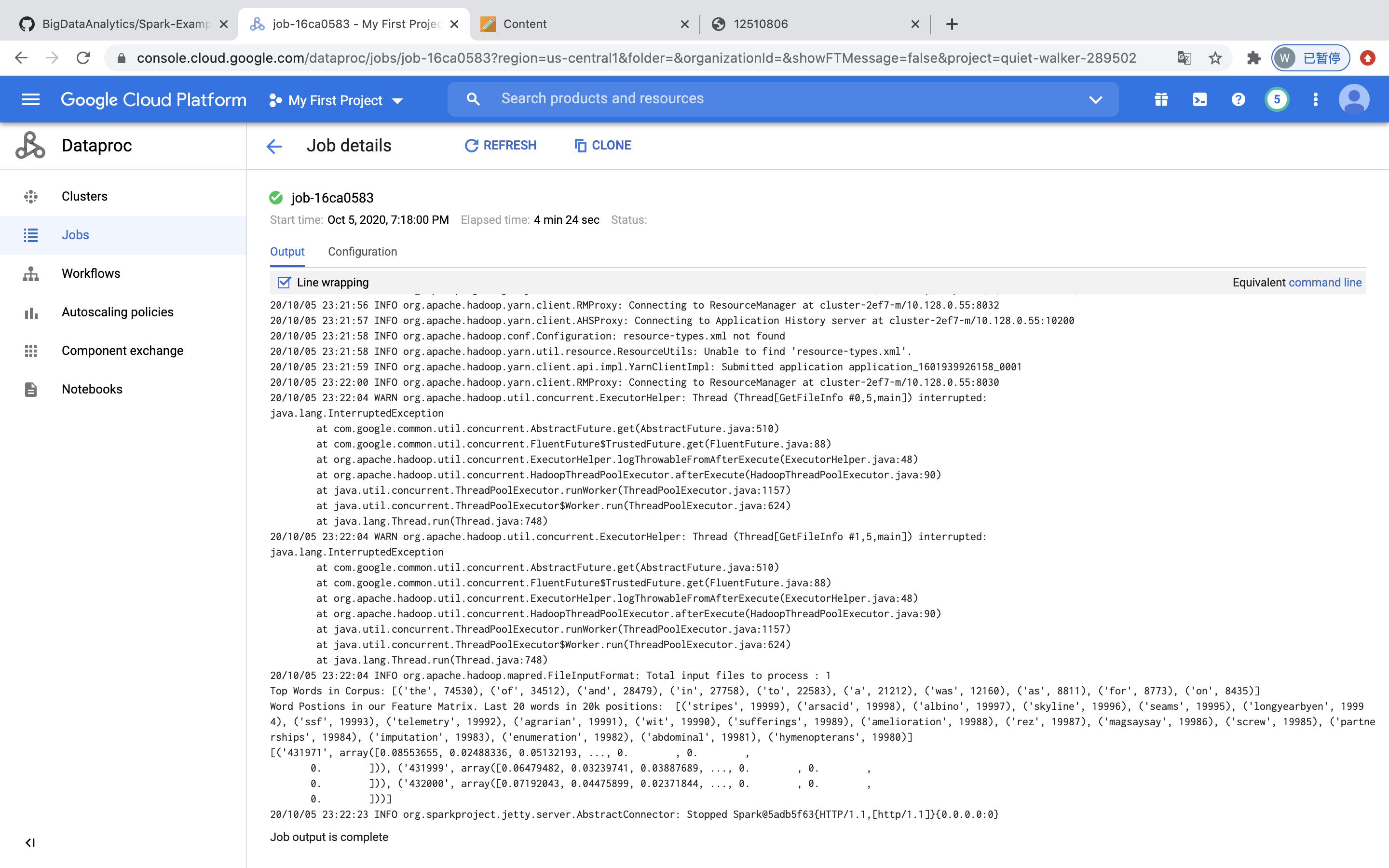
