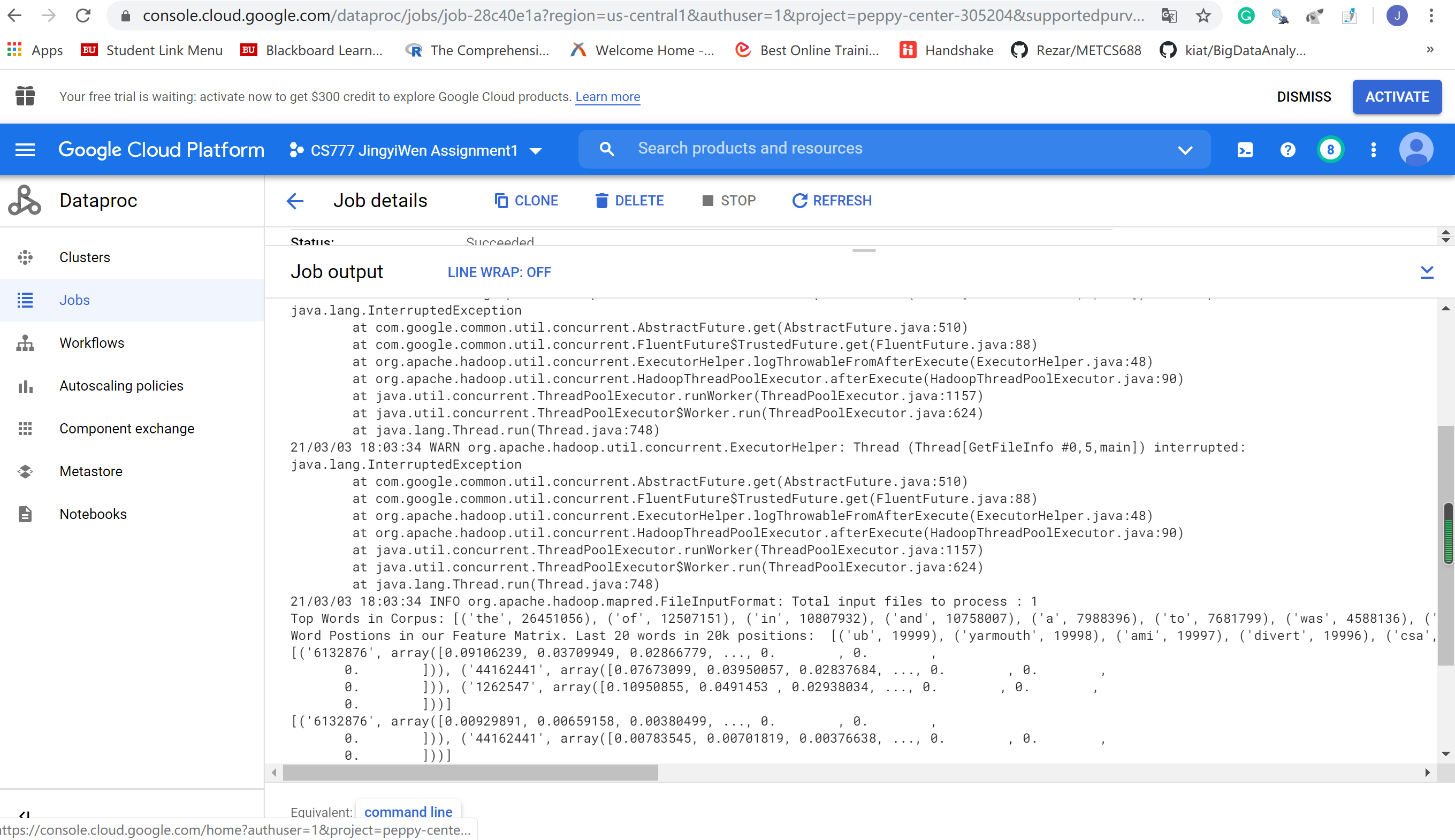
Assignment 2 Answer File

Jingyi Wen

Task 1:

print(allDocsAsNumpyArrays.take(3))

print(allDocsAsNumpyArraysTFidf.take(2))

