CSCI 4253 Project Proposal

Evan James
Mario Ramirez
Johnathan Kruse

Contents

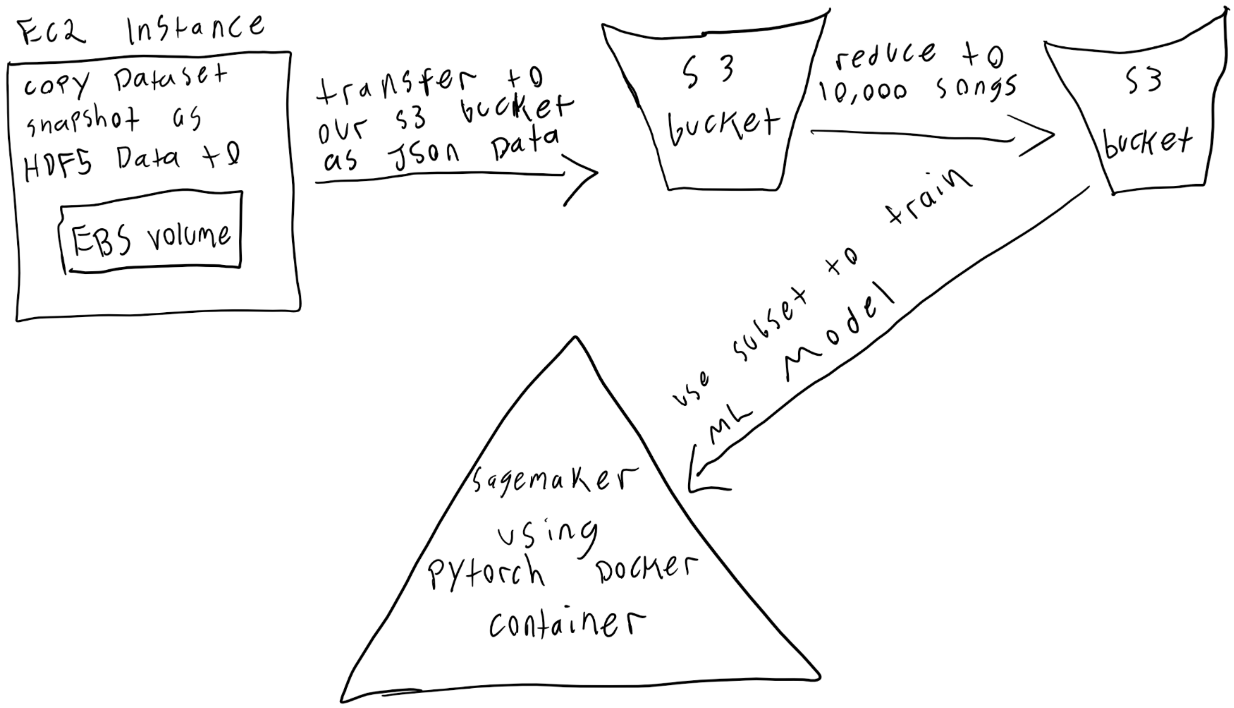
Problem description

Globally, the music industry is valued at approximately $16 billion (Forbes, 2017). Furthermore, the number of songs downloaded and streamed each year is estimated to be well over a billion and is growing each year (Nielsen 2016 Music Year-End Report). As a musician, the market has become increasingly difficult to stay relevant. One of the greatest challenges is creating a song which will become “popular” and thus incur more consumption by listeners. Our goal is to help musicians around this problem by determining both if their song will be popular, and characteristics of what would make their song more popular.

For our project, we will utilize the Million Song Dataset -- found at https://labrosa.ee.columbia.edu/millionsong/ -- to train a machine learning algorithm to predict whether songs will be popular or not. The MSDS contains over one million songs along with information of the song taken from The Echo Nest API. The song attributes include tempo, time signature, key, energy, danceability, and many others. The most interesting metric is “hotttness” (note the spelling is intentional), which is a value which represents how popular the song was on a scale from 0 to 1. We will have a machine learning algorithm ingest the various attributes of each song and use it to be trained to guess whether any given song will be popular or not. Once the ML algorithm is trained, we will be able to query the Spotify API to get the necessary artifacts from a song and feed it into the algorithm to attempt to predict whether the song will be hot or not.

This is an interesting problem for many reasons. It can easily be seen how an algorithm like this would be useful in the music industry. It could be used to not only predict whether a song will be popular or not but could also be used to give insight into what kind of characteristics cause a song to be popular or not. The types of song attributes and their hotness can also be tracked over time, showing what kind of songs were popular each year and what kinds of trends happened over time with the sound of music. Having insight as to what makes music popular would be extremely useful for music streaming companies, such as Spotify, but also to record labels who have to decide on what kind of music to invest money and resources into.

High-level solution architecture



We plan to use Amazon SageMaker to implement the Machine Learning Model using the Pytorch framework. We will then want to use an s3 bucket that will contain the data from the million song dataset to train and test our Machine Learning Model on. SageMaker will be important for us to use because it can be directly connected to an s3 bucket which is where we will store our data. However, before we can implement this SageMaker ML Model to work with our s3 bucket we need to transform this data into a form that is more easily workable.

