MAPPING GENRE TAXONOMIES WITH HIERARCHICAL CLUSTERING

Thomas Nuttall
Music Technology Group, Universitat Pompeu Fabra, Barcelona
thomasnuttall01@estudiant.upf.edu

ABSTRACT

Here we present a mapping of musical sub-genres as specified in four music databases. Sub-genre similarity is determined using co-occurrence across 247,716 tracks and hierarchical clustering analysis applied to group and visualise the similarity structure between them.

1. TOOLS AND METHODOLOGY

We use the Kikundi package [6] developed specifically for this analysis. Our pipeline is designed to be dataset independent with all parameters specified in a YAML configuration file, passed to the package on the command line.

Each run is assigned a unique id using Python's UUID library and considers only the intersection of our datasets.

We train our hierarchical model and output a dendrogram alongside a graph of our chosen evaluation metric (see below) at all values of k between 2 and 100, where k is the number of clusters. Both plots are stored under the unique run id.

All configuration parameters, the unique run id, our evaluation metric computed at k=2:100 and the git describe of the code used on a particular run are stored in a results csv every run.

See the Kikundi documentation for further information on package operation and structure.

2. DATASET

Four datasets, detailed in Table 1, supply genre and sub-genre annotations for a total of 1,458,447 tracks. Annotations are of the format "parentgenre—subgenre", we discard the parent genre for this analysis.

We concern ourselves only with the intersection of all four datasets, providing 1238 unique sub-genres across 247,716 tracks, and apply the following transformations to each:

- Hyphens, spaces, apostrophes and commas are removed so as to unify sub-genres like "Blues-Rock", "Bluesrock", "Blues Rock", "Blue's Rock" and "bluesrock".
- Sub-genres including a "/" are sorted alphabetically and the hyphen removed so as to unify those like "acid/jazz" and "jazz/acid".
- Sub-genres are forced to lower case.

Resulting in a total of 1064 sub-genres for our analysis.

3. SUB-GENRE CO-OCCURRENCE AND SIMILARITY COMPUTATION

We use sub-genre co-occurrence as a method of computing sub-genre:sub-genre similarity, hence we remove our track ids and reduce each of our 247,716 tracks to a set of sub-genre labels. A co-occurrence matrix is computed, each row a sub-genre, each column a sub-genre, each value the number of times the two have co-occurred.

Every sub-genre is now described by a distribution across all other sub-genres. We consider sub-genres that have occurred together often as being similar. This should make sense intuitively but is also evidenced in [1], [2], [3].

We compare every pairwise combination of distributions using cosine similarity. The result is a metric between 0 and 1 for each sub-genre pair quantifying how similar they are, 0 being that they are identical and 1 being that they are completely different. Table 2 shows the top 5 most similar sub-genres.

It is worth noting that a sub-genre that occurs only once will have a cosine similarity of 1 with all co-occurring sub-genres, this is undesirable and misleading since we do not have enough observations to quantify it's relationships with certainty. Therefore a threshold for inclusion is required, we discuss this further in Section 4.

4. AGGLOMERATIVE CLUSTERING

We perform an agglomerative hierarchical clustering analysis on our similarity matrix. Ward linkage is used so as to minimize the within cluster variance of our groupings. To evaluate our clustering we use the David Bouldin Index [4], equation (1).

$$DBI = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{M_{ij}}{S_i} \tag{1}$$

Where M_{ij} is the separation between the ith and jth cluster. S_i is the within cluster variance of the ith cluster and N is the number of clusters. It should be obvious that this equates to the ratio of between cluster variance and within cluster variance, averaged over all clus-

Table 1. Summary of datasets [7]

Origin	Number of Records	Unique Sub-Genres	Dataset Name
Last.fm	566,710	297	acousticbrainz-mediaeval2017-lastfm-train.tsv.bz2
Tagtraum	486,740	265	acousticbrainz-mediaeval2017-tagtraum-train.tsv.bz2
Discogs	904,944	284	acousticbrainz-mediaeval2017-discogs-train.tsv.bz2
AllMusic	1,353,213	744	acousticbrainz-mediaeval2017-allmusic-train.tsv.bz2

