



# Million Song Dataset

[Home](#)
[Getting the dataset](#)
[Code](#)
[Tutorial](#)
[Tasks / Demos](#)
[More data](#)
[Forum](#)
[FAQ](#)
[Contact / Cite](#)
[Blog](#)
[Home](#) » [Blogs](#) » [millionsong's blog](#)

## Deriving a genre dataset

Submitted by millionsong on Mon, 02/28/2011 - 18:34

Exchanging emails with [Dianne Cook](#), we pondered the idea of creating a simplified genre dataset from the Million Song Dataset for teaching purposes.

DISCLAIMER: I think that genre recognition was an oversimplified approximation of automatic tagging, that it was useful for the MIR community as a challenge, but that we should not focus on it any more.

That said, as a master student, I loved working on the [GZTAN genre dataset](#). It was simple enough to clearly understand the task; we could argue the label of a particular track, but they were still reasonable; and it was more complex than a trivial binary classification. Plus, for a machine learning or stat class, isn't it great to work on popular music data?

The idea is to use artist tags in the MSD that describe typical genres. From the artists tagged by these, we extract simple features from all their tracks. The Echo Nest terms are a little too complicated and diverse to be used for that purpose. To get a handful of genres, we would have to handpick a large number of tags and merge the related ones into genres classes. For instance, 'us pop', 'pop', 'indie pop', 'american pop', ... could all be merged into 'pop'.

The musicbrainz tags are more proper for such a task. We have less of them, but they were applied by humans and are usually very descriptive. They also tend to be standardized, as musicbrainz contributors care for consistency. From the top 50 most popular musicbrainz tags, we chose the following 10 ones that loosely mimic the GZTAN genres:

classic pop and rock, folk, dance and electronica, jazz and blues, soul and reggae, punk, metal, classical, pop, hip-hop

We consider all artists that have been tagged with these, but we remove artists that were also tagged with another word from the top 50 musicbrainz tags. It is a little extreme, but we want to avoid confusing artists, e.g. those that span more than one genre. Yes, it is disappointing, that is why we should work on automatic tagging instead of genre recognition.

As for features, we use the simple ones from The Echo Nest: loudness, tempo, time\_signature, key, mode, duration, average and variance of timbre vectors. Not that we provide the artist name and title for each of the songs, so students can make sense of the data.

Evidently, building such simplified dataset implies huge flaws! The main one is the unbalancedness of the data. The 'classic pop and rock' class is represented by 23,895 tracks, while the 'hip-hop' one has 434 tracks. It is bad. But real data sometimes does not behave well. Otherwise, there is still overlap, between 'classic pop and rock' and 'pop' for instance. Finally, we rely on musicbrainz tags, which could be wrong or incomplete for many artists.

That said, this data is still fun if you want to provide your students with realistic music data for a homework or project. The [python code](#) to create that dataset is provided, and here is the actual [MSD genre dataset](#).

This could also be improved. If you have suggestions, or have other such dataset in mind for your students, let us know!

-TBM

[millionsong's blog](#)
[Login to post comments](#)

### News

**April 25, 2012**

The [MSD Challenge](#) has launched!

**October 20, 2011**

We release the [Last.fm](#) dataset of tags and similarity!

**April 12, 2011**

We release the [musiXmatch](#) dataset of lyrics!

**March 15, 2011**

We release the [SecondHandSongs](#) dataset of cover songs!

**February 8, 2011**

We release the dataset! (and get Dan to [blog](#))

### Quick links

[LabROSA](#)  
[The Echo Nest](#)  
[Musicbrainz](#)  
[Infochimps](#)  
[7digital](#)  
[Last.fm](#)  
[musiXmatch](#)  
[SecondHandSongs](#)

### Main contact

[MSD mailing list](#)

### Random track

[new track](#)

### Random ISMIR paper

[new paper](#)

### Subscribe to MSD Mailing List

Email:

[Subscribe](#)


[Google Groups](#)





Search the site

(NOT THE DATA):  
  

Search

AddThis

 SHARE

    ...

User login


\* Username:

\* Password:

Log in

[Log in using OpenID](#)

[Request new password](#)



http://labrosa.ee.columbia.edu/millionsong/blog/11-2-28-deriving-genre-dataset

2/2