Analysis of data generated by the corpus of chords project. Complete with spelling and grammar mistakes, grossly unprofessional anecdotes, and the manic effect of students on a deadline trying not to go too deep into a really interesting rabbit hole.

I. The generator.

The degree to which the generator “sounds like music” is a plausible if inexact metric for the quality of the word embeddings. Nonfunctional word embeddings would result in generated sequences indistinguishable from randomly picking chords from the dataset. The strongest piece of evidence for this not happening is the relative lack of large jumps (pitchwise). Even Though it is clearly capable of it, and it does make some, in general it sticks to a range, and when it does jump it sticks with that range. This is not behavior we would expect from a nonfunctional set of embeddings. This cannot simply be a result of most of the dataset existing in one range, because when adding seven semitone transpositions of the dataset this behavior remains (the jumps become more common, but it still sticks to a range for a while before jumping, whereas the expected behavior would be a frequency distribution of ranges matching the dataset).

In terms of if one chord to another is generally “making sense” this is more difficult to determine empirically without a somewhat exhaustive amount of analysis by hand, although both myself and Janice (being classically trained musicians) believe we hear a conspicuous number of coherent progressions, granted we also hear a plentiful, and in my opinion delightful, amount of dissonant nonsense. Dissonance should not necessarily be seen as a sign of a bad embedding, the dataset in question makes frequent, although judicious, use of dissonance. The embedding won't necessarily represent the frequency of certain progressions, just their validity, thus a more even ratio of dissonance to consonance is to be expected. Furthermore adding transpositions to the dataset *seems* to be increasing the amount of dissonance, this makes sense as transpositions will give the embedding model a greater amount of contexts for off key notes, which are often perceived as dissonances.

Personally, with the output so far generated, I find the embeddings based generator to be a useful artistic tool, which is the ultimate goal for the generator. The clusterings however, are another matter.

II. The clusterings.

Looking at the data generated by the clusterings, the first thing which jumps out to me is that we have several clusters in which the tonic triad has a disproportionately low representation. This is fantastic. This is exactly what we hoped for. This conforms to the classical concept of “tonic harmony”, “dominant harmony”, and “intermediate harmony”. The clustering is recognising that there are chords which are distinctly not tonic harmony. Note for example that across most of the clustering models, flat notes appear in the same clusters, nothing in the key of C. In general, off key chords are showing up in the same clusters. The tonic itself shows up in some frequency in nearly every cluster. This is simply indicative of the fact that the tonic will be the most common note in general. Furthermore we also see lots of clusters that primarily contain notes in the key, and a minority of clustering containing even amounts of both. Looking at the top two chords in each cluster on the 48-cluster produced by sklearn, the most common intervals are perfect fifths, perfect fourths, and thirds of the appropriate quality. Very few seconds and sevenths. One top 3 for a cluster is Eb major, Bb major, F dominant 7th. This is frankly incredible. We have the appropriate tonic, dominant 7th and subdominant chords for the key of Bb. Similar stories are told across a staggering majority of the clusters. It makes me very sad that I ended up not having the time to incorporate the clustering into the generator. I am however, beyond pleased with this result and supremely confident that meaningful categories are being drawn.