gedit moviemapper.py

#!/usr/bin/python

import sys

def mapper():

for line in sys.stdin:

fields=line.strip().split()

id=fields[0]

tag=fields[2]

print "{0}\t{1}".format(id,tag)

mapper()

gedit movies.txt

movieid rating tag

1 3 action

1 3 adventure

1 3 thriller

2 2.5 romantic

2 2.5 comedy

2 2.5 drama

3 4 sci-fi

3 4 thriller

3 4 psycological

cat movies.txt | python moviemapper.py

gedit moviereducer.py

#!/usr/bin/python

import sys

def reducer():

cmid=None

tags=set()

for line in sys.stdin:

mid,tag=line.strip().split('\t')

if mid=='movieid':

continue

if mid==cmid:

tags.add(tag)

else:

if cmid:

print "{0}\t{1}".format(cmid,','.join(tags))

cmid=mid

tags=set()

tags.add(tag)

if cmid:

print "{0}\t{1}".format(cmid,','.join(tags))

reducer()

cat movies.txt | python moviemapper.py | sort | python moviereducer.py

hdfs dfs -put movies.txt movies.txt