Assignment 8: Ghost Protocol

Ankita

11/08/2017

Objective

Given a million song dataset,perform iterative computations, computations on a graph, learning basic clustering algorithms like K-means and Hierarchical Agglomerative Clustering.

Preparing Data

- Checks done to remove 'N/A', '0' in the dataset.
- Check done to dis-regard non-float entries in `duration`,`loudness`, `tempo`, `song_hotness` columns.

Implementation

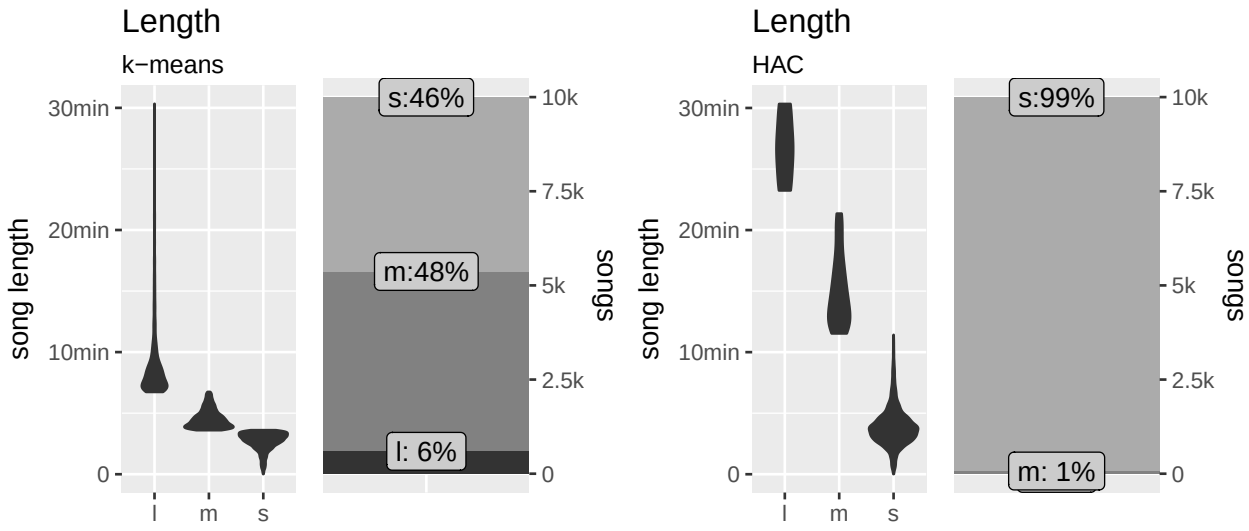
-K-Means Clustering: For clustering using k-means, my solution first takes the file 'song_info.csv' and cleans it to remove dirty data. Then for choosing K, I have used the first 3 values as k=3 from the dataset as another approach like choosing the `minimum, maximum and mean` might actually have values that are not real data points. Thus after the inital centroids are chosen,K-mean algorithm is run for n= 10 times. In each step, based on its distance form the centroids, the points are classied to be in one of the three clusters. In each steps I take the mean of the data points in the cluster and make it the new centroid. This process repeats for 10 steps or till the clusters converge.

-HAC Clustering: Here initially I sort the dataset and then each data point is considered one cluster and I go on repeating the HAC till it converges to 3 cluster.For that,in each iteration I cluster two points that are closest(based on Euclidean distance) in the entire dataset.Then I combine the closest pair with the rest of the cluster points and iterate over to find the next pair.

