APSTA-2017: Process Journal

William Spagnola

3/14/2019

3/14/19 *Happy π Day*

What I've done this week

1) I learned the basic functions in the RSelenium package, and I created a webcrawler to scrape chord information from a dynamic webpage (www.hooktheory.com)
2) I learned how to scroll down to load svg images located toward the bottom of page. Alan helped me to understand the basic concept of webcrawlers during our group discussion on Monday. Previously, I could not understand why I was only able to scrape the first two svg images from each page. The reason was that only the first two images load unless you scroll down.
3) I learned how to scrape songs and links from hooktheory.com, so I can determine which songs are available for each artist. I can now create a vector of links for the exact locations of the pages containing the chords of each song.

Example Dataframe

```
##           artist                        song   song_parts
## 1      Tom Petty                Free Fallin'       Chorus
## 2      Tom Petty     Into The Great Wide Open        Verse
## 3      Tom Petty                 Wildflowers       Chorus
## 4      Tom Petty                  Zombie Zoo       Chorus
## 5   Led Zeppelin                   Tangerine        Verse
## 6   Led Zeppelin                   Tangerine       Chorus
## 7   Led Zeppelin          Stairway to Heaven        Intro
## 8   Led Zeppelin      Communication Breakdown        Verse
## 9   Led Zeppelin      Communication Breakdown       Chorus
## 10  Led Zeppelin                   Thank You        Intro
## 11  Led Zeppelin               Ten Years Gone Instrumental
## 12 Elvis Presley   Can't Help Falling In Love       Chorus
## 13 Elvis Presley I Can't Help Falling In Love       Chorus
## 14 Elvis Presley               Love Me Tender        Verse
## 15 Elvis Presley               Love Me Tender       Chorus
## 16 Elvis Presley                    Hound Dog        Verse
## 17 Elvis Presley               Don't Be Cruel        Verse
## 18 Elvis Presley             Suspicious Minds        Verse
## 19 Elvis Presley   You Were Always On My Mind        Intro
## 20 Elvis Presley   You Were Always On My Mind        Verse
## 21 Elvis Presley               If I Can Dream       Chorus
## 22 Elvis Presley               That's Alright        Verse
## 23 Elvis Presley               That's Alright   Pre-Chorus
## 24 Elvis Presley               That's Alright       Chorus
## 25 Elvis Presley               That's Alright       Bridge
## 26   Ray Charles            Hit The Road Jack        Intro
## 27   Ray Charles                      Georgia        Intro
## 28   Ray Charles                      Georgia        Verse
## 29   Ray Charles                      Georgia       Bridge
## 30   The Beatles                     Hey Jude        Verse
```

```
## 31   The Beatles                     Hey Jude       Chorus
## 32   The Beatles                     Hey Jude        Outro
## 33   The Beatles           A Day in the Life        Verse
## 34   The Beatles           A Day in the Life       Bridge
## 35   The Beatles                   Let It Be        Verse
## 36   The Beatles                   Let It Be       Chorus
## 37   The Beatles                   Blackbird        Verse
## 38   The Beatles                   Blackbird       Chorus
##                                                                                       chords
## 1                                                  F-Bb(add9)-F-Csus4-F-Bb(add9)-F-Csus4
## 2                                                 G-C-Dsus4-G-em-D-am-G-C-Dsus4-G-F-em-A
## 3                                                          f#m-A-E-f#m-A-E-f#m-A-E-f#m-A
## 4                                                                  em7-Asus2-A-em7-A-G6
## 5  am-asus4-am-am(add9)-G-D-am-asus4-am-am(add9)-G-D-C-G-am-G-D-Dsus4-D-D(add9)-D-C-D-G
## 6                                                                      G-D-C-D-G-D-C-D
## 7       am-E-C-D-Fmaj7-G-am-E-C-D-Fmaj7-G-am-C-D-Fmaj7-am-C-G-D-C-D-Fmaj7-am-C-G-Fmaj7
## 8                  E-D-A-D-E-D-A-D-E-D-A-D-E-D-A-D-E-D-A-D-E-D-A-D-E-D-A-D-E-D-A-D
## 9                                                                  A7-B7-E-D-A-D-E-D-A-D
## 10                                                                       D-C-G-D-C-G-D
## 11                                                                           A-G-D-F-A
## 12                                                D-f#m-bm-G-D-A7-G-Abm-em-D-A7-D
## 13                                                 D-f#m-bm-G-D-A7-G-Abm-em-D-A-D
## 14                                                                   D-E7-A7-D-E7-A7
## 15                                                 D-F#7-bm-D7-G-gm-D-Dbm7-E7-A7-D
## 16                                                                         C-F-C-G-F-C
## 17                                                                          D-G-D-em-A-D
## 18                                             G-D-em-D-C-D-G-D-em-D-A-C-G-C-G-am-G-em
## 19                                                              C-am7-C-am7-C-am7-C-am7
## 20                          C-am-F-G-C-am-F-G-C-F-C-F-E-am-C-am-F-am7-dm7-G-dm7-G
## 21                                                                               A-D-E
## 22
## 23                                                                 G-C-D-C-G-C-D-C-D-C-G
## 24                                                                 C-Gbm-C-D-em-bm-C-D
## 25                                                                 C-Gbm-C-D-em-bm-C-D
## 26               g#m-F#-E-D#7-g#m-F#-E-D#7-g#m-F#-E-D#7-g#m-F#-E-D#7-g#m-F#-E-D#7
## 27                                                              Gbm-em-cm-c#o-G-am7-D7
## 28                                          G-B7-em-G7-C-c#o-G-E7-A7-D7-F7-E7-A7-D7
## 29                       em-am7-A7-am7-em-am7-em7-A7-em-am7-em-F#-F#7-bm7-E7-A7-D7
## 30                                                       F-C-C7-C7sus4-C7-F-Bb-F-C-F
## 31                      F7-Bb-dm7-gm7-dm-C7-F-Fmaj7-F7-Bb-dm7-gm7-dm-C7-F-F7-C7-F
## 32                                                                         F-Eb-Bb-F
## 33                                       G-em-C-am7-f#m7(b5)-Gbm-em-C-F-em-C-F-em-C
## 34          E-D-E-B(add9)-E-B(add9)-B-E-D-E-B(add9)-E-B(add9)-C-G-D-A-E-C-G-D-A-E-D-C
## 35                                                      C-G-am-Fmaj7-F6-C-G-F-C-dm-C
## 36                                                                am-G-F-C-G-F-C-dm-C
## 37             G-am7-G-C-c#o-D-d#o-em-Eb-D-c#o-C-cm-G-A7-D7sus4-G-C-G-A7-D7sus4-G
## 38                                            F-C-dm-C-Bb-C-F-C-dm-C-Bb-A-dm7-gm
```