Table 2. Most similar sub-genres

Sub-genre 1	Sub-genre 2	Cosine Similarity
technicaldeathmetal	brutaldeathmetal	0.013477
contemporaryrnb	urbancrossover	0.025773
oldschoolsoul	classicsoul	0.026649
neoromantic	moviesoundtrack	0.028763
melodictrance	vocaltrance	0.031634

ters. Minimising this value is equivalent to optimising for tight, distant clusters (preferable to disperse, nearby clusters as would be observed when clustering random points in space).

Our pipeline outputs a graph of the David Bouldin Index at all values of k between 2 and 100. We inspect this for 27 values of our co-occurrence threshold (mentioned in Section 3.) [1] .

Figure 1 is typical of our evaluation graphs at all thresholds, it is obvious that after around k=20, the increase in performance for greater k is marginal.

Figure 2 shows the David Bouldin Index of a k=20 clustering at varying thresholds. Though it is obvious that the harsher the threshold, the better clustering, the choice is not quite as clear-cut as with our value for k, so we proceed with a threshold of 8000 since it provides a nice balance of performance and number of sub-genres in our final analysis, 304.

5. TOWARDS A MORE INFORMED GENRE TAXONOMY

It is of interest to us whether our decision for threshold and k produces a clustering more intuitive or nuanced than the parent genre classifications in our original data. We manually inspect the sub-genre groupings at k=20 and map these to their original parent genres. There are 64 parent genres in our original dataset, this is reduced to 36 for our subset of 304 sub-genres post-threshold. See section 6.2 for further discussion.

[1] It is important to note that our threshold is concerned with how many times a sub-genre co-occurs with other sub-genres, *not* with how many times it appears overall, therefore one sub-genre will be counted n times for each occurrence alongside n other sub-genres.

6. RESULTS

6.1 Similarity Structure

To visualise the similarity structure in our sub-genres we use a dendrogram. With 304 sub-genres to inspect, it is infeasible to display the entire plot, instead we present 3 interesting snapshots and attach alongside this document the full visualisation. There seems to be no big surprises at the chosen threshold.

6.1.1 Metal Music

Figure 3 is a snapshot of our dendrogram containing all metal sub-genres. There exists no metal sub-genres outside of this section. The sub-splits match intuition, with more