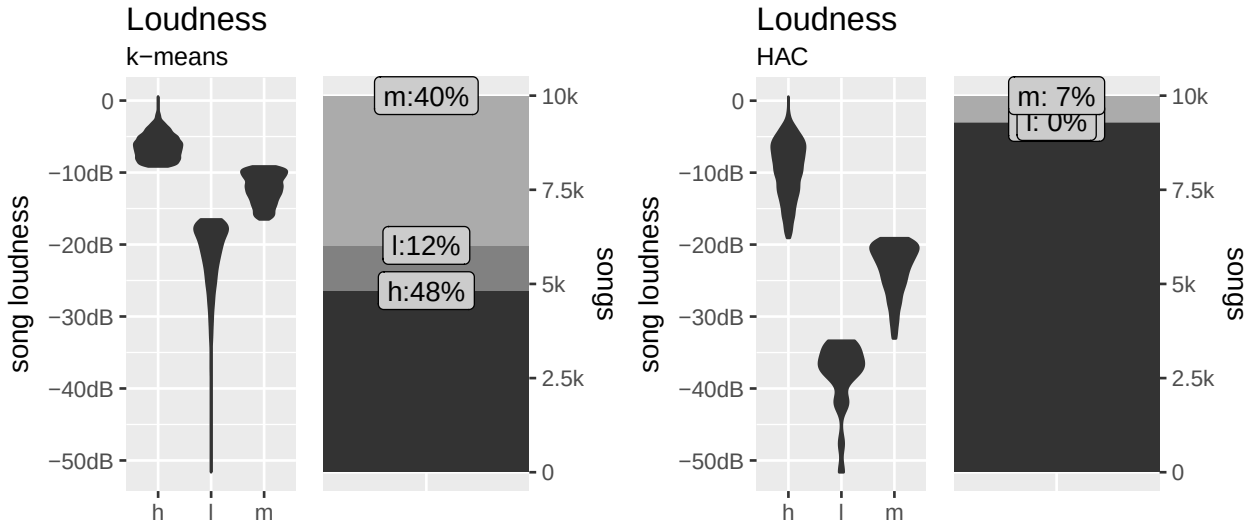
Results

The program is run on the small dataset and the results can be seen below:

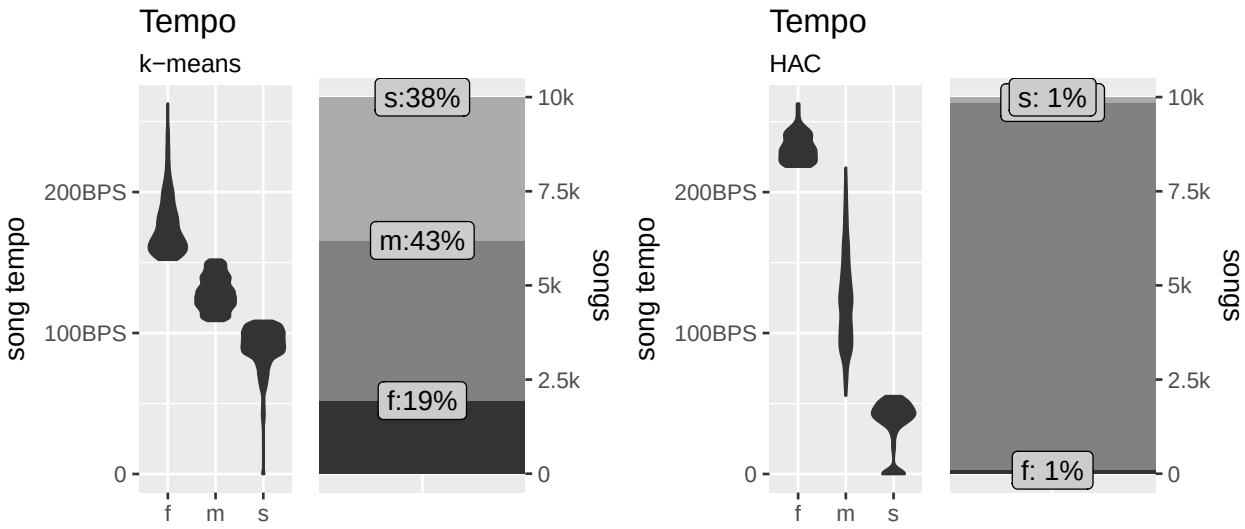
1. In the graph below, l='Long', m='medium', s="small". It is seen that for K-means majority songs are small and medium while for HAC, as the values are near close, all the values go in the small cluster.



