Make sure to choose pieces with only piano music! Some might be stuff like **quartet, concerto, or symphony**: these will contain other instruments (which might mess up data?) Make sure to choose piano-only.

Maybe split our piece into 30-second chunks as data? That way each data point is consistent, and we also get more data from each composer.

From Each Time Period: Let’s try to find ~5 pieces from each composer from this list:

Baroque:

Johann Sebastian Bach

Domenico Scarlatti

Handel

Telemann

(other, wont use: Corelli, Vivaldi)

Classical:

Clementi

Haydn

Mozart (pick 10)

Transition (not used): Beethoven, Schubert

Romantic (includes late romantic?):

Chopin

Liszt

Robert Schumann

Felix Mendelssohn

(others, won’t use tho: Rachmaninoff (late romantic), Brahms, Dvořák, Grieg)

(We will be skipping impressionist)

Modern:

Prokofiev

Debussy

Ginastera

(others, won’t use tho: Gershwin, Copland, Shostakovich, Barber, Kabalevsky, Medtner, Scriabin, Stravinsky)

TODO list

1. Open the files and read into a dictionary {“romantic”, music21.midiObjects, “classical”, music21.midiObjects}
2. Process each midi object

Word = notes being played at the same time (for example, if a chord is playing, our word would be the all the notes *playing* at that instant)

Sentence = 10 seconds of the song (hyperparameter)

Things to remember/decisions we made

1. Not doing it by seconds
   1. Doing it by measures instead
   2. Its a consistent rate
   3. So, if a song has more *words* then it has more measures.
   4. More cohesive/scales across better
   5. Takes care of *sample rate* issue
2. Not keeping track of individual notes
   1. Not keeping track of single notes
      1. Would not work for naive Bayes (There is no order information, all songs would seem the same)
      2. Would lose valuable information
         1. Duration of notes
         2. Which notes are played together/very unreliable order
         3. Intensity
         4. Tempo/rhythm
      3. If we tried to combine notes together, it wouldn’t work
         1. If we did it by measures, there are 88^4 combinations
         2. Difficult to keep track of which notes are still playing
         3. Very sparse, would not generalize well to test set
      4. We will use feature extraction
         1. Write more here
      5. We will split our songs up by measures to keep some order information intact