

Task 1 & Task 2:

I tried to print the answers in my code, but I don't know how to grab it from the outputs.

Task 3.1:

('Max number of wikipedia categories of wikipedia pages is Row(max(Counts)=375092)',
'Average number of wikipedia categories of wikipedia pages is
Row(avg(Counts)=13.044221049294068)', 'Standard Deviation of wikipedia categories of
wikipedia pages is Row(stddev(Counts)=688.6245334448599)')

Task 3.2:

Top 10 mostly used Wikipedia categories are:

```
Row(Category='All_stub_articles', Counts=375092)
Row(Category='Articles_with_short_description', Counts=234722)
Row(Category='Coordinates_on_Wikidata', Counts=176867)
Row(Category='Living_people', Counts=138238)
Row(Category='Wikipedia_articles_with_VIAF_identifiers', Counts=114141)
Row(Category='Wikipedia_articles_with_WorldCat-VIAF_identifiers', Counts=114137)
Row(Category='Articles_with_'species'_microformats', Counts=87438)
Row(Category='Wikipedia_articles_with_LCCN_identifiers', Counts=86909)
Row(Category='Wikipedia_articles_with_ISNI_identifiers', Counts=75960)
Row(Category='Webarchive_template_wayback_links', Counts=70764)
```

Task 4

```
from pyspark.ml.feature import StopWordsRemover
remover = StopWordsRemover()
allWords = remover.transform(allWords)
```

Task 4.1

Since this is not a sentiment analysis, I think removing stop words will not make heavily change for the results. It will only reduce the size of the doc and increase the running time.

I think stemming will change the results heavily, since it increased the similarity of each words, which will cause more false positive in the prediction.

The screenshot displays the Databricks Event Timeline for cluster-8865. The timeline is divided into two main sections: Executors and Jobs. The Executors section shows the lifecycle of individual executors, with blue boxes indicating when an executor was added and red boxes indicating when it was removed. The Jobs section shows the lifecycle of jobs, with blue boxes indicating when a job succeeded, red boxes indicating when it failed, and green boxes indicating when it was running. A specific job, 'runJob at PythonRDD.scala:166 (Job 15)', is highlighted as running. The timeline spans from Monday, September 20, 22:00 to Tuesday, September 21, 02:00.

Executors Timeline:

- Added:** Executor 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15.
- Removed:** Executor 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15.

Jobs Timeline:

- runJob at PythonRDD.scala:166 (Job 15):** Running from approximately 22:00 on Monday to 01:00 on Tuesday.
- Other Jobs:** Several other jobs are shown as failed (red boxes) or succeeded (blue boxes) at various points in time.