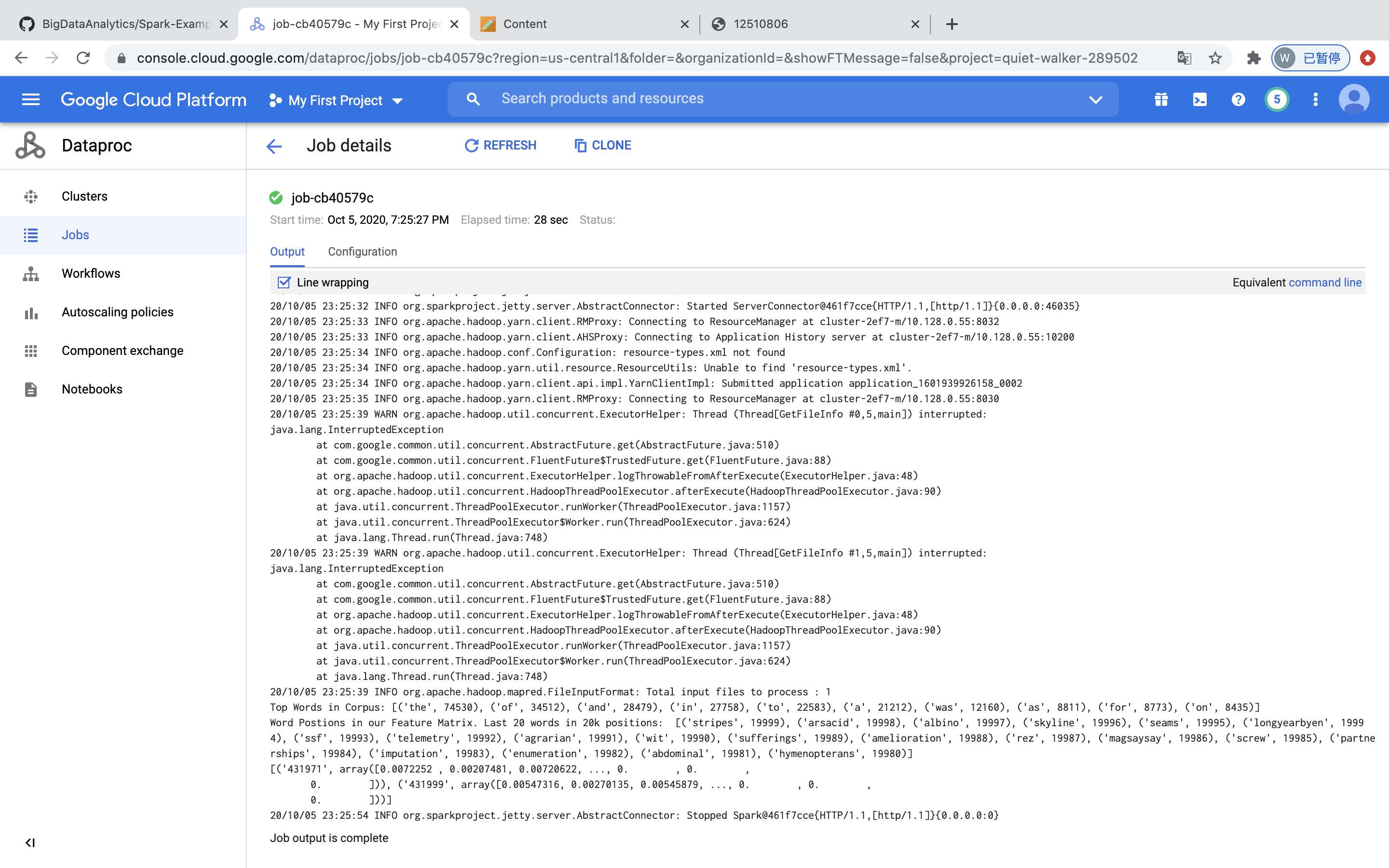
Small set by RDD:

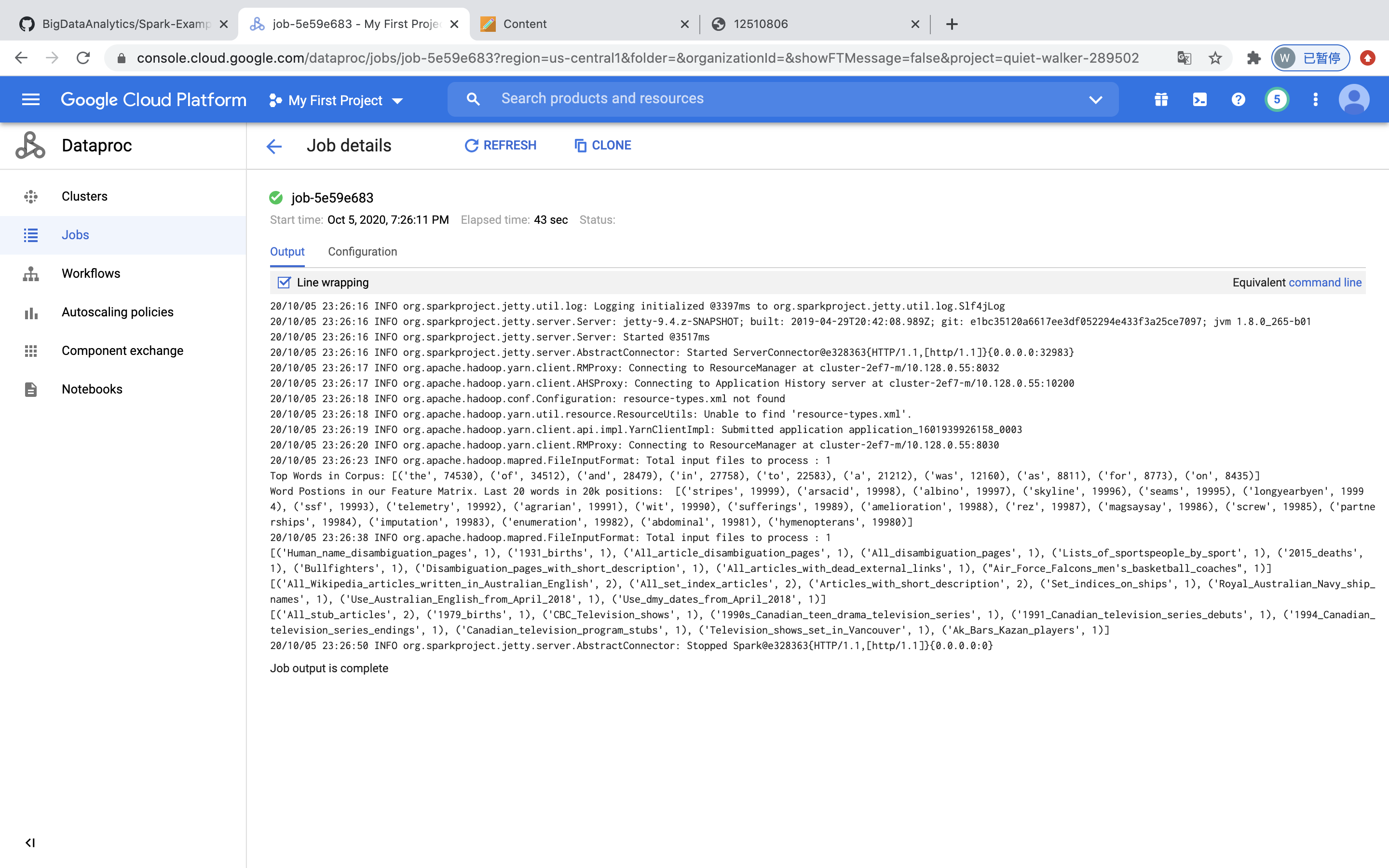
Task1:



