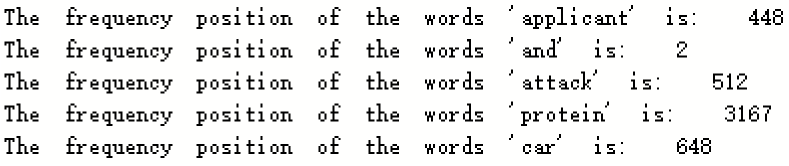
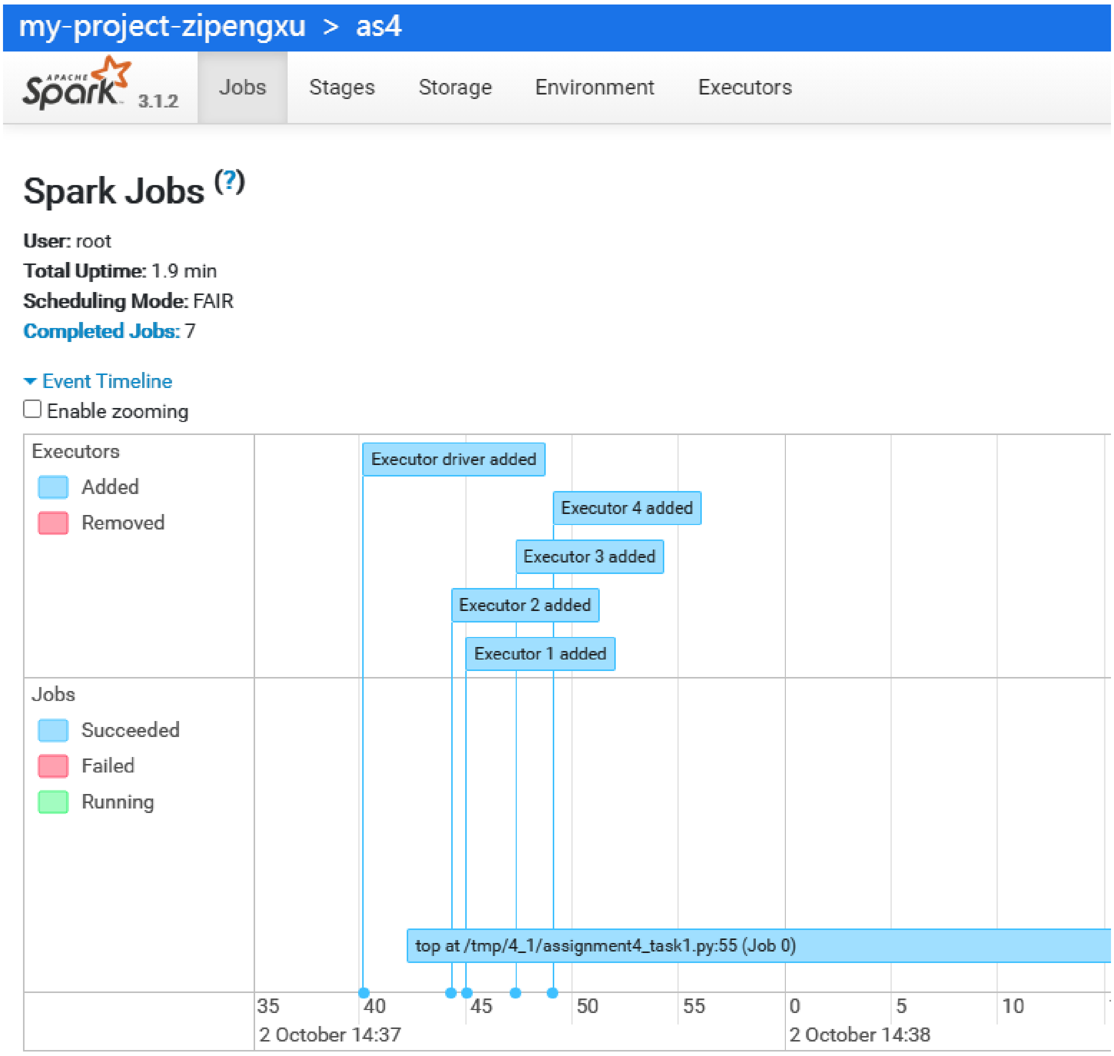
Assignment 4

