Big Data and Analytics
BTech.CE.SemVII

Categorising Books (Big Vs Small)(Total counts based on page count)

Prof. Jigar Pandya • Jul 15

Provided total numbers of pages of each book line by line in multiple data files across hdfs, we need to find out how may total large books and small books we have got. We may define large book to be having page count above certain limit i.e. 300

pages.txt
Text

CategorizeBook.java
Java

1 class comment

Prof. Jigar Pandya Jul 15
Input Lines showing number of pages of certain book:
350

250

150

450

120

Mapper Output. For every line map function is called. Given the page count as value of the kv pair. Output of all multiple calls to map function combined into context as below:

<"Big Books",1>
<"Small Books",1>
<"Small Books",1>
<"Big Books",1>
<"Small Books",1>

Sort and Shuffle (Bucket Generated for similar keys)
Bucket for "Big Books" entries
<"Big Books",1>
<"Big Books",1>

Bucket for "Small Books" entries
<"Small Books",1>
<"Small Books",1>
<"Small Books",1>

Note that input to is a key and set of values grouped into a list. For every key reducer is called separately. Here, because two unique key are there reducer function is running two times.

Reducer given following:
<"Big Books", [1,1]>
<"Small Books", [1,1,1]>

Simple logic of summation very similar to unique word count will fetch us statistics.
Output of reducer to files
<"Big Books", 2>
<"Small Books", 3>

Add class comment...