Task2:



Task3:



Task4 :

Small set by DataFrame:

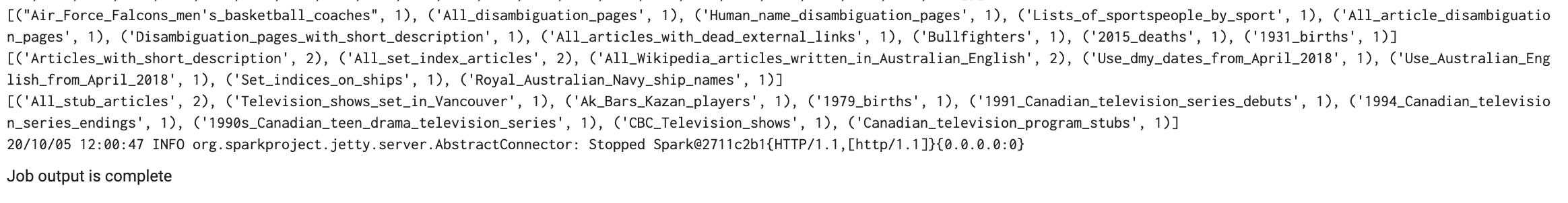
Task1:

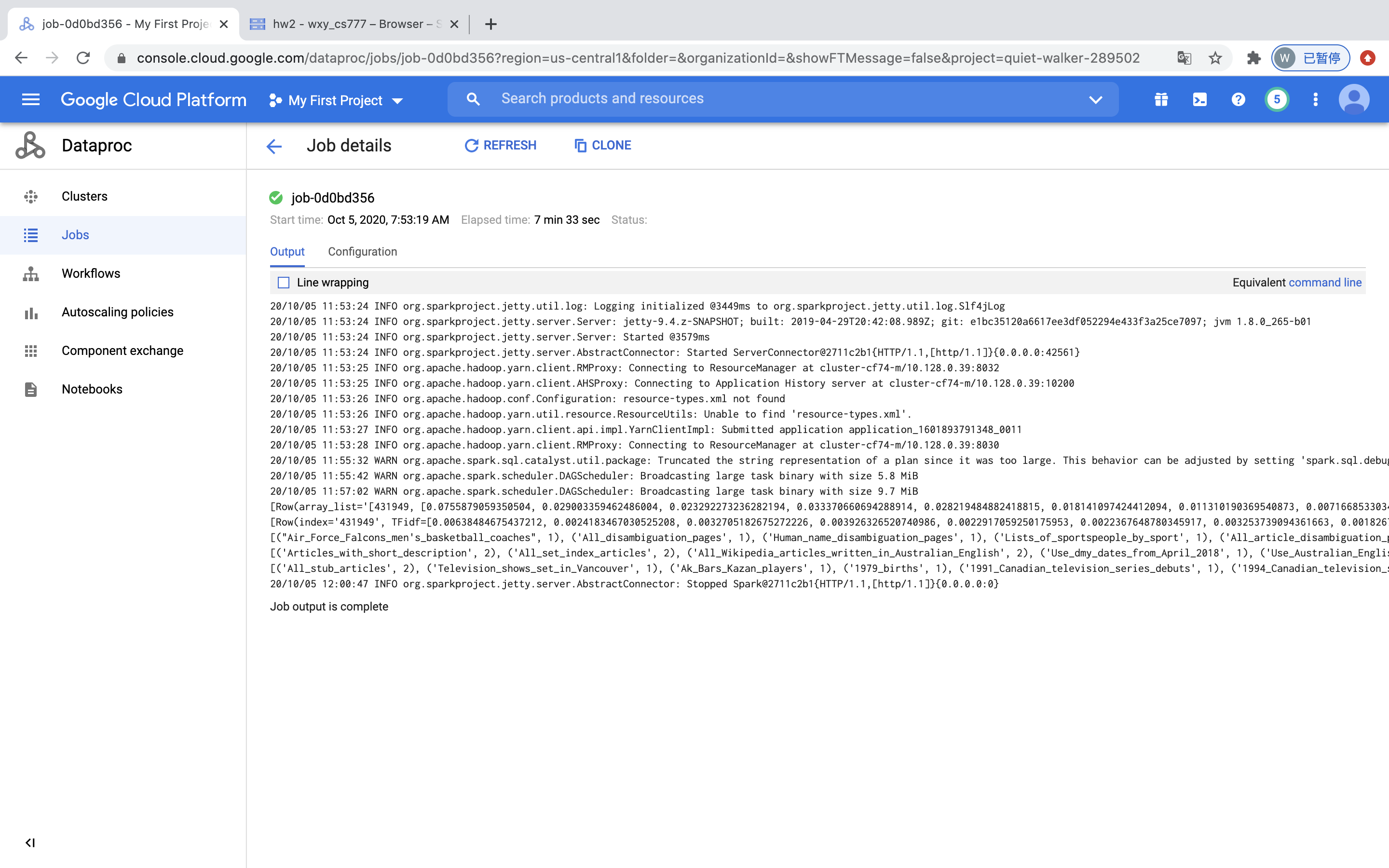


Task2:



Task3:





Big set by RDD:

I have some problems in running my job on large dataset: after I submitted my codes,

My task will break up in 30 minutes ,though I use 8v,32ram machine. The error massage said it is machine problem but not my codes, in addition, Google cloud doesn’t allow me to build larger machine so I don’t know how to run large dataset.

Task5:

5.1:

In my opinion, stemming those stop words will change my result a little, because I use small training dataset, but as size of training size becomes larger, the impact is going to be less. Cause in real world, when dataset becomes larger, the frequencies of occurrences of these stop words will be closed to each other. Also, these words have no contribution to the classification, because they have poor ideographic ability and literally have no clear meaning.

However, when dataset is small, there can be scenarios that the stop words shows more times in a certain category and when remove them, the result could have changed sometimes.

5.2:

In my opinion, stemming those words to their root word will not change my result heavily, because they have the same meaning no matter which tenses they are in.So when change their tenses, their ideographic will not change.