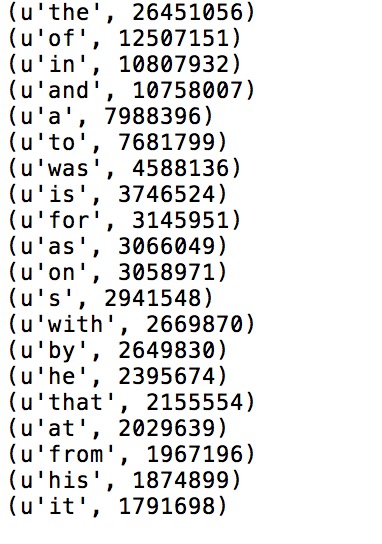
Kun Zhao

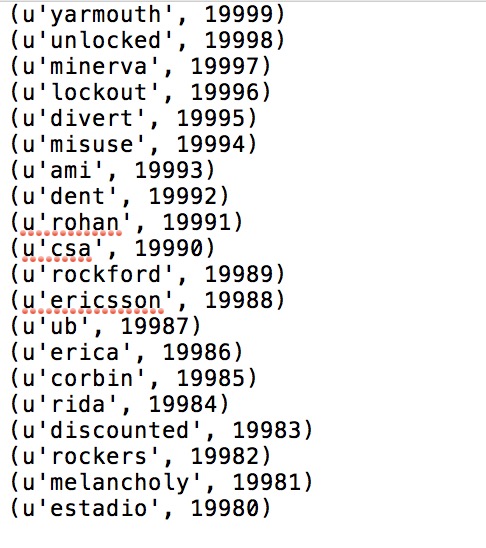
Github username: kunkunbu

**Task 1:**

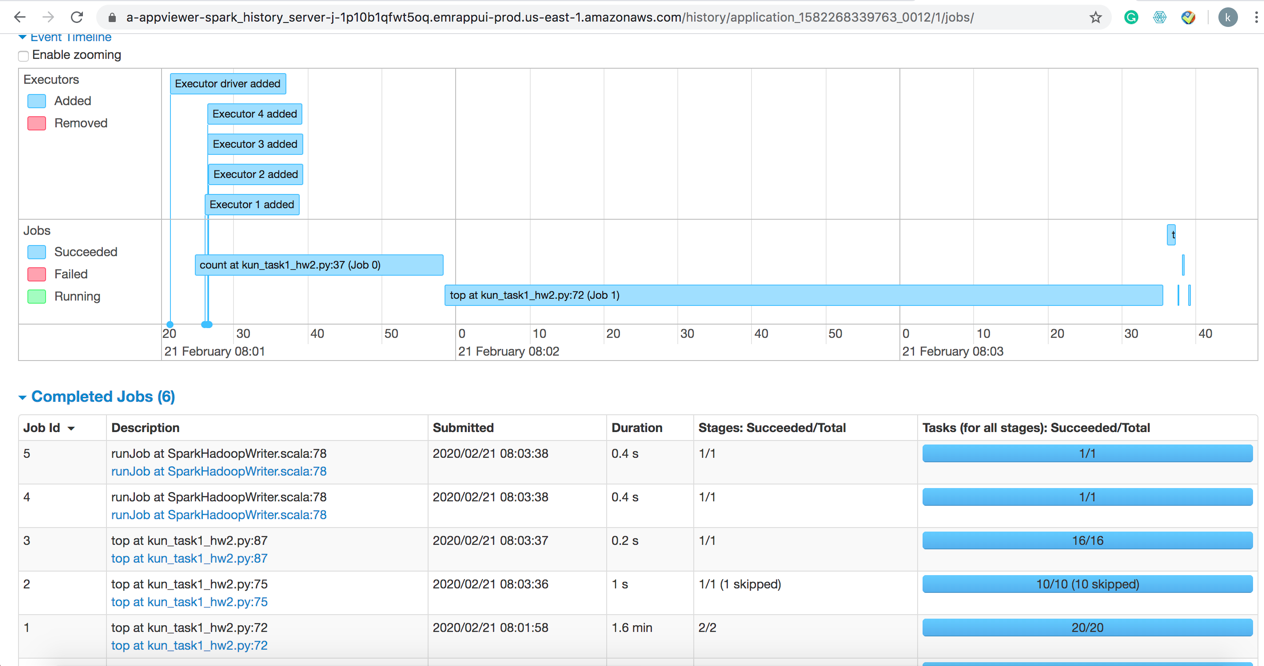
Result 1: top 20 most frequent words

