Assignment 7: Clustering

Ankita, Jiangtao

11/03/2017

Objective

Given a million song dataset,perform iterative computations, computations on a graph, learning basic clustering algorithms.

Dataset

We are using a dataset based on the metadata in the Million Song Database.

Preparing Data

- Checks done to remove empty entries in the dataset.
- Check done to dis-regard non-float entries in `duration` , `tempo` , `song_hotness` columns.

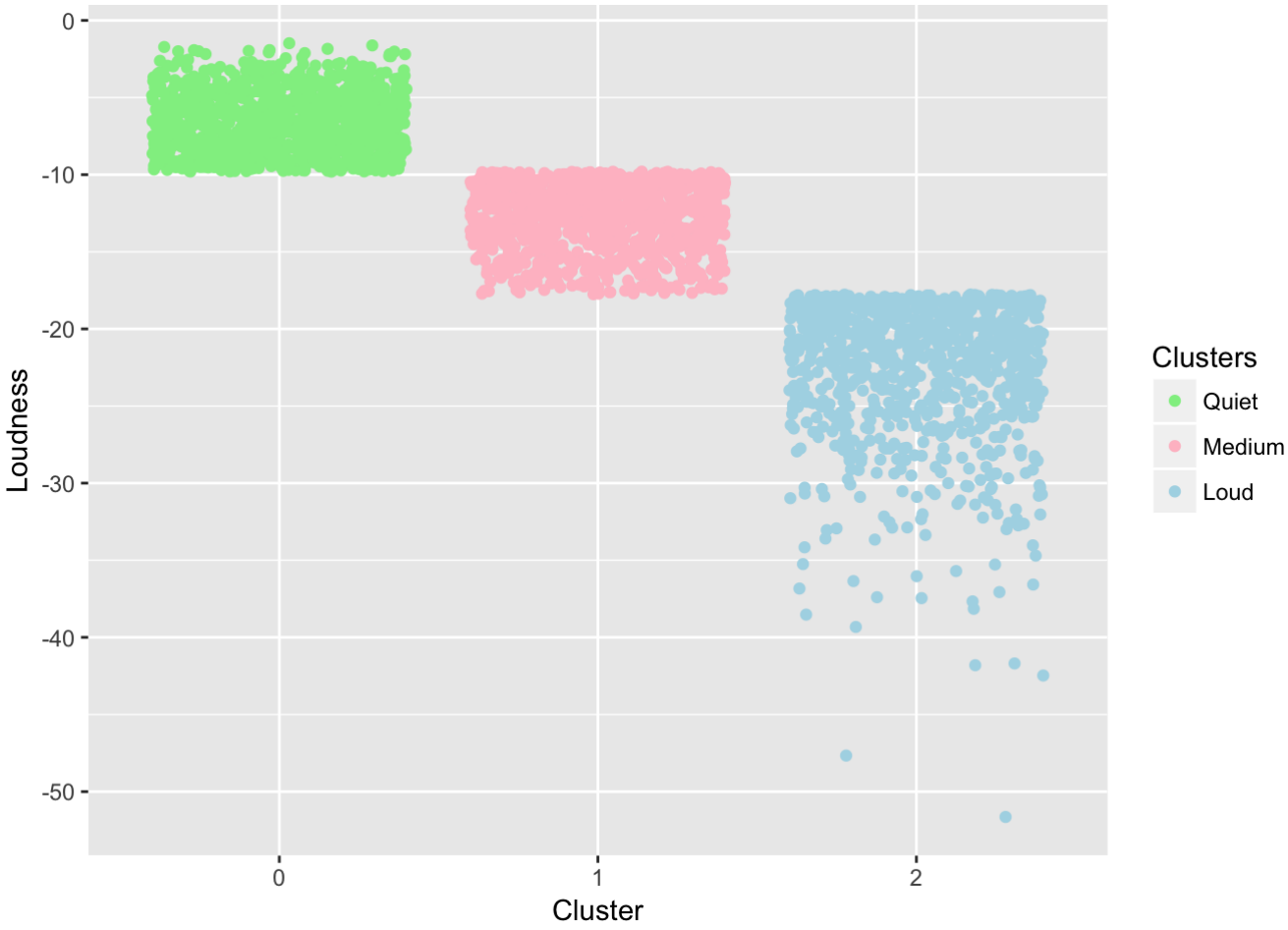
Implementation

Subproblem 1 : Clustering

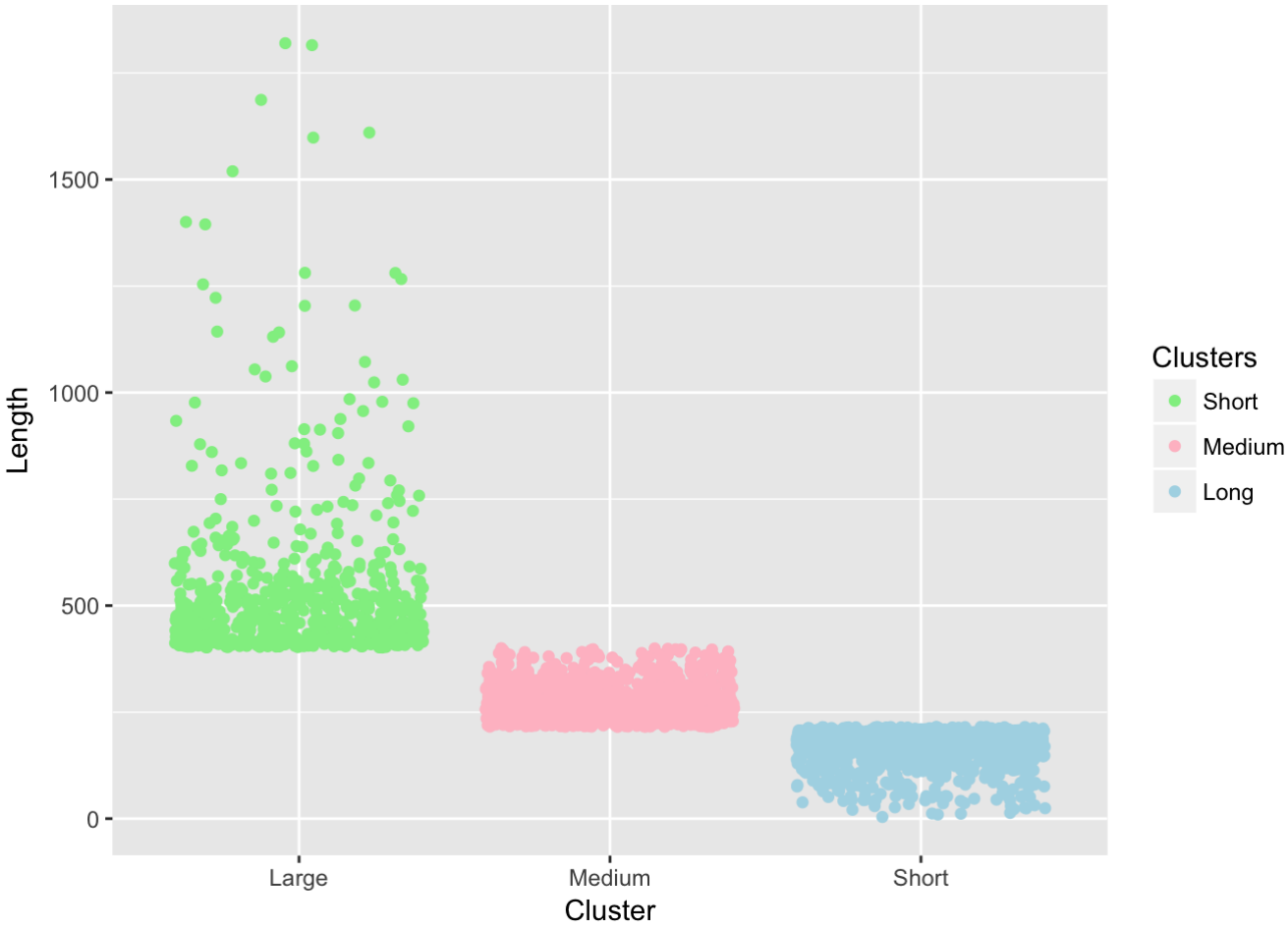
K-Means Clustering - - For clustering using k-means, we Our solution first takes the two files ‘song_info.csv’ and ‘artist_terms.csv’ and converts it into `mapPartitions` that converts each partition of the source RDD into multiple elements of the result. - After that, for each task like finding the distincts and top’5, we select only the required column from the RDD thus avoiding reading all the columns of the dataset everytime. This improves the execution time as well. Hierarchical Agglomerative Clustering -

Output

Fuzzy loudness:



Fuzzy length:


1500
1000
500
0
Length
Large
Medium
Short
Cluster
Clusters
Short
Medium
Long

Fuzzy tempo:

Local Execution Environment Specifications:

- Macintosh 2.5Ghz i7 Quad Core
- 16 GB RAM
- macOS Sierra Version 10.12.6
- Java version : 1.8
- Scala version : 2.11.11
- Spark version : 2.2.0