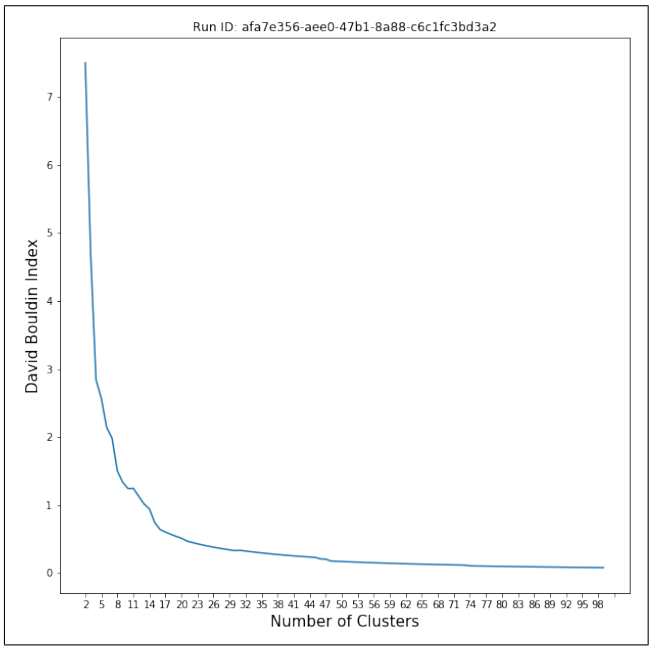

Figure 1. Evaluation of clustering at 0 threshold

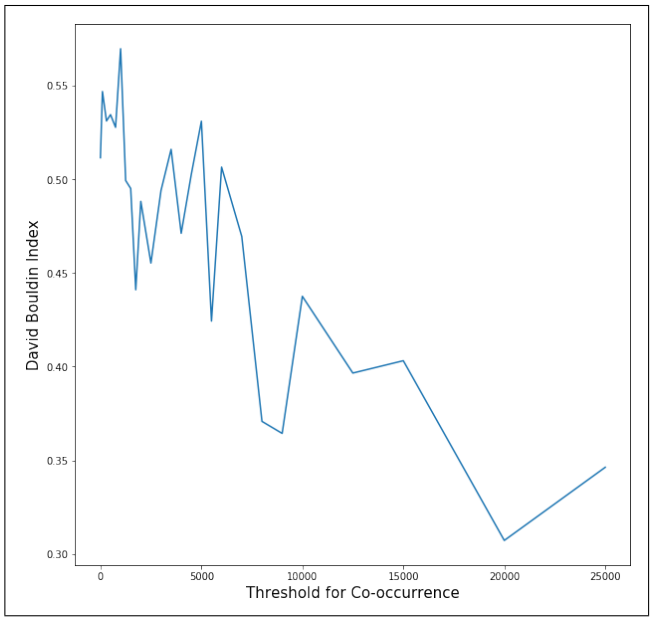

Figure 2. David Bouldin at k=20 for varying thresholds

traditional, harder metal (thrash, speed, gothic, death) identified as a separate sub-group to lighter metal variants (nu, rap, funk).

The metal sub-genres are some of the most similar in our dataset and make up one of the sub-group's last to join with others. The observations here are not only unsurprising from an experiential perspective, but corroborate other studies into genre similarity and grouping (see Table 6 and Figure 4 in [5]).

6.1.2 Hip-Hop and Rap

Figure 4 shows the Hip Hop and Rap portion of our dendrogram. Other than rap-metal, this sub-group completely captures all Rap and Hip Hop sub-genres. The sub-splits again make sense intuitively, with harder more gangster orientated sub-genres (hardcore, gangsta) separated from socially conscious, underground, jazz sub-genres (underground, alternative)

6.1.3 Jazz music

Our final sub-group is interesting because it includes both some of our most and least similar sub-genres, Figure 5 demonstrates.

Beginning at the top, the showtunes sub-genre is one least similar to others. Merriam Webster defines a show tune as simply "a song from a musical" [8]. This vague definition affords liberal application of the tag. For example, Hamilton - the musical - consists primarily of Hip Hop and Rap songs, We Will Rock You offers the Rock songs of the band, Queen and The Sound of Music contains more traditional vocal pieces typical of popular theatre in the 50s and 60s. Probably it is this broad scope that prohibits the sub-genre from pairing too closely with any one other sub-genre and hence accounts for it's similarity score.

Ultimately though, it is with the Jazz sub-section of our dendrogram that showtunes is affiliated, and perhaps unsurprisingly given Jazz's famously difficult quantification/categorization. Nevertheless, the remaining sub-genres in Figure 5 make for a much more intuitive grouping, with sub-sections correctly grouping early Jazz variants (Big Band, Swing) of the early 20th century; What western musical traditions would refer to as world-jazz (World Fusion, Latin, Brazilian, Bossa); Later more groove focused Jazz (Funk, Groove, Soul) and harder variants (Bop, Post Bop, Bebop).

Aptly though, the Jazz sub-section is not without it's surprises. It is difficult to level the grouping of traditionally slower, softer Cool Jazz with the hard and chaotic style characteristic of Bop. Also a mystery is how Traditional Pop falls alongside the early Jazz sub-genres, Swing and Big Band. Slightly oxymoronic, we must assume that Traditional Pop refers to popular music of the 1930s and 40s where artists like Jimmy Dorsey or Glenn Miller dominated the charts - both appropriately classified as Big Band/Jazz artists also. [9]