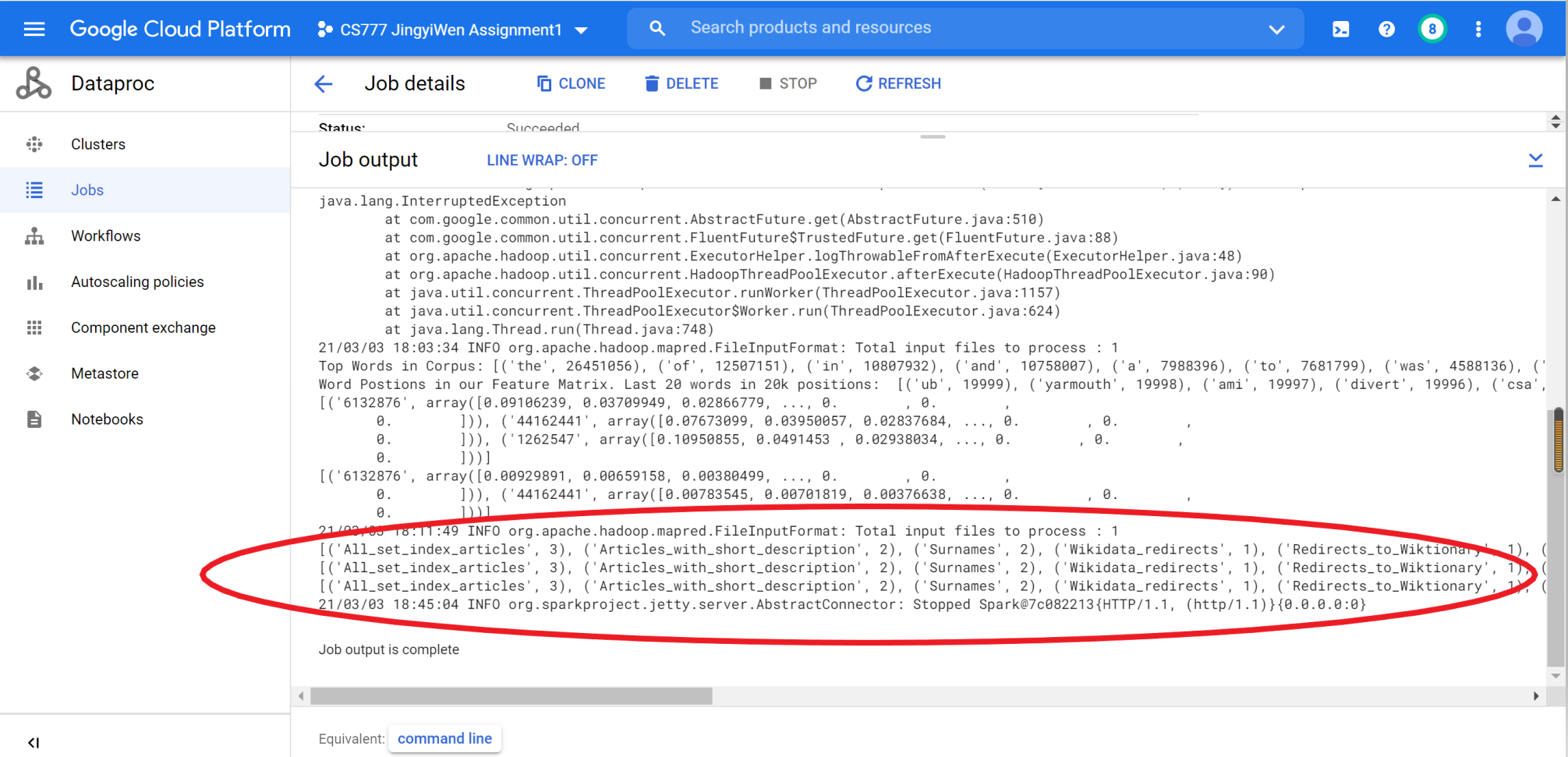


Task 2

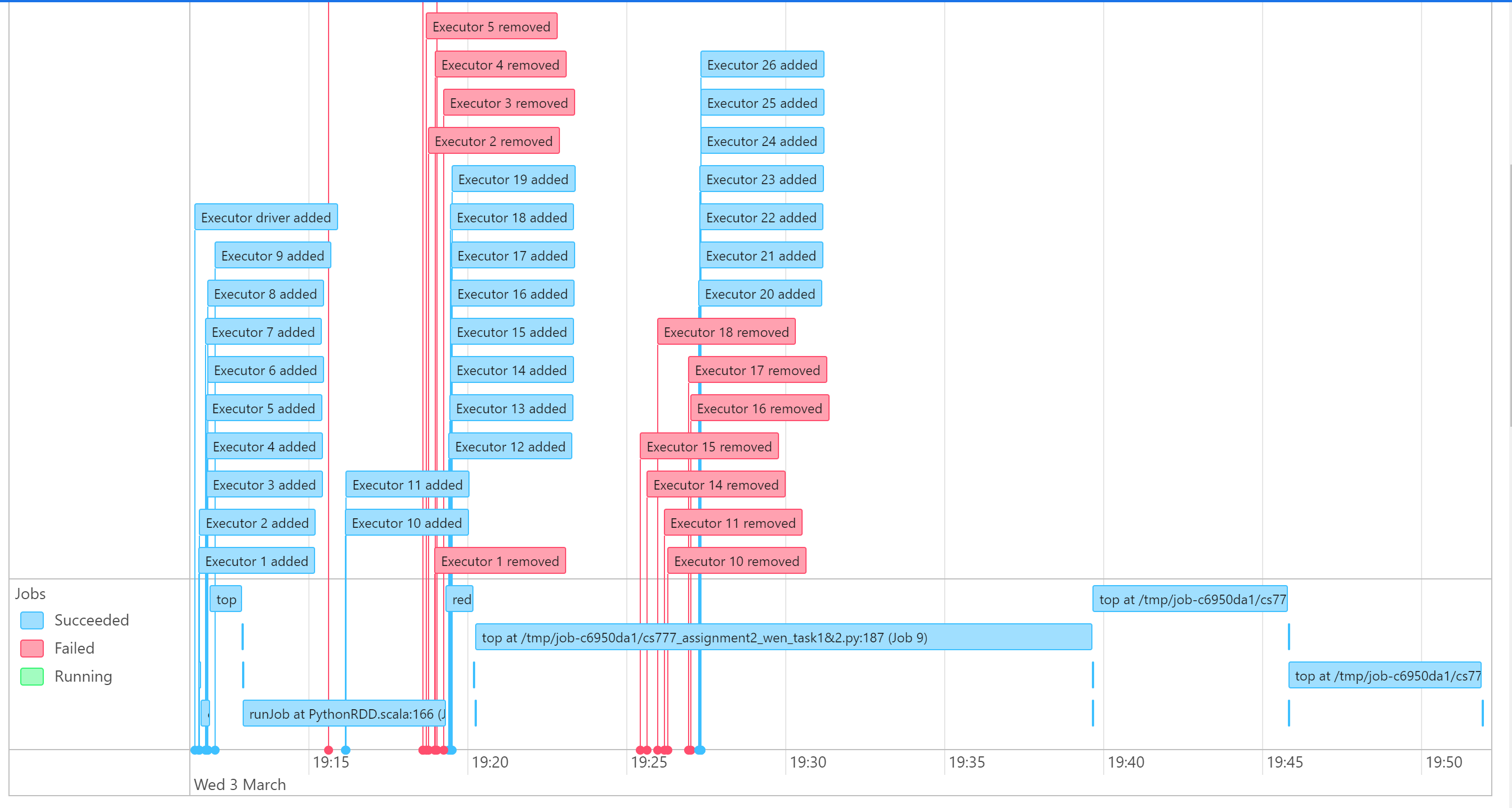
print(getPrediction('Sport Basketball Volleyball Soccer', 10))

