gedit bookmapper.py

#!/usr/bin/python

import sys

def mapper():

for line in sys.stdin:

fields=line.strip().split("\t")

t,a,py,ac,l,pgs=fields

print "{0}\t1".format(py)

mapper()

gedit books.txt

title author publishedyear country language pages

b1 a1 2000 c1 english 100

b2 a2 2000 c2 spanish 150

b3 a3 2001 c3 french 250

b4 a4 2002 c4 GERMAN 175

b5 a5 2000 c5 mandarin 200

cat books.txt | python bookmapper.py

gedit bookreducer.py

#!/usr/bin/python

import sys

def reducer():

cy,my=None,None

cc,mc=0,0

for line in sys.stdin:

fields=line.strip().split('\t')

y,c=fields

if y=='publishedyear':continue

if y==cy:

cc+=int(c)

else:

if cy:

print "{0}\t{1}".format(cy,cc)

if cc>mc:

mc=cc

my=cy

cy=y

cc=int(c)

if cy:

print "{0}\t{1}".format(cy,cc)

if my:

print " Max year\t",my

reducer()

cat books.txt | python bookmapper.py | sort | python bookreducer.py

hdfs dfs -put books.txt books.txt