Goals for next week

1) Adjust webcrawler scrolling to load 4, 5, and 6 part songs. Currently I have been able to load up to 3 parts of a song if a song has 3 parts.
2) Match songs scraped with website with another database in order to determine genre, year produced, and possibly key. Song keys can be determined based on the chords, but it might be easier if I can just

get this info from another source such as the Spotify API.

3) Engineer features to assess songs. I was thinking of creating indicator variables of 1-grams, bigrams and trigrams. This would include information on common chord changes in each of the songs. Then I could apply some machine learning algorithm to find clusters of songs based on the chord change features.

4) I'm not sure if I want to include slash chords yet. Slash chords are simply chords where the bass note (lowest note) is different from the root. For example, a C-major chord is constructed using C-E-B with C (the root) typically being the lowest note in the chord. If I played the same chord with E or B as the lowest note, I would write the chord either as C/E or C/B respectively. However, it still would be considered a C major chord.
5) Convert chords into scale degrees using **roman numerical analysis**. This will help compare two songs that feature similar chord progressions but are written in different keys.

Challenges

1) Hooktheory.com includes around 12,000 songs in total. This is a lot of data; however, it doesn't include all the songs in the Billboard Hot100 from the past 50 years.
2) Hooktheory.com is skewed toward modern songs. For example, it contains over 50 songs by Taylor Swift but only around 8 songs for Elvis. The sparsity of songs from earlier decades may make it difficult to compare trends across different decades.

3) Another issue is that the same song can be written differently. Furthermore, sometimes the listed artist name includes the band name as well as the singer's, and other times it only includes the lead singer's (e.g. 'Tom Petty' vs. 'Tom Petty & the Heartbreakers'). These differences can complicate the task of linking this dataset to other sources.

4) Need advice on how to scrape data without overloading server. What is an appropriate wait time betwen each request?

Idea for Analysis

1) Start by taking the Hot100 Billboard dataset (Jaejin found a csv of this) and determine the top 5 artists from each decade (1950s to 2010s) measured by number of appearances in Hot100 grouped by decade and artist. Then build a datset with hooktheory data based around these artists.