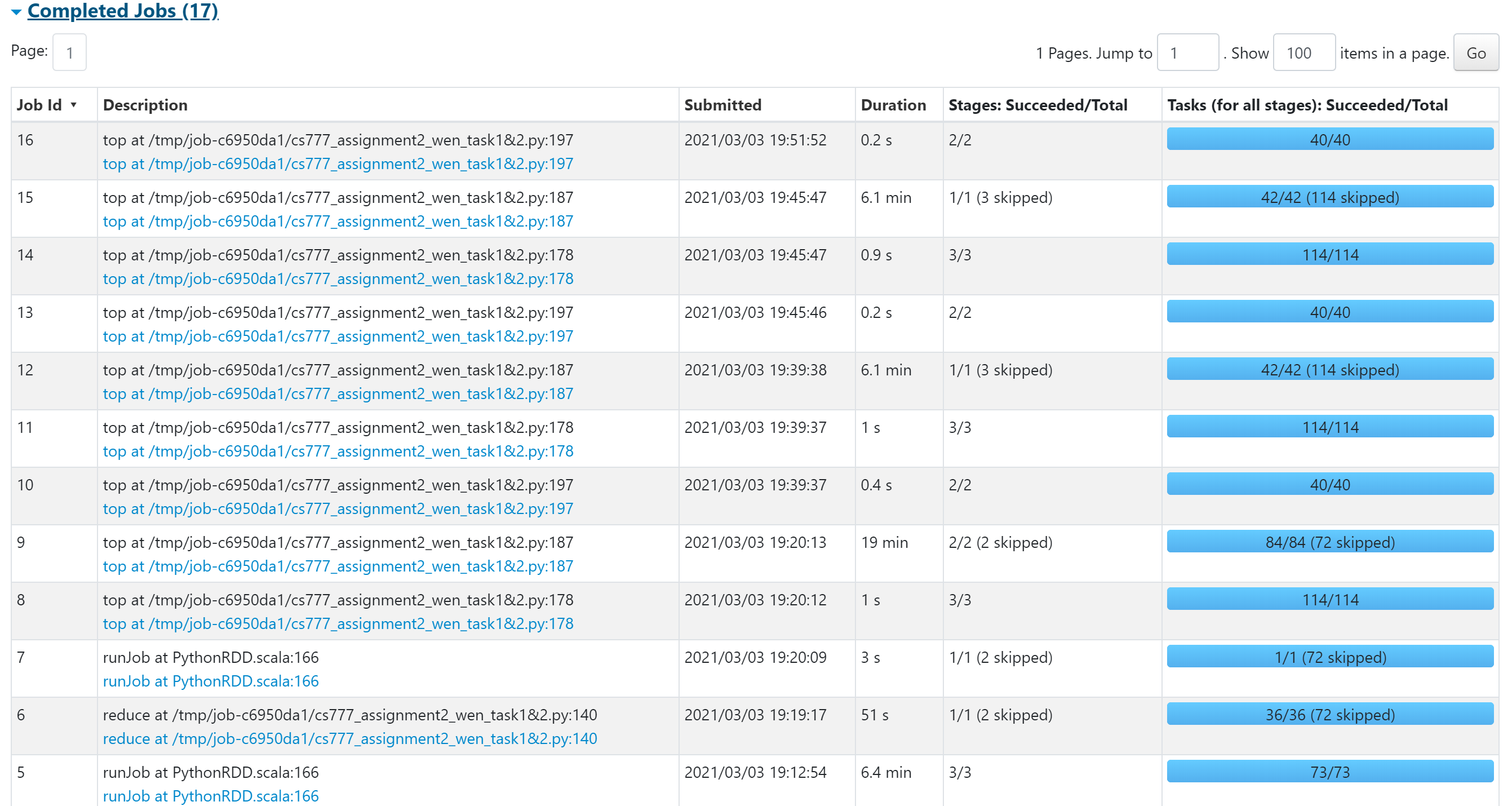
print(getPrediction('What is the capital city of Australia?', 10))

print(getPrediction('How many goals Vancouver score last year?', 10))