Task1



```
The frequency position of the words 'applicant' is:  448
The frequency position of the words 'and' is:  2
The frequency position of the words 'attack' is:  512
The frequency position of the words 'protein' is:  3167
The frequency position of the words 'car' is:  648
```

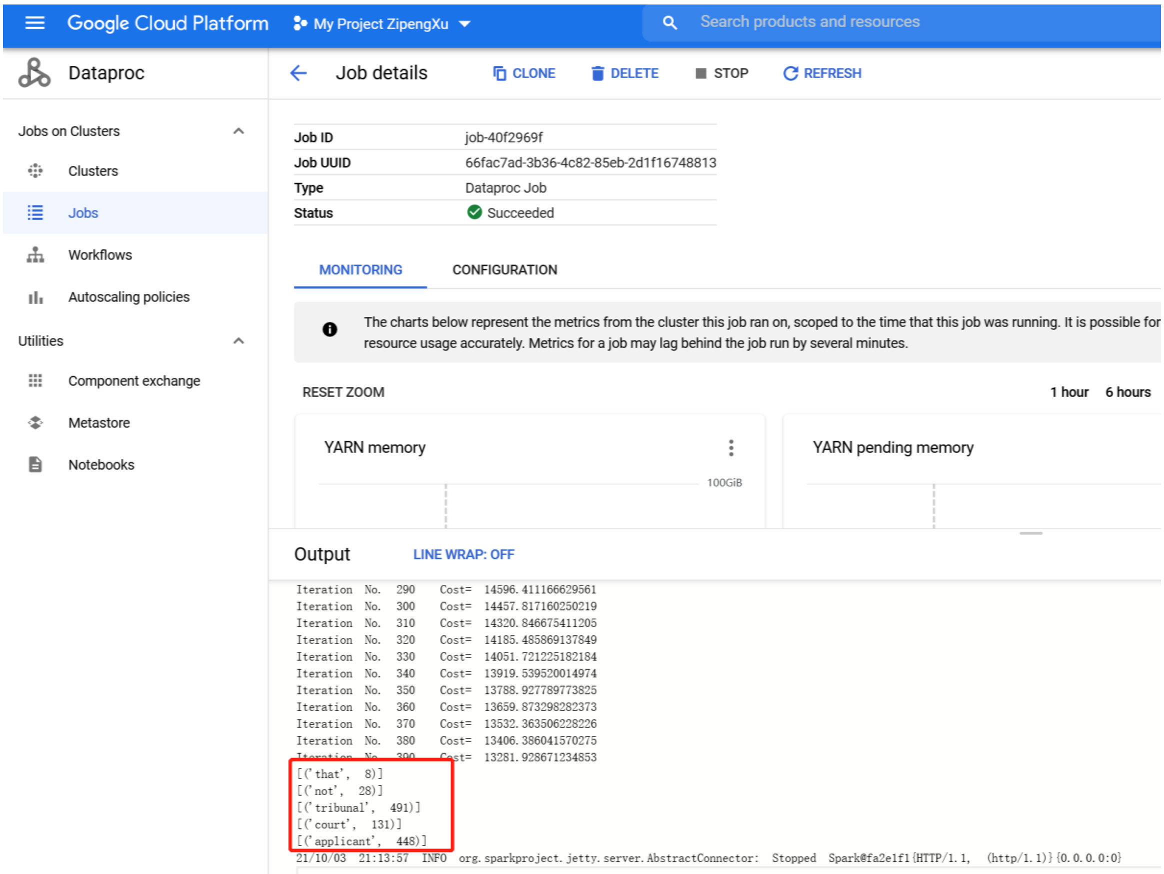


Spark Jobs (?)