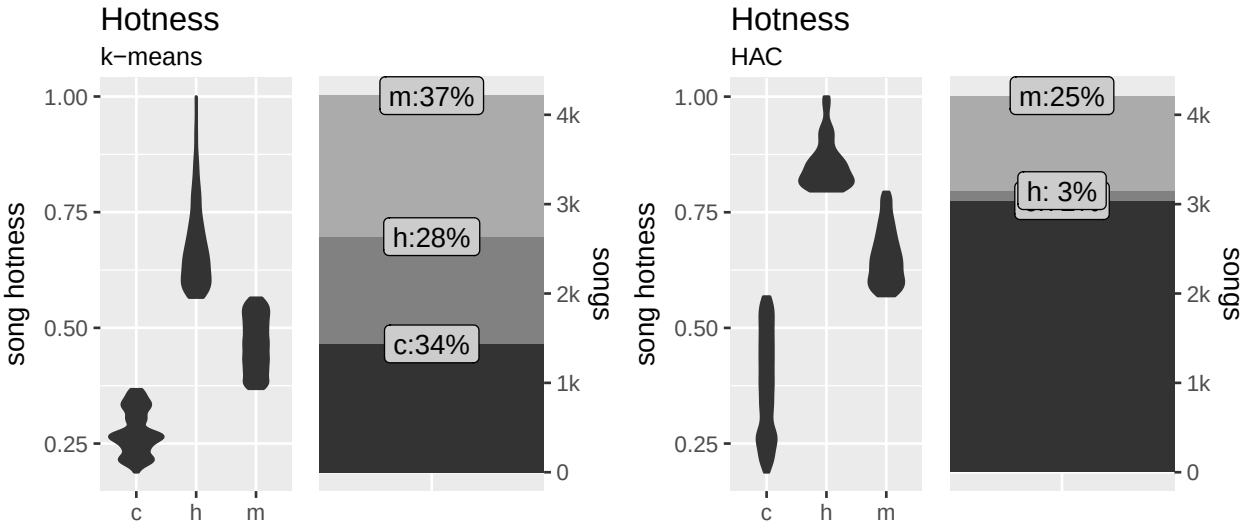
2. In the graph below, l=‘Low’, m=‘medium’, h=“high”. It is seen that for majority songs are high for both Kmeans and HAC.



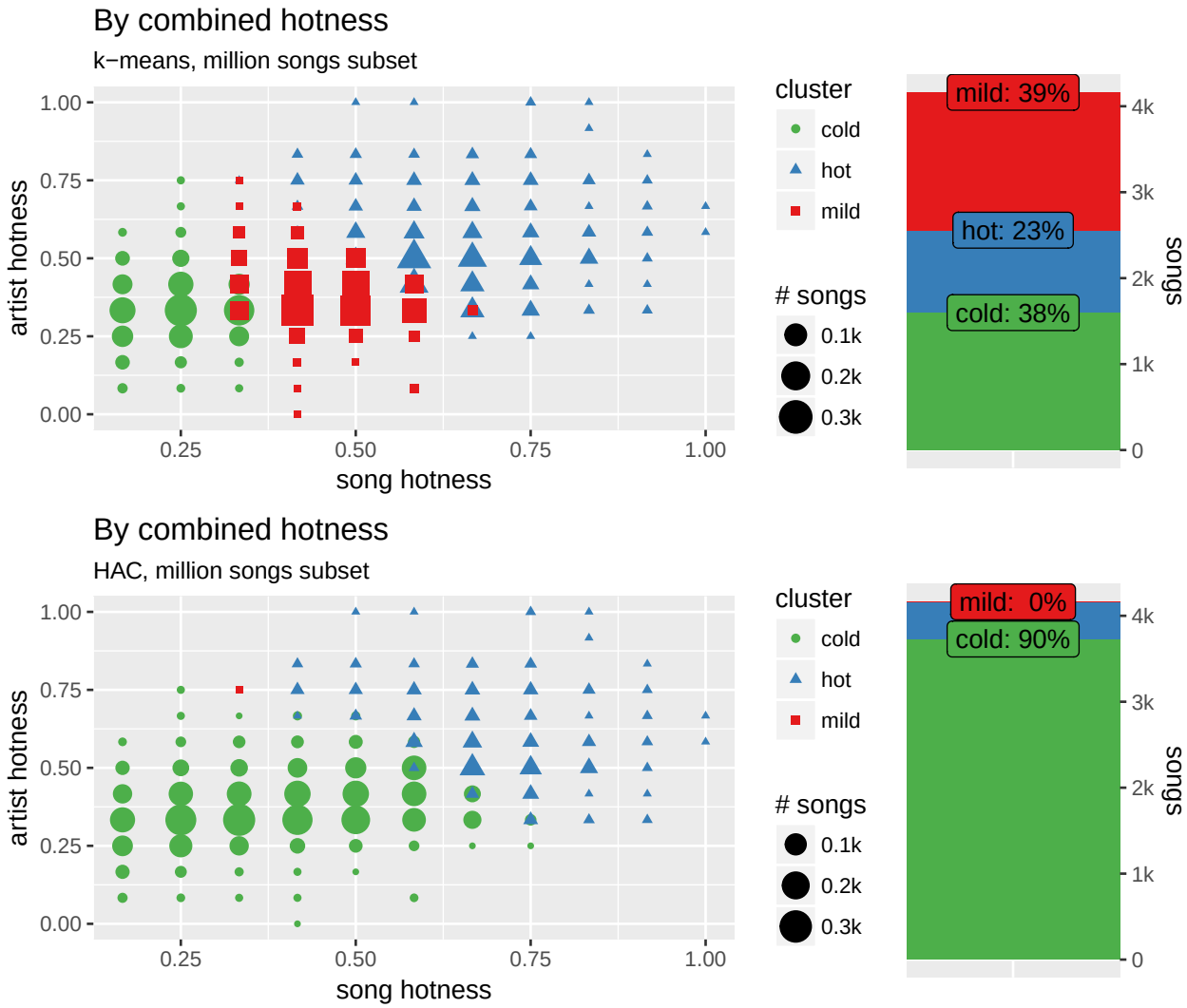
3. In the graph below, s=‘slow’, m=‘medium’, f=“fast”. It is seen that for majority songs are high for both Kmeans and HAC.



