Music Classification Assignment: Task 1 Ninad Puranik 27/02/2019

Description of task : The task consisted first the creation of a track-subgenre occurence matrix from the predominantly occuring subgenres in the intersection set of 4 large datasets containing the musicbrainz id and subgenre annotations. This was to be followed by doing heirarchial clustering using the pairwise cosine distance between the colums of occurence matrix. This would help us understand which subgenres are similar to each other. The clustering was to be visualied using a dendrogram.

Methodology:

1) The methodology was to first read all the mbids in each of the four datasets into four sets. The simple set intersection operation done sequentially for the four sets of mbids will give the set of mbids present in all four datasets.
2) For each dataset, read the lines corresponding to the mbids in the intersection set just obtained. Identify a subgenre by finding the 3-hyphen, '---' , sequence. Populate a dictionary (called my_dataset in the python implementation) containing a mapping of mbid : list of subgenres. The subgenres are stored in 'source---genre---subgenre' format.
3) During step 2 also populate a set of all subgenres present in the 4 datasets as well as keep a count of number of occurences of all subgenres.
3) Remove the subgenres occuring less than a threshold number of times, here threshold chosen is 500 times. This is done to remove insignificant subgenres to speed-up the further process.
4) Finally we get 247716 mbids and 711 subgenres from which an occurence matrix is to be constructed.
5) Iterating over the my_dataset dictionary populate the occurence matrix marking '1' where a subgenre exists for a given mbid.
6) Simultaneously a csv file 'my_dataset.csv' is is written line by line.
7) From the occurence matrix , the cosine distance is calculated using the pdist module from scipy librray.
8) The cosine distance is used to do a heirarchial clustering and creating a dendrogram for visual representation. The 'col_names' which is the list of the 711 subgenres is used to label the dendogram. The 'svg' file format is used for the dendrogram to get a theoretically infinite resolution.
9) Some important patterns are observed from the dendrogram which can be seen in the included in 'results_images' folder.

Results and Discussion:

1) While the different datasets follow different taxonomies, it can be seen from the dendrogram that similar genre--subgenres are grouped together. Hence, broadly speaking, the results obtained here seem satisfactory.
2) For example in the image file 'cluster_of_electronic_industrial', we can see the the tags electronic-industrial from lastfm and discogs, to be clustered with tagtraum-industrial-* subgenres and allmusic-pop/rock-industrial.
3) From the image 'cluster_of_hiphop-rap', we can see that all subgenres with hiphop and rap as the main genre are clustered together. Many people consider hiphop and rap to be the same genre.
4) In the image 'cluster_of_metal' , we can see that all subgenres containing the phrase 'metal' irrespective of dataset and taxonomical structure are clusterred together.
5) A system of aliasing can be devised so that all genre names from different taxonomies that fall in the same higher level cluster in this heirarchial clustering are considered the same genre. Similarly all subgenres which are clustered at the lower level in the heirarchial clustering can be aliased as the same subgenre. This system can help us to understand that 2 different genre-subgenres from different taxonomical structure are the same.

Includes:
1) Jupyter Notebook with code: *Task_1_submission_Ninad.ipynb*
2) The dataset of mbids present in all four original datasets: *my_dataset.csv*
3) Dendrogram result: *dendro.svg*
4) Folder containing screenshots of specific parts in the dendrogram: *results_images*
5) Code execution state in jupyter notebook: *code_execution_snapshot.html*
6) The .tsv files for all the original datasets should be present in same folder as the above for the jupyter notebook to execute properly.