User: root
Total Uptime: 1.9 min
Scheduling Mode: FAIR
Completed Jobs: 7

▾ Event Timeline
□ Enable zooming

Task2



Job ID	job-40f2969f
Job UUID	66fac7ad-3b36-4c82-85eb-2d1f16748813
Type	Dataproc Job
Status	Succeeded

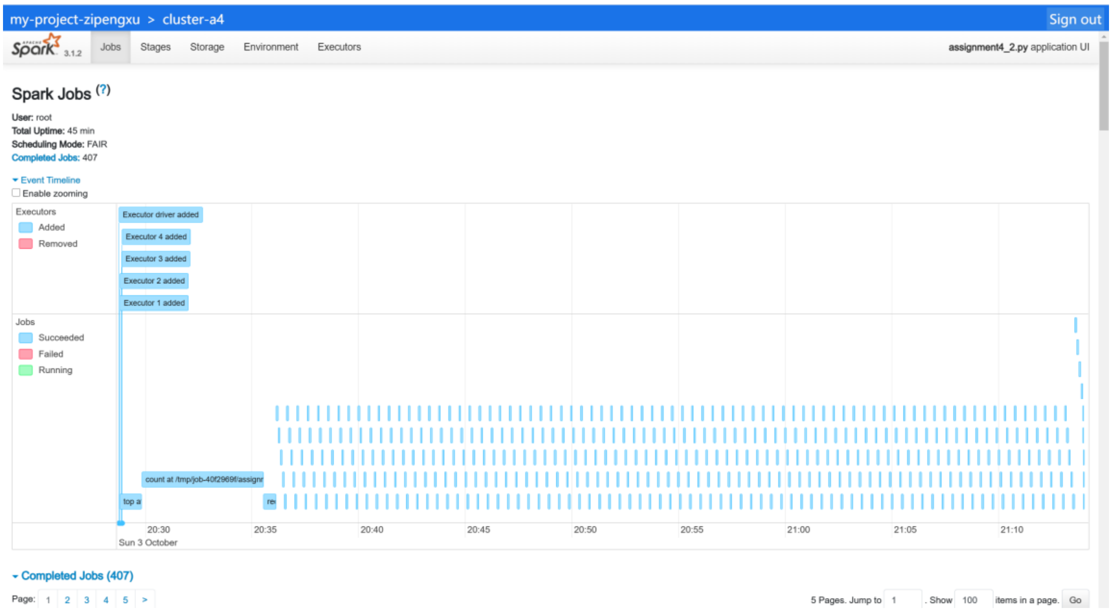
The charts below represent the metrics from the cluster this job ran on, scoped to the time that this job was running. It is possible for
resource usage accurately. Metrics for a job may lag behind the job run by several minutes.

Output

```
Iteration  No.  290   Cost=  14596.411166629561
Iteration  No.  300   Cost=  14457.817160250219
Iteration  No.  310   Cost=  14320.846675411205
Iteration  No.  320   Cost=  14185.485869137849
Iteration  No.  330   Cost=  14051.721225182184
Iteration  No.  340   Cost=  13919.539520014974
Iteration  No.  350   Cost=  13788.927789773825
Iteration  No.  360   Cost=  13659.873298282373
Iteration  No.  370   Cost=  13532.363506228226
Iteration  No.  380   Cost=  13406.386041570275
Iteration  No.  390   Cost=  13281.928671234853
[('that',  8)]
[('not',  28)]
[('tribunal',  491)]
[('court',  131)]
[('applicant',  448)]
21/10/03  21:13:57  INFO  org.sparkproject.jetty.server.AbstractConnector:  Stopped  Spark@fa2e1f1{HTTP/1.1,  (http/1.1)}{0.0.0.0:0}
```

Top 5 words:

[‘that’, ‘not’, ‘tribunal’, ‘court’, ‘applicant’]



Spark Jobs (?)

