Latent Dirichlet Allocation

Yuxuan Zhang

We will first run the script to display the result

In [10]:
```
!python ./lda_test.py
```

```
['bass', 'bass', 'bass', 'tuba', 'tuba', 'tuba', 'deep', 'bass', 'horn', 'horn',
'tuba', 'horn', 'tuba', 'deep', 'deep', 'horn', 'deep', 'horn', 'deep', 'bass',
'horn', 'deep', 'horn', 'deep', 'deep', 'tuba', 'horn', 'horn', 'tuba', 'bass',
'horn', 'horn', 'horn', 'horn', 'deep', 'horn', 'deep', 'tuba', 'bass', 'horn',
'deep', 'horn', 'bass', 'deep', 'tuba', 'tuba', 'horn', 'tuba', 'bass', 'horn',
'horn', 'horn', 'tuba', 'bass']
[0.33333334 0.33333334 0.33333334]
[(0, '0.361*"pike" + 0.272*"bass" + 0.197*"deep" + 0.071*"horn" + 0.064*"catapul
t" + 0.035*"tuba"'), (1, '0.255*"bass" + 0.221*"horn" + 0.183*"deep" + 0.129*"tub
a" + 0.127*"pike" + 0.085*"catapult"'), (2, '0.376*"pike" + 0.263*"catapult" + 0.
209*"horn" + 0.077*"deep" + 0.059*"bass" + 0.016*"tuba"')]
```

For inferred topic 0, we have

- bass - 0.272
- pike - 0.361
- deep - 0.197
- tuba - 0.035
- horn - 0.071
- catapult - 0.064

We see that this topic has dominant words "pike", "bass", and "deep", and the two words with low probability are "horn" and "tuba". Hence, this should be mapped to the true topic 0.

For inferred topic 1, we have

- bass - 0.255
- pike - 0.127
- deep - 0.183
- tuba - 0.129
- horn - 0.221
- catapult - 0.085

We see that this topic has dominant words "bass", "horn", "deep", and "tuba", so this topic should likely be mapped to the true topic 2. However, we also see a bit of noise coming from "pike", which appears in this inferred topic with a lower probability.

For inferred topic 2, we have

- bass - 0.059
- pike - 0.376

- deep - 0.077
- tuba - 0.016
- horn - 0.209
- catapult - 0.263

We see that this topic has dominant words "pike", "catapult", and "horn". The word "deep" also has a probability of 0.077, which is close to 0.1. Therefore, this topic should be mapped to the true topic 1.

In conclusion, the mapping between our inferred result and true topics is

Inferred topic 0 → True topic 0

Inferred topic 1 → True topic 2

Inferred topic 2 → True topic 1

To reduce the noise, we can increase the number of iterations of the 'ldaModel'. From the documentation of gensim, we see that this parameter is set to a default value of 50, which may be the reason for the noise in our result.