Question 1:

I used violin, guitar and trumpet same as the sounds I used for the previous assignment 9 as I can clearly distinguished the difference between these 3 sounds audibly. Further for I added bassoon, clarinet, cello, snare drum, flute which I also could distinguish the difference of these sounds by listening to them.

Finally, i added naobo and xiaoluo instead of mridangam and daluo. As I can hear a lot of similarites between snare drum and mridangam as maybe their belong to the drum kind of instruments. And I did note select daluo as I heard more similarities between the sounds of daluo and naobo and daluo and xiaoluo as compared naobo and xiaoluo.

I used 0-5 secs for the duration of all the sounds as the active regions of most singe note sounds only takes a period of 1-3 seconds to complete and to obtain results with the minimum silence duration. Most of the suitable results can be obtained by using this setting.

By extending the maximum duration to 6-10 secs, the search results may include sounds that have longer period of silence at the start and the end which are not very useful features for classification of the sounds. And setting the duration to 6-10 secs might produce search results consisting of multiple notes in a single sound.

By setting the tag to 'single-note' for violin/trumpet/clarinet/cello/flute, '1-shot for guitar/snare drum' or 'non-vibrato' for bassoon is sufficient to get most results that matches a single-note sound. I did not set any tags for naobo and xiaoluo as there are not many search results produced regarding these 2 instruments and most of the single note results can be obtained by using their instrument names and using a duration of 0-5 secs without using any tags.

```
queryText='violin',
duration=(0,5),
tag='single-note'

queryText='guitar',
duration=(0,5),
tag='one-shot'

queryText='trumpet',
duration=(0,5),
tag='single-note'

queryText='bassoon',
duration=(0,5),
tag='non-vibrato'

queryText='clarinet',
duration=(0,5),
tag='single-note'

queryText='cello',
duration=(0,5),
tag='single-note'

queryText='snaredrum',
duration=(0,5),
tag='1-shot'

queryText='flute',
duration=(0,5),
tag='single-note'

queryText='naobo',
duration=(0,5),
tag=''

queryText='xiaoluo',
duration=(0,5),
tag=''
```

Question 2:

I used the all the descriptors (0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16) and I achieve a 78.2% accuracy average over 10 runs which is the best results which I got with a combinations of descriptors.

The results after the runs shows that for this case instead of using a few chosen descriptors that best separates the sounds, it is better to use as many descriptors as possible to obtained the best accuracy.

Accuracy using single descriptors:

0: 46.50%

1: 45.00%

2: 36.50%

3: 46.00%

4: 39.50%

5: 42.50%

6. 43.00%

7. 42.50%

8. 44.50%

9. 45.00%

10. 44.00%

11. 41.50%

12. 56.00%

13. 49.00%

14. 41.50%

15. 33.50%

16. 44.00%

Using descriptors (12, 13) i got an accuracy of 68%.

Using descriptors (0, 1, 3, 9, 12, 13) which got 45% or more accuracy for their single feature classification accuracy i got an accuracy of 75%.

Using descriptors (2, 3, 4, 5, 7, 8, 9, 12, 15, 16) which got less than 45% accuracy for their single feature classification accuracy i got accuracy of 69.50%.

Using descriptors (1, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14) i got accuracy of 77.50%.

Using descriptors (3, 5, 10, 13) i got accuracy of 72.00%.

Using descriptors (3, 13) i got accuracy of 72.00%.

Using descriptors (12, 13) i got accuracy of 68.00%.

Question 3:

Part 1. Using better and more features

By using Extractor function in Essentia, I extracted lowlevel, rhythm, sfx and tonal types of descriptors. By trying out the single feature accuracy of the many descriptors, I find that the pitch_instantaneous_confidence achieved the highest single feature accuracy of 53% among the descriptors tested.

Descriptors used:

0: 'lowLevel.mfcc.mean.0'
1: 'lowLevel.mfcc.mean.1'
2: 'lowLevel.mfcc.mean.2'
3: 'lowLevel.mfcc.mean.3'
4: 'lowLevel.mfcc.mean.4',

5: 'lowLevel.mfcc.mean.5'
6: 'lowLevel.spectral_strongpeak.mean.0'
7: 'lowLevel.pitch_instantaneous_confidence.mean.0'
8: 'lowLevel.spectral_rms.mean.0'
9: 'lowLevel.spectral_crest.mean.0'
10: 'lowLevel.spectral_spread.mean.0'
11: 'lowLevel.barkbands.mean.0'
12: 'lowLevel.barkbands.mean.1'
13: 'lowLevel.barkbands.mean.2'
14: 'lowLevel.barkbands.mean.3'
15: 'lowLevel.barkbands.mean.4'
16: 'lowLevel.barkbands.mean.5'

By combining descriptors 0,1,2,3,4,5,6,7,11 I achieved an accuracy of 84.6%

Part2. Computing the descriptors stripping the silences and noise at the beginning/end

By using Energy function in Essentia, I extracted the overall single energy of all the frames of the sound. After plotting energy vs time graph of different instruments I find out that the maximum energy of the silence frames occur at 0.05 – 0.07. I stripped off the frames which energy level is below a certain threshold from the overall extraction.

By setting several values of the threshold and using the mfcc descriptors I find that setting the threshold values to 0.05 and 0.07 accuracy does not change. But setting the threshold to 0.1 and above reduces the accuracy. Probably the threshold it set too high that some of frames that contained useful active information are also stripped off.

Descriptors used:

0: 'lowLevel.mfcc.mean.0'
1: 'lowLevel.mfcc.mean.1'
2: 'lowLevel.mfcc.mean.2'
3: 'lowLevel.mfcc.mean.3'
4: 'lowLevel.mfcc.mean.4',

5: 'lowLevel.mfcc.mean.5'
6: 'lowLevel.spectral_centroid.mean'
7: 'lowLevel.dissonance.mean'
8: 'lowLevel.hfc.mean'
9: 'sfx.logattacktime.mean'
10: 'sfx.inharmonicity.mean'
11: 'lowLevel.spectral_contrast.mean.0'
12: 'lowLevel.spectral_contrast.mean.1'
13: 'lowLevel.spectral_contrast.mean.2'
14: 'lowLevel.spectral_contrast.mean.3'
15: 'lowLevel.spectral_contrast.mean.4'
16: 'lowLevel.spectral_contrast.mean.5'

By combining descriptors 0,1,2,3,4,5,6,7,11,13,14 I achieved an accuracy of 79.6%