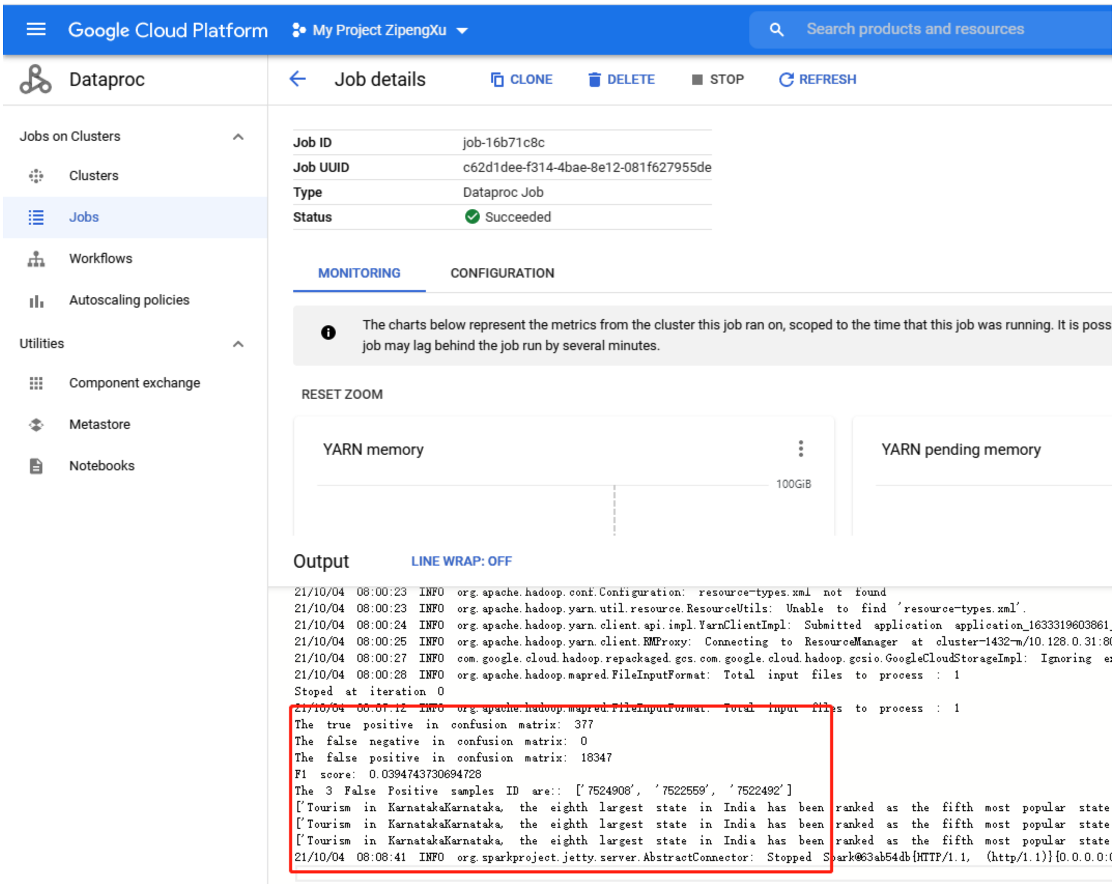
User: root
Total Uptime: 45 min
Scheduling Mode: FAIR
Completed Jobs: 407

☐ Enable zooming

Completed Jobs (407)