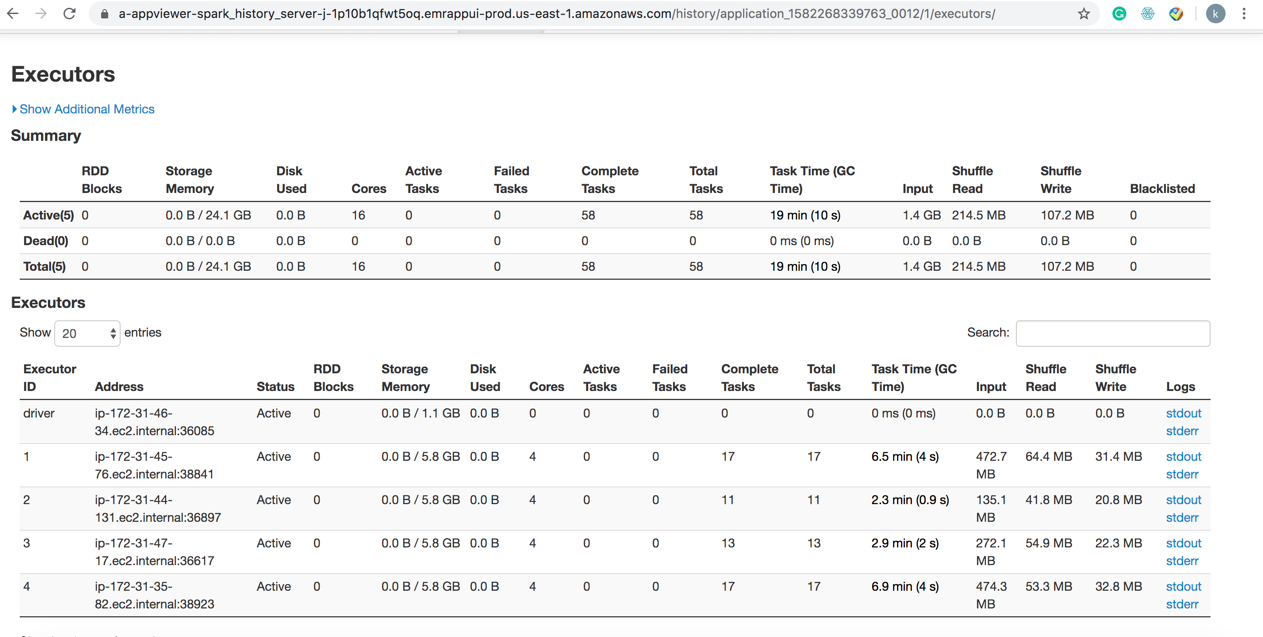


Result 2: last 20 words in your 20k feature vector



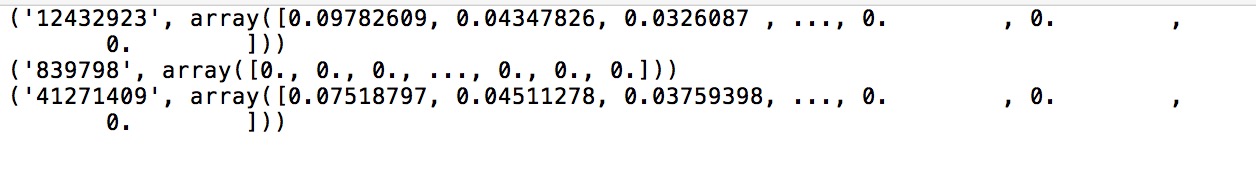
Spark history:





**Task 2:**

Result 1: allDocsAsNumpyArrays.take(3)