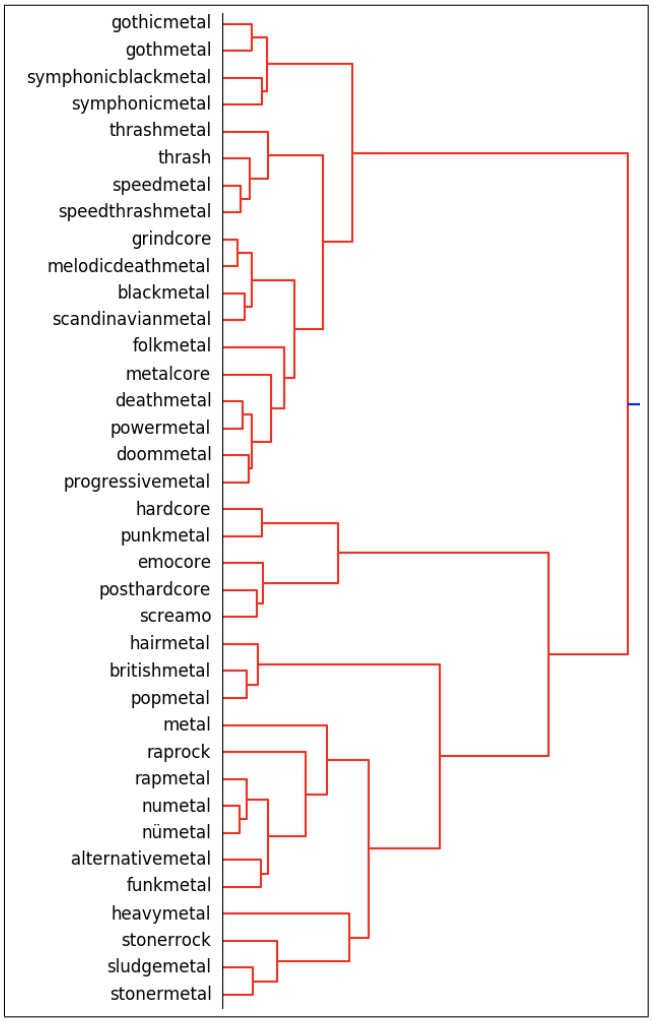

Figure 3. Dendrogram snapshot of metal sub-genres, threshold = 5000

6.2 Clustering Results

As mentioned in Section 5, we explore 99 different groupings (k=2:100) of our 304 sub-genres at a threshold of 8000.

Figure 1 shows the performance as measured by the David Bouldin Index (equation (1)) at all 99 values of k. We proceed with k=20 as our optimum grouping given the negligible increase in performance with greater k.

Table 3 shows a snapshot of our k=20 clustering. The full clustering can be found alongside this file. The clusters agree with instinct and correctly identify nuance such as the distinction between rock and roll oriented sub-genres vs art-rock/prog-rock sub-genres.

Jazz remains largely unchanged except for the breakaway of world and latin influenced jazz, this seems justified.

The new taxonomy is somewhat redundant in some instances, such as cluster 17 and 18 where all religious and festive sub-genres are correctly grouped, hence our model adds no additional information. The Electronic parent genre however is split across multiple new clusters, this is good, electronic is a very broad term and not a particularly useful genre classification. Cluster 12 is another example of sub-genres that are instinctively similar yet not at all well defined by their parent genre.

