14 Appendix: Frequent sequence mining

We want to check, among our set of frequent items extracted with **sequential pattern mining**, which of them appear as one contiguous block and with what support; so, we are doing what we're calling **frequent sequence mining** starting from candidate sequences which are the top patterns extracted with sequential mining. These, the first 5 patterns that showed the highest support in the pattern mining section, are:

1. [9, 9, 9, 9, 9, 9, 9, 9, 9, 9] occurs 406 times (as a sequence),
2. [10, 9, 10, 10, 10, 9, 9, 9, 9, 9] occurs 125 times,
3. [9, 9, 9, 9, 9, 9, 10, 9, 10, 10] occurs 135 times,
4. [9, 10, 10, 10, 9, 9, 9, 9, 9, 9] occurs 108 times,
5. [9, 9, 10, 9, 9, 9, 9, 9, 9, 9] occurs 341 times.

There are some interesting results: we can see that the support obtained in the sequential pattern mining (chap. 10, basically the same support for all 5 patterns) doesn't necessarily follow the same distribution of occurrence of that pattern in this sequence mining (shown in the list just above, with highly variable supports). The first sequence occurs 406 times in our set of time series and this represents the 5% of the total number of time series. We can say that this sequence of length 10 appears frequently.
In order to try and give some characterization to this task we observed the *track genre* distributions for the time series that contain one of the 5 sequences in our list. In particular we noticed that there might be some correlations between the 1th and 5th sequences. These two sequences showed interesting features as they both have a similar **distribution of genres** and similar **support**. We plotted their genre distributions in the following histograms:

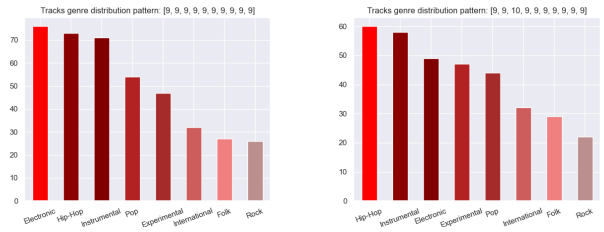

Figure 76: Genre distribution for Time Series that contain 1th and 5th sequences
As specified in chapter 10 we couldn't increase the pattern length, so the fact that 1st and 5th show similar distributions suggests to us that they could be both characterizing a particular distribution and therefore be part of the same, larger, pattern (longer than 10, therefore not discovered with sequential pattern mining).
We also noticed from the genre distribution that if the symbol "10" appear in the sequence extracted, the number of times series of genre Electronic decreases. That's shown in the following histograms of the 1th and 2th sequences:

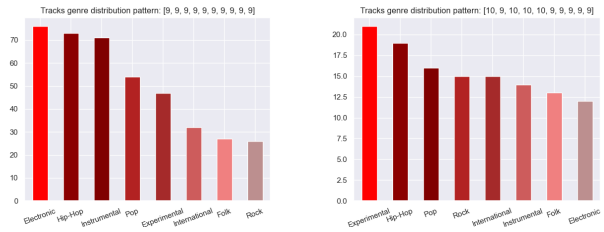

Figure 77: Genre distribution for Time Series that contain 1th and 2th sequences
This can be observed also from the previous comparison. From these two observations it looks like the frequency of genre *electronic* and the frequency of symbol 10 are inversely proportional.

✍

Aknowledgments:
Fabio: I would like to thank my two teammates, it was an absolute joy and super fun to work with them.
Marianna: Thank you for the time spent in reading our work. We wish you enjoy it.
Saverio: It was a fun challenge, I hope you liked our hard work!