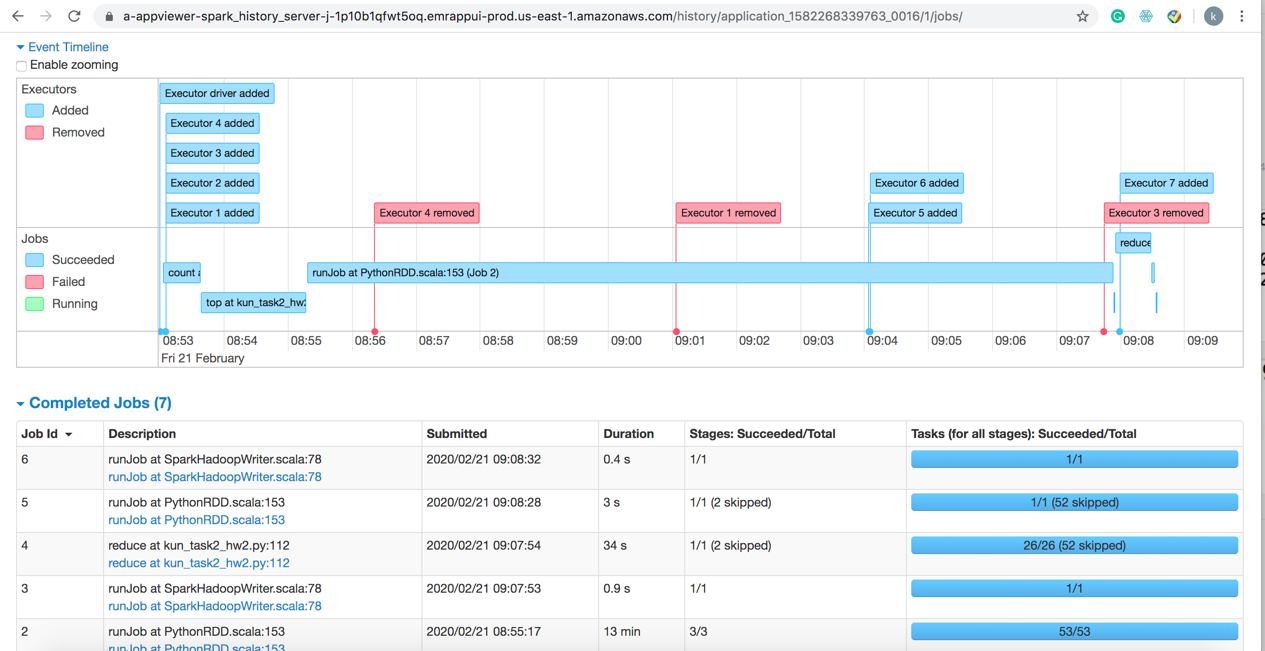


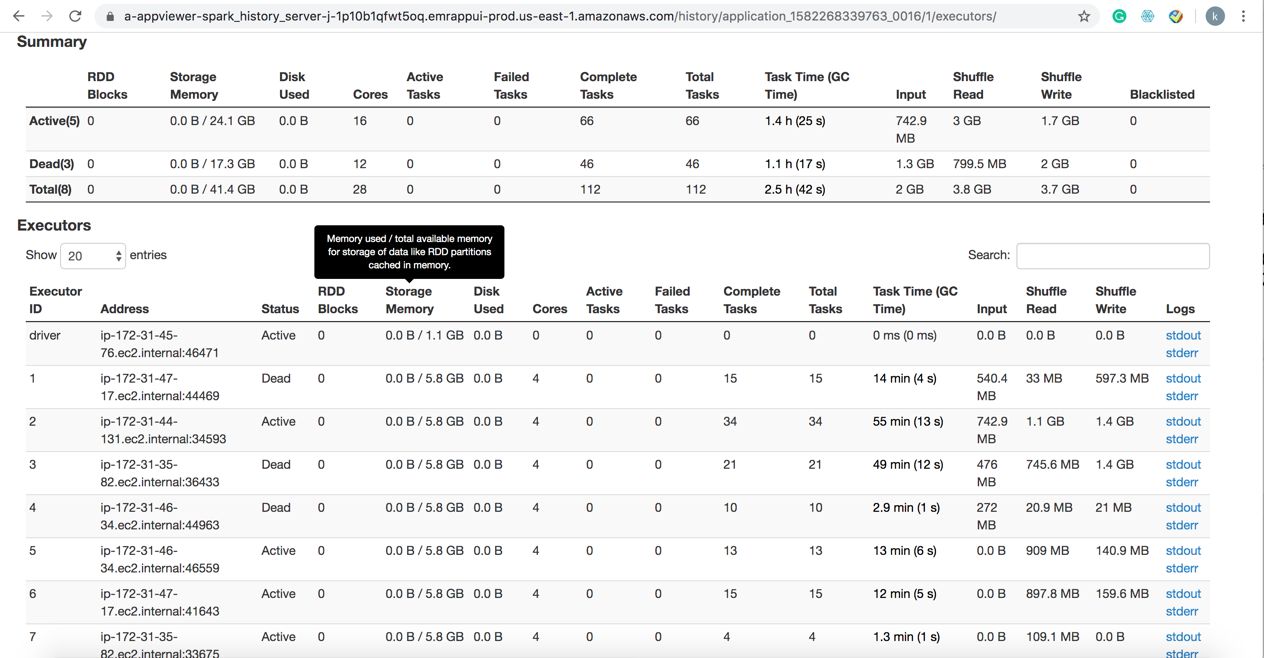
Result 2: allDocsAsNumpyArraysTFidf.take(2)

A screenshot of a cell phone

Description automatically generated

Spark history:





**Task 3:**

**Results:**

1.getPrediction('Sport Basketball Volleyball Soccer', 10)

A screenshot of a cell phone

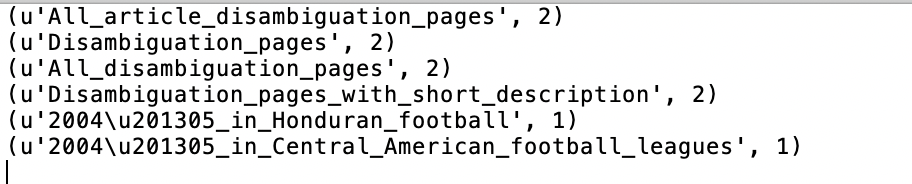
Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

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



**Spark history:**

A screenshot of a cell phone

Description automatically generated

**A screenshot of a cell phone

Description automatically generated**

**Task 4: using dataframe to predict category:**

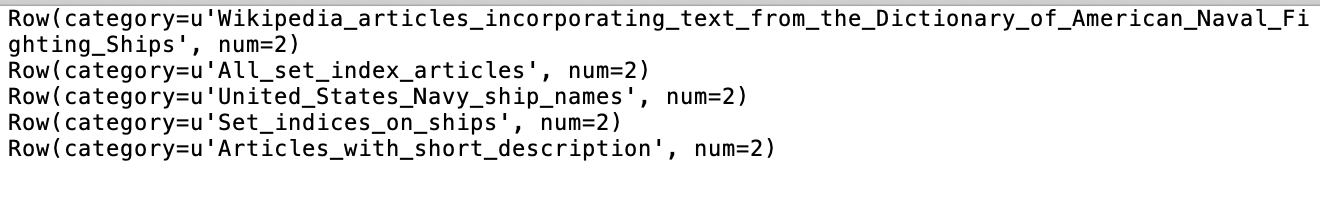
**Small dataset Results:**

1.getPrediction('Sport Basketball Volleyball Soccer', 10)

A screenshot of a cell phone

Description automatically generated

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



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

A screenshot of a cell phone

Description automatically generated

Spark history:

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

**Large data set:**

I met memory error when I run the code. I have tried a lot of methods like reduce(), toPandas(). However, I can’t figure this problem now. I will continue doing this!!

**A screenshot of a cell phone

Description automatically generated**

Error reason:

**A close up of a logo

Description automatically generated**