First, to get the raw data we will use the Amazon public dataset snapshot to copy the data into an EBS volume. This can then be copied into our s3 bucket which we will manipulate from the original HDF5 format to JSON format. Once we have the data in a format we like we will reduce the size of the data to 10,000 songs to be more easily workable. After the data is set up correctly we can then connect our ML Model directly to our s3 bucket to start training and testing.

Dataset

For the project, we will use the Million Song Dataset (MSDS). The dataset contains 273 GB of song information compiled by Lab ROSA using data provided by The Echo Nest. The Echo Nest, a Spotify owned company, is a data collection organization whose focus is specifically on music. LabROSA is a Columbia University research organization whose focus is on intelligent sound recognition. LabROSA selected songs based on familiarity of the artist, song similarity and popularity. The dataset was last updated in November of 2015.

LabROSA has stored the data in an HDF5 format. HDF5 is a file storage technology which allows very large heterogeneous data sets, file compression and search optimization. For each song, there are over fifty features (data about the song), all of which are not needed. Hence there will be preprocessing for eliminating irrelevant data features, which will also reduce the total file size. Depending on the overall reduced size, the processed data will be transferred to local storage (e.g. as JSON) to eliminate redundant preprocessing computation. Please see the table below for some sample fields.

Subset of Dataset Fields:

Field Name	Type	Description
Title	String	Name of the song
Loudness	Float	How loud the song is overall (dB).
Song Hotttness	Float	An index rating between 0 and 1 (inclusive) of how popular a song was when released (e.g. 1 means very popular).
Tempo	Float	Song tempo (BPM).

More information can be found at: https://labrosa.ee.columbia.edu/millionsong/pages/field-list.

The MSDS is a static dataset, which can be accessed through a couple of different ways. The Open Science Data Cloud (OSDC) hosts the data as 26 tar files which can be downloaded locally. Alternatively, the data is available as an Amazon Public Dataset snapshot free of charge. All that is required is an Amazon Web Services (AWS) account which each member of the team already has access to. Therefore, we can retrieve the data by using an EC2 (Elastic Cloud Compute) virtual machine instance, attaching the data as AWS EBS (Elastic Block Storage) and then importing it.

During our initial development phase, we will store a small subset of the processed data (provided by LabROSA) as a CSV file (can import into HDFS too) in an S3 bucket. During our final development phase, we will process the data, then store it in EBS or potentially an S3 bucket if costs exceed our expense limit. In both of these phases, the data will be imported in order for our machine learning algorithm to train on.

Challenges

The problem is difficult to solve because there is no straight-forward algorithm for making a “popular” song. There are many different factors that go into making a song and gauging the perception that will be generated by the public. For example, one factor that is hard to determine is that as time progresses the genre of “popular” songs changes with mainstream listeners. A song that may have been popular half a century ago would likely not be popular today.

Some of these factors may be too random to determine, in fact, while we might have a plethora of data at our disposal, we are not guaranteed to get any meaningful success from our machine learning algorithm. Regarding our solution, some of the main challenges will be evaluating, optimizing and selecting the best algorithm and accompanying parameters for our use case. Deciding which algorithm we select will impact our final results. Considering there are many algorithms at our disposal, this can be a challenging task. Some of the algorithms to consider are regression, neural networks, and more.

Furthermore, our group is not widely experienced with machine learning, especially at a large scale, so learning and understanding the concepts may be difficult. However, we will look to use libraries such as TensorFlow and other technologies to make this step easier. Additionally, each of these algorithms will have hyperparameters that will affect prediction accuracy. Learning the tuning process will be a challenge.

Timeline

Right now it is too early and we do not have enough information about Machine Learning Modeling to be able to effectively split up responsibilities. Machine Learning is something none of us have any experience in so that will be the main part that we all put our focus on. The other main part is how we will transform the HDF5 data from the million song dataset into JSON data to be used by our PYTorch ML Model. All of these parts are very new to us so we still need to dig deeper into these tools to find exactly how to implement them to properly split up the responsibilities.

For a timetable, we have a few goals that we want to meet before the due date and before the two checkpoints for the final project. Before checkpoint 1, we want to have the million song dataset reduced and transformed to JSON data so it is ready to be used by our ML Model. Also, we want our ML Model to be properly implemented in Amazon SageMaker so it is ready to be trained on our reduced dataset. Next, Before checkpoint 2, we want to have successfully trained our ML Model on the smaller, reduced dataset. Also, we want to have the larger subset of the data ready for use of training the ML Model. Finally, by the due date, we want to have successfully trained our ML Model on the entirety of the Million song dataset, or close to it, to effectively predict the potential “hotttness” of a new song.

Sources

Million Song DataSet:

Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere.
The Million Song Dataset. In Proceedings of the 12th International Society
for Music Information Retrieval Conference (ISMIR 2011), 2011.

LabRosa:

Million Song Dataset, official website by Thierry Bertin-Mahieux,
available at: http://labrosa.ee.columbia.edu/millionsong/

The Echo Nest .... http://the.echonest.com/

HDF5 …………... https://support.hdfgroup.org/HDF5/whatishdf5.html

AWS ……………. https://aws.amazon.com/datasets/million-song-dataset/

Forbes Value of the Music Industry: https://www.forbes.com/sites/hughmcintyre/2017/04/25/the-global-music-industry-grew-by-6-in-2016-signalling-brighter-days-ahead/#5dc3a83163e3

Nielsen 2016 Music Year-End Report:
https://www.nielsen.com/us/en/insights/reports/2017/2016-music-us-year-end-report.html