5 Pages. Jump to 1 . Show 100 items in a page.

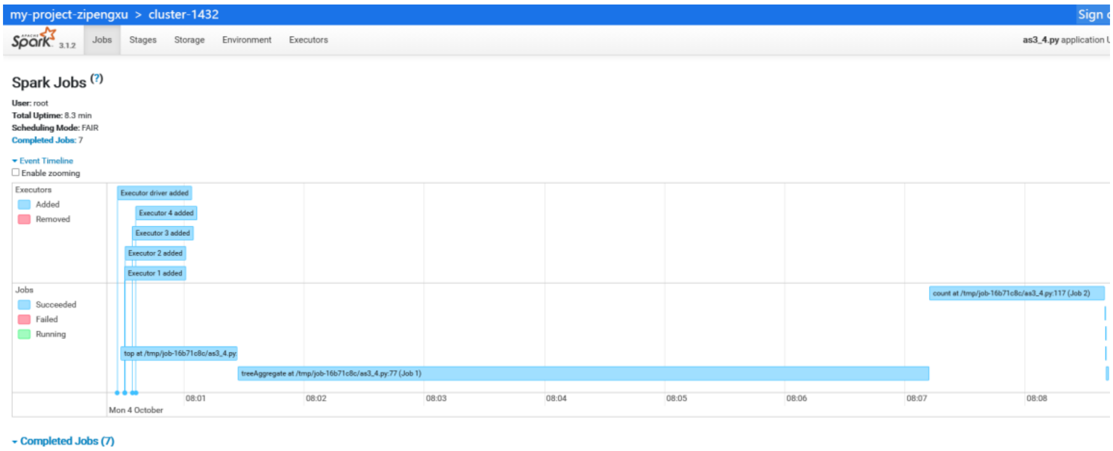
Task3



Job ID job-16b71c8c
Job UUID c62d1dee-f314-4bae-8e12-081f627955de
Type Dataproc Job
Status Succeeded

The charts below represent the metrics from the cluster this job ran on, scoped to the time that this job was running. It is poss
job may lag behind the job run by several minutes.

```
21/10/04  08:00:23  INFO  org.apache.hadoop.conf.Configuration:  resource-types.xml  not  found
21/10/04  08:00:23  INFO  org.apache.hadoop.yarn.util.resource.ResourceUtils:  Unable  to  find  'resource-types.xml'.
21/10/04  08:00:24  INFO  org.apache.hadoop.yarn.client.api.impl.YarnClientImpl:  Submitted  application  application_1633319603861_
21/10/04  08:00:25  INFO  org.apache.hadoop.yarn.client.RMProxy:  Connecting  to  ResourceManager  at  cluster-1432-m/10.128.0.31:80
21/10/04  08:00:27  INFO  com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageImpl:  Ignoring  e
21/10/04  08:00:28  INFO  org.apache.hadoop.mapred.FileInputFormat:  Total  input  files  to  process  :  1
Stoped  at  iteration  0
21/10/04  08:07:12  INFO  org.apache.hadoop.mapred.FileInputFormat:  Total  input  files  to  process  :  1
The  true  positive  in  confusion  matrix:  377
The  false  negative  in  confusion  matrix:  0
The  false  positive  in  confusion  matrix:  18347
F1  score:  0.0394743730694728
The  3  False  Positive  samples  ID  are::  ['7524908',  '7522559',  '7522492']
['Tourism  in  KarnatakaKarnataka,  the  eighth  largest  state  in  India  has  been  ranked  as  the  fifth  most  popular  state
['Tourism  in  KarnatakaKarnataka,  the  eighth  largest  state  in  India  has  been  ranked  as  the  fifth  most  popular  state
['Tourism  in  KarnatakaKarnataka,  the  eighth  largest  state  in  India  has  been  ranked  as  the  fifth  most  popular  state
21/10/04  08:08:41  INFO  org.sparkproject.jetty.server.AbstractConnector:  Stopped  Spark@63ab54db{HTTP/1.1,  (http/1.1)}{0.0.0.0:
```



Spark Jobs (?)

User: root
Total Uptime: 8.3 min
Scheduling Mode: FAIR
Completed Jobs: 7

F1-score of my classifier:
0.0394743730694728

I sample 3 articles that my model though were Australian court case (id are :['7524908', '7522559', '7522492']). The reasons that my model was fooled are as followed:

1. I pick these words which are highly correlative with courts articles are basicially included in legal or court cases article, which means it is unlikely to take 'Australian' court cases article precisely from the huge Wikipedia articles dataset. If I weight some words higher,

like ‘Australian’, in my model, I may be more likely to pick Australian court cases precesly.

2. The learning rate I select is based on the training model using small unbalance dataset. I might take a relatively large learning rate in my model, which means I fail to get the local minimum variable in gradient descent. So my model could be improved by adapting parameters.