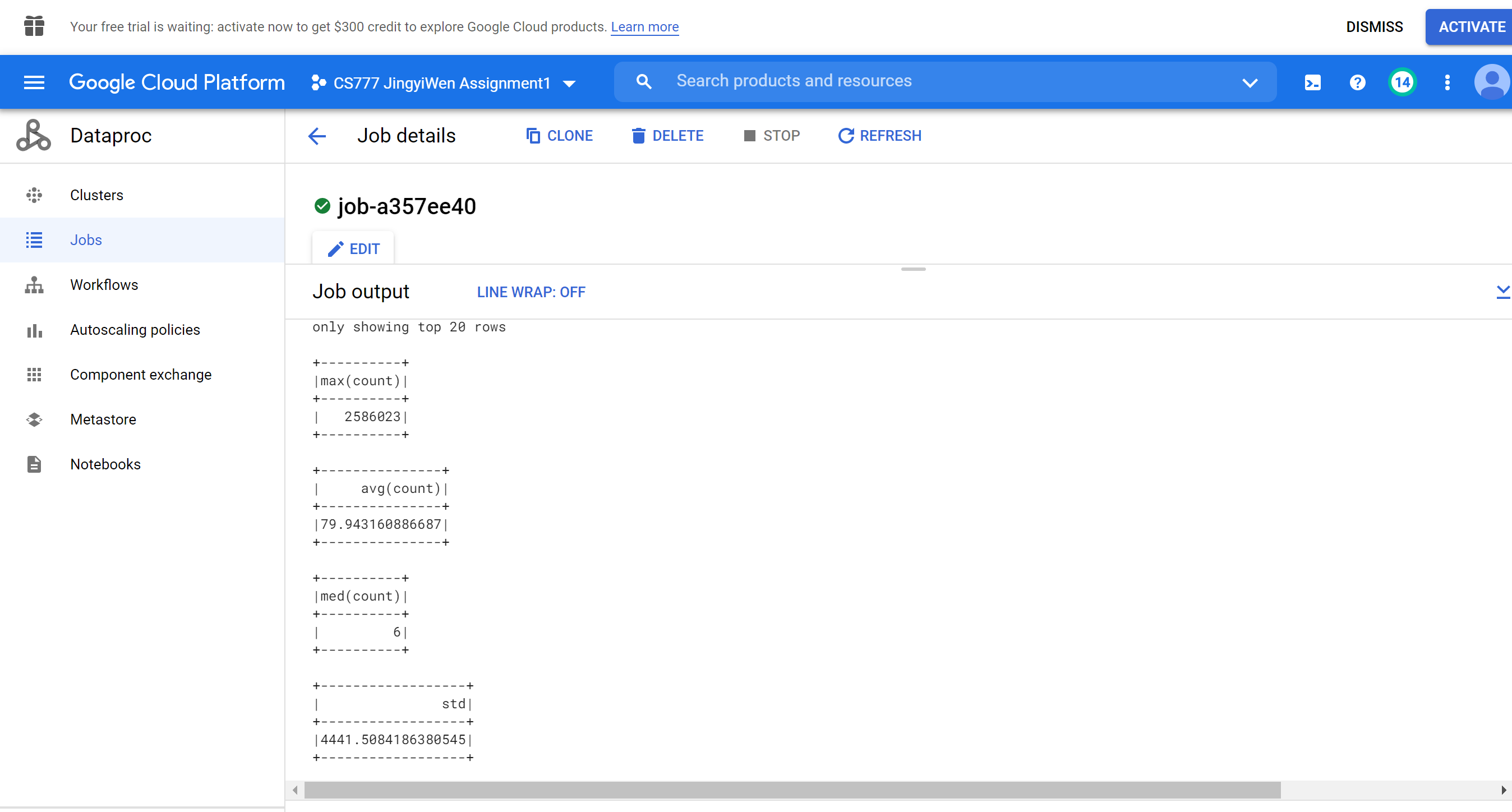


Spark History:

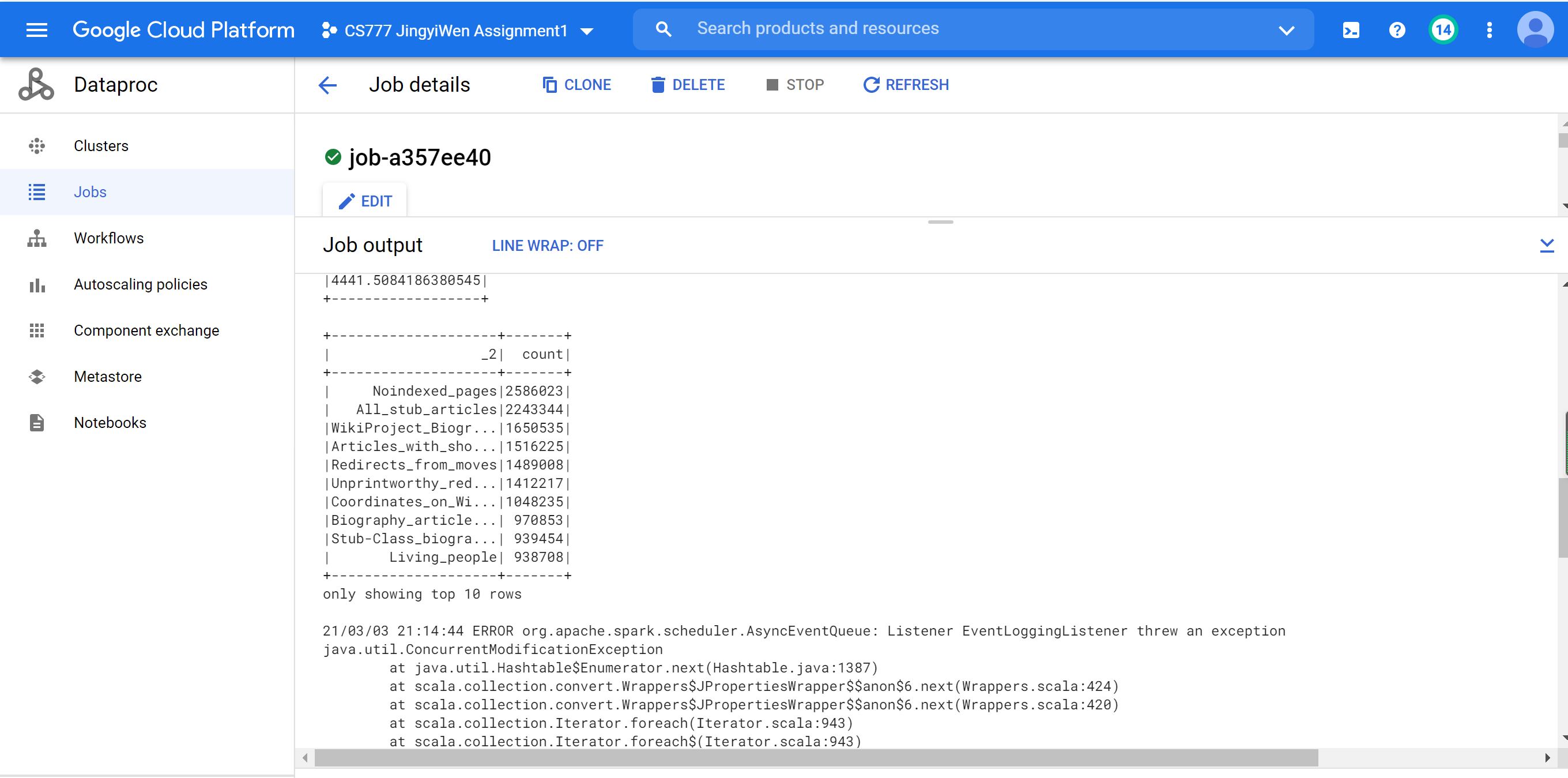




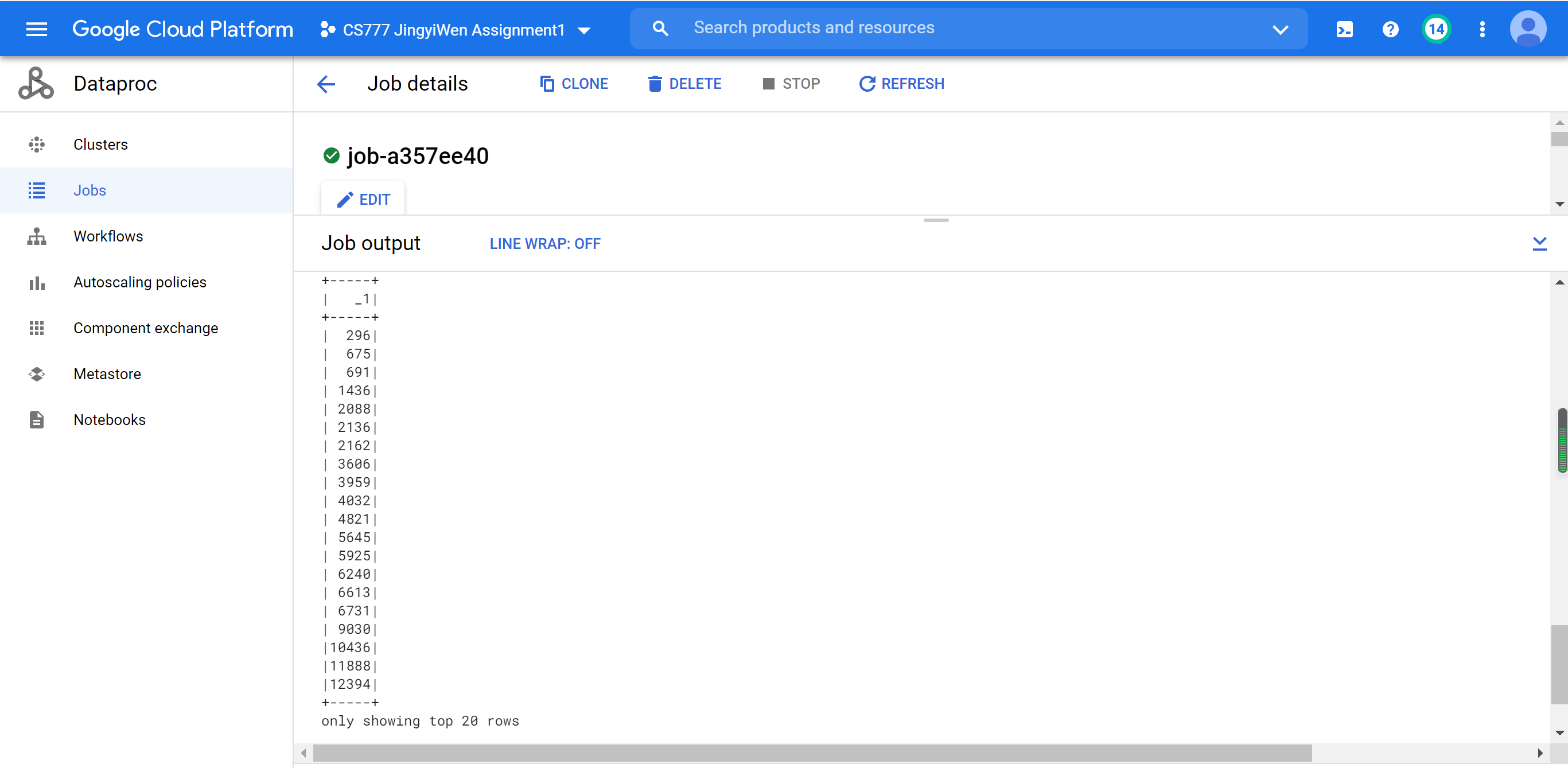
Task3.1



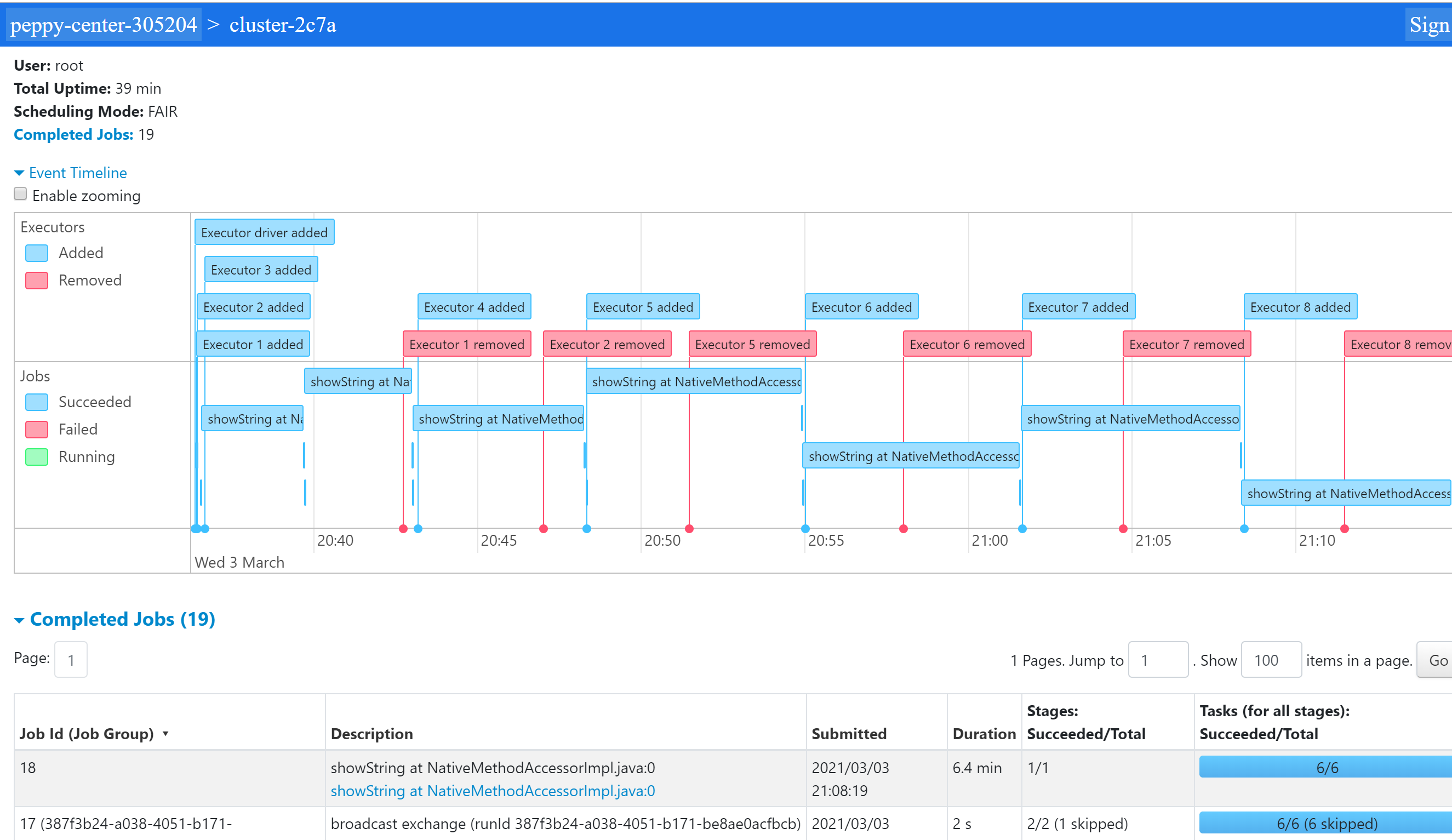
Task 3.2



Task 3.3



Spark History:



Task 4.1

No, the results won’t be changed heavily. The reason is that stop words are appears almost equally in each document, which means the tf-idf value won’t be changed if I cleaned stop words in all documents.

Task 4.2

Yes, my results will be changed heavily. The reason is if I stemming the words, it will reduce the weight of some key words in each document, which means the distances between each pair of documents will change heavily.