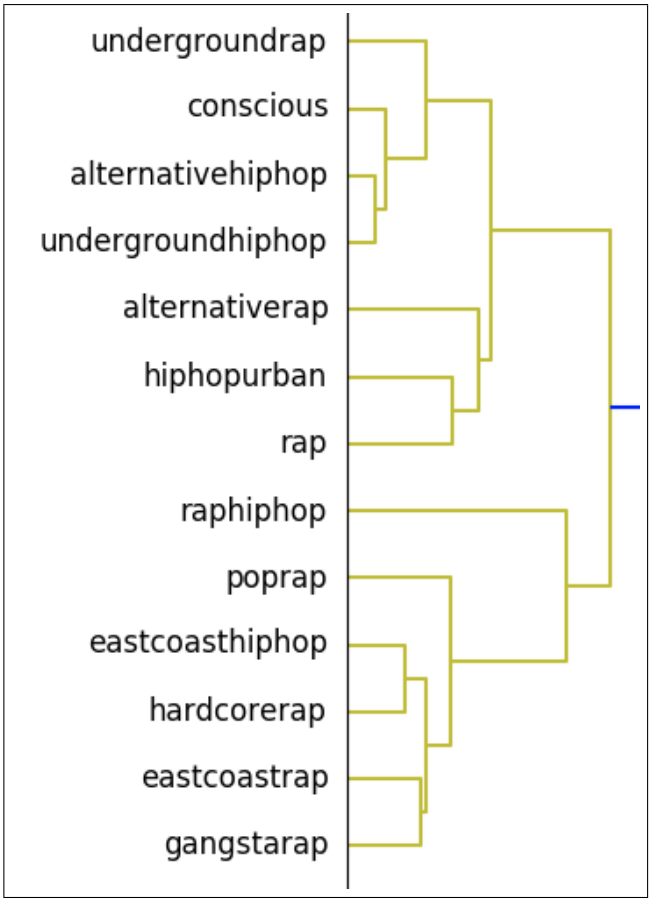

Figure 4. Dendrogram snapshot of Hip-Hop and Rap sub-genres, threshold = 5000

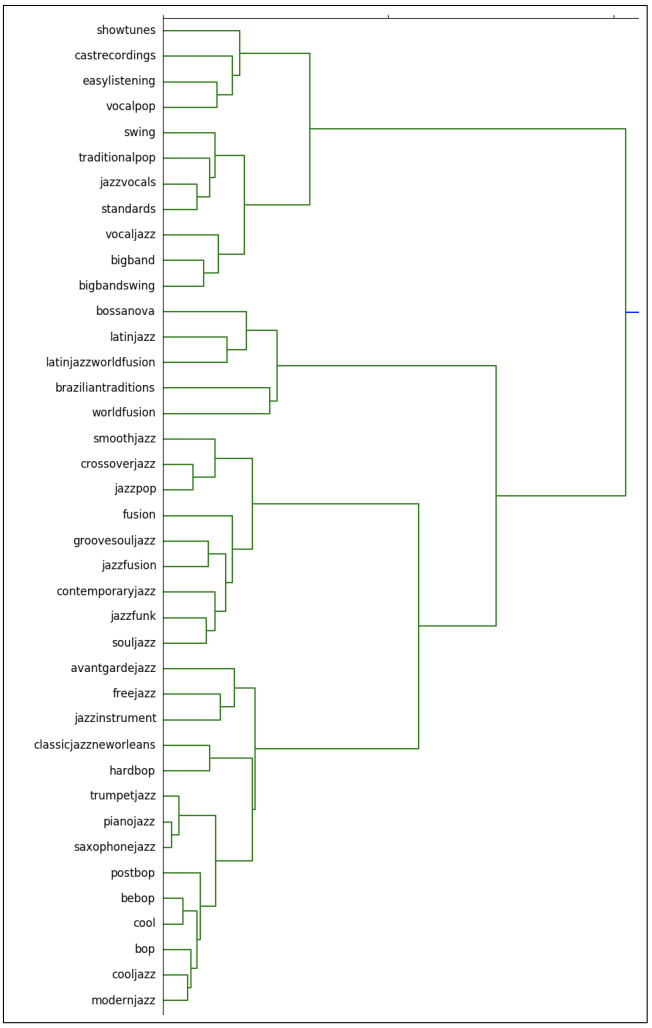

Figure 5. Dendrogram snapshot of Jazz sub-genres, threshold = 5000

7. CONCLUSION

We have demonstrated how the analysis of sub-genre tags can provide a more nuanced and informative genre classification whilst maintaining parent classifications where necessary. It is important to note that our final solution contains less groups (20) than the original data (36). In some instances the parent genre remains largely unchanged, such as religious songs, festive songs, jazz and hip hop. But in other more broad genres like rock n roll, country or electronic, hidden structure is uncovered.

