K-means using Spark

Data

The Iris data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters.

Downloaded from here: https://archive.ics.uci.edu/ml/datasets/iris

TODO

1. Import the Iris dataset.
2. Build a K-means model to classify the species of Iris. You can choose a k value randomly at this step.
3. Report the original performance using Silhouette score.
4. Try to improve the performance of the original model by trying at least 10 different k values.
5. Select the best k based on step 4 and print out the following sentence in your code:

   “k=xx gives the best performance, Silhouette =xx "

   (replace xx with your own numbers)