4. In the graph below, c='cool', m='mild', h="hot". It is seen that for majority songs are mild for both Kmeans while cool for HAC.



5. In the next 2 graphs below,clustering is done on 2D datasets,and c='cool', m='mild', h="hot".From the first graph on left, majority points seems to be hot but are instead mild. And for the next graph, majority points are cool.



Observation:

Performance of KMeans and HAC on small dataset: KMeans - 6 sec and HAC - 1 minute

On comparing song titles, loudness, tempo and length of the song classified above, following trends can be seen based on the table below: a. Loudness v/s Hotness: If the loudness is high then song hotness tends to be low. b. Duration v/s Hotness: If the duration of song is more, we cannot firmly say that the hotness will be high or low. Row 2 and 3 supports the ambiguity. c. Tempo v/s Hotness: If the tempo of song is more, we cannot again firmly say that the hotness will be high or low. Row 1 and 3 supports the ambiguity.

Duration	Loudness	Tempo	Title	Song.hotttnesss
314.1742	-18.674	104.803	Suar Agung	0.0637250
237.2436	-15.777	118.713	Transformation	0.0256891
269.6355	-5.388	104.038	Nothin' On You [feat. Bruno Mars] (Album Version)	1.0000000
145.0575	-6.544	150.569	Immigrant Song (Album Version)	1.0000000
232.2020	-7.375	130.060	Such Grand Ideas	0.1878950

Conclusion:

K-means is linear in the number of data objects i.e. O(n), where n is the number of data objects.The time complexity of the hierarchical clustering algorithms is quadratic i.e. O(nˆ2). Therefore, for the same amount of data, hierarchical clustering takes quadratic amount of time. Like for the bigger dataset it will take ~3 hours.As the number of records increase the performance of hierarchical algorithm goes decreasing and time for execution increased .K-mean algorithm also increases its time of execution but as compared to hierarchical algorithm its performance is better.Thus,as a general conclusion, k-mean algorithm is good for large dataset and hierarchical is good for small datasets.

Execution Environment Specifications:

OS	Architecture	Cores	RAM	Model	Processor
macOS10.12.6(16G29)	x86_64	4	16GB	MacBookPro14	2.5Ghz i7 Quad Core