Clustering of Songs with Amazon Web Services

Introduction

The objective of this project is to group a set of songs allocated in Spotify that have some similarities among them.

Every person has her/his preferences regarding the kind of music that he/she likes, but sometimes this person does not listen to different songs because the songs are not actually available or the person does not know the singer or the band. In this sense, the project offers the opportunity to group songs that are similar to facilitate the creation of your playlist of favourite songs.

Dataset

This dataset is composed of three files retrieved from Spotify with audio features for over 20k songs. The dataset can be downloaded from [1].

Every file contains 18 attributes among which are strings and integer attributes such as artist name, track id, danceability, energy, etc.

Missions

The final goal is to find out new songs that can be part of your list of favourite songs. To accomplish this task, different ways of using a clustering approach are going to be compared. On the one hand, through a clustering approach following a distributed cluster-computing framework, and on the other hand, through the use of Amazon RDS.

To do that, some tasks need to be accomplished:

1. Understand the content that is available in the dataset and clean the dataset if it is necessary.
2. Clustering following a distributed cluster-computing framework.
    a. Choose the best clustering algorithm taking into account the attributes that the dataset offers. In order to decide the best approach, it could be useful to have an overview of the different perspectives explaining the advantages and drawbacks of your decision.
    b. Evaluate your model statistically.
3. Clustering in the cloud using AWS.
    a. Create a database using Amazon RDS.
    b. Create tables to load the data of the dataset.
    c. Export the data into S3 and apply a cluster analysis in AWS using a different clustering algorithm.
    d. Analyze the outcomes.
    e. (Optional task) Put the final results into RDS for future access.
4. Use the different packages of visualization explained during the course to visualize clusters based on your findings.
5. Compare the findings of both methods using metrics, the visualizations, etc.

Deliverables

To carry out the assessment of the project, the group has to submit the following:

- A report using Google Doc and explaining the concept about the project solution and the expected division of the tasks regarding the components of the group. This document should not be longer than 10 pages (including cover, table of contents, etc).
- Collaborative work using Git with commit + push changes on a daily basis.
- A presentation explaining the thought process, your approach and the reason for this choice, your findings and the real task division in your group at the end of the project. Every group has 15 minutes per presentation and there will be 5 minutes of questions.

[1] https://www.kaggle.com/rafaelnduarte/spotify-data-with-audio-features?select=2020_spotify.csv