Running Kmeans Clustering in Spark

3 questions

1.

After running the Kmeans pyspark script, with the number of clusters k=3: What is the SSE?

- 1.96
- 2.03
- 4.54

2.

Change the number of clusters to k=4 and rerun the training command and get the SSE (or rerun the whole script).

What is the approximate SSE for k=4?

(note that the random data is generated with seed values set so it should be the same across runs)

- 4.96
- 1.65
- 1.50

3.

What is the relationship between SSE when k=1 and the variance of each dimension?

- SSE = total variance of each dimension
- SSE is not related to variance of each dimensin

SSE is 2 times the variance of each dimension

Submit Quiz