base file were very little. First, a checkpoint directory was added so that the *updateStateByKey* function could be used. Then, a new function called *update* was created, which sums the values of each key with its new value obtained from the curent context.

Then, this *updateStateKey* is applied on the current context obtained counts, which sums the historical values with the new ones. These final values are then transformed by applying a *sortBy* function that sorts in decreasing order of values. This is then passed to the *pprint* function with a parameter 100, thus printing the top-K of each iteration taking into account the whole history since the program started.

A more efficient way would be to apply a *take* function instead of sorting, but the code got incredibly complex and having some spark errors, so I find this implementation enough for this homework. The files are relatively small too.