In future it would be nice to match the results of this study to similar conclusions drawn by other methods. How these results can be applied in the field of Music Information Retrieval such as music recommendation or discovery would also be worth investigating.

8. REFERENCES

[1] Topic detection by clustering keywords. In *IEEE Computer Society*, pages 54–58, 2008.

[2] A. Hotho G. Stumme C. Cattuto, D. Benz. Semantic grounding of tag relatedness in social bookmarking

systems. In *International Semantic Web Conference*, pages 615–631, 2008.

[3] R. Brussee C. Wartena. Instance-based mapping between thesauri and folksonomies. In *International Semantic Web Conference*, pages 356–370, 2008.

[4] D.L. Davies and D.W. Bouldin. A cluster separation measure. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 224–227, 1979.

[5] Paul Lamere. Social tagging and music information retrieval. In *Journal of New Music Research*, pages 101–114, 2008.

[6] T. Nuttall. *Kikundi*. Gitlab, `https://gitlab.com/tnutz/kikundi/tree/develop`, fcfbd01bf74380d954dde157eef290d1efe7daca. 2019.

[7] A. Porter. https://www.dtic.upf.edu/˜aporter/amplab/. 2019.

[8] Merriam Webster. `https://www.merriam-webster.com/dictionary/show%20tune`. 2019.

[9] Wikipedia. `https://en.wikipedia.org/wiki/List_of_Billboard_number-one_singles_of_the_1940s`. 2019.

Table 3. Clustering snapshot with parent genre information. k=20, threshold=8000

Name	# Sub-genres	Example Sub-genres	Parent Genre Breakdown
0	11	minimal, worldbeat, ambienttechno	9% avant-garde / 73% electronic / 9% new age / 9% international
1	25	trance, electronicamainstream, nujazz	4% trance / 76% electronic / 4% jazz / 4% reggae / 12% dance
2	28	stonerrock, metalcore, heavymetal	46% pop/rock / 14% rock/pop / 11% rock / 29% metal
3	15	punkrock, punkrevival, emo	67% pop/rock / 13% rock / 13% reggae / 7% rock/pop
4	5	neoprog, symphonicrock, artrock	60% pop/rock / 40% rock/pop
5	23	jazzpop, jazzfusion, classicjazzneworleans	100% jazz
6	11	easylistening, showtunes, traditionalpop	27% vocal / 9% easy listening / 45% jazz / 18% stage & screen
7	9	countrypop, country, traditionalcountry	89% country / 11% folk, world, & country
8	5	braziliantraditions, worldfusion, bossanova	60% jazz / 40% international
9	6	gospel, christianrock, ccmcontemporarygospel	17% christian / 83% religious
10	3	holiday, christmas, holidays	67% holiday / 33% christmas/holiday
11	10	soundtracks, filmmusic, originalscore	20% soundtrack / 30% classical / 50% stage & screen
12	21	alternativefolk, americana, contemporarysingersongwriter	5% folk, world, & country / 10% country / 5% international / 29% folk / 38% pop/rock / 14% rock
13	10	modernelectricblues, regionalblues, acousticblues	10% pop/rock / 90% blues
14	12	europop, latinpop, spanish	33% pop / 25% pop/rock / 8% international / 8% r&b / 25% latin
15	16	popsoul, funk, quietstorm	12% funk / soul / 6% hip hop / 12% soul / 44% r&b / 25% rnb
16	36	rock&roll, bluesrock, boogierock	64% pop/rock / 19% rock / 3% r&b / 3% jazz / 11% rock/pop
17	30	noisepop, shoegaze, indiepop	3% pop / 53% pop/rock / 7% electronic / 30% rock / 7% rock/pop
18	15	industrial, electropop, electro	27% pop/rock / 33% electronic / 7% rock / 13% gothic / 20% industrial
19	13	conscious, eastcoastrap, alternativerap	8% hip hop / 38% rap / 54% hiphop