APSTA-2017: Process Journal

William Spagnola

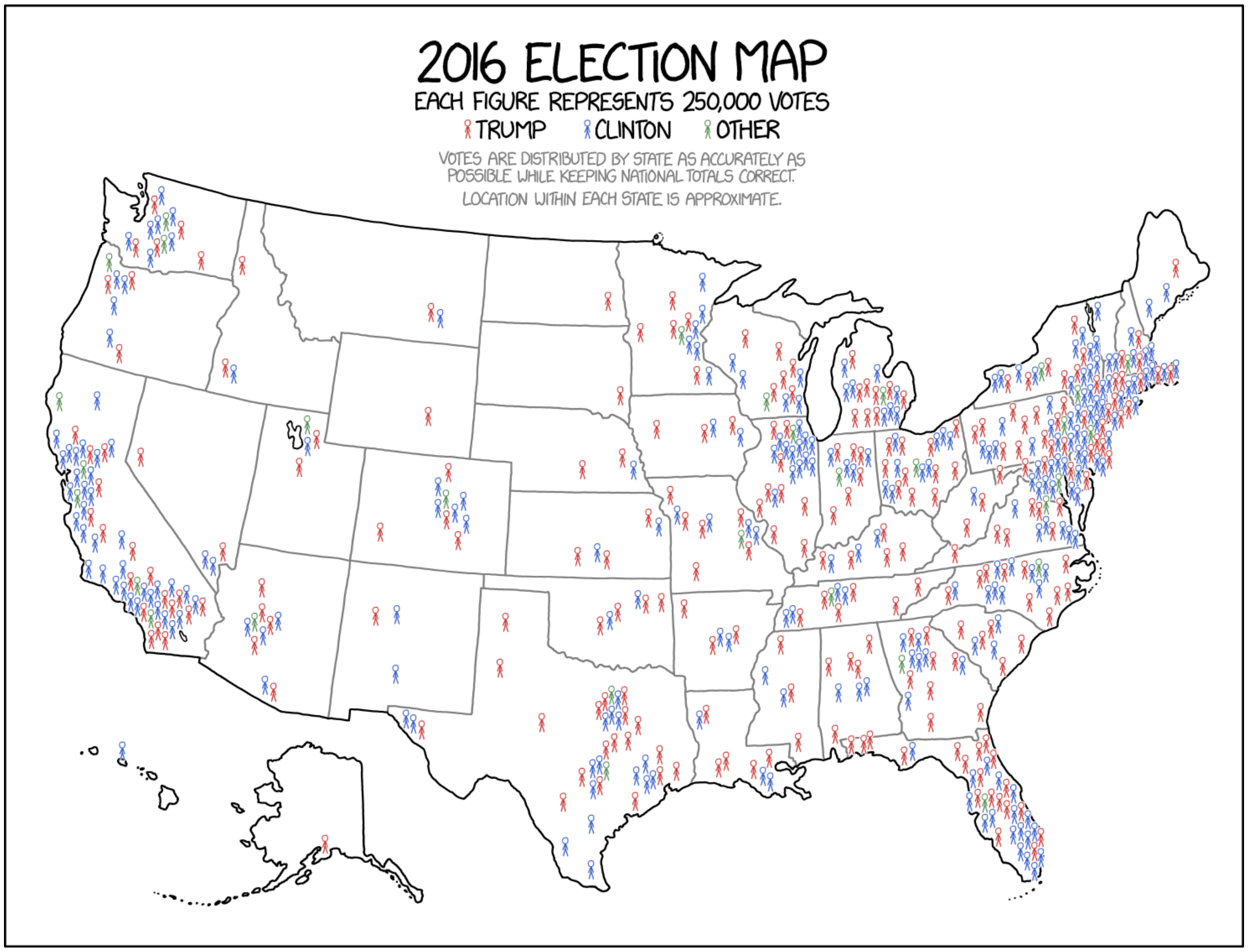
Spring 2019 Semester

4/16/19 Data Visualization Examples

Here is a link to a data visualization example **2018 House Forecast** I won't tell if I think that one is good or bad.

I also think the graph below is interesting.

```
include_graphics('img/2016_election_map.png')
```


2016 ELECTION MAP
EACH FIGURE REPRESENTS 250,000 VOTES
TRUMP
CLINTON
OTHER
VOTES ARE DISTRIBUTED BY STATE AS ACCURATELY AS POSSIBLE WHILE KEEPING NATIONAL TOTALS CORRECT.
LOCATION WITHIN EACH STATE IS APPROXIMATE.

4/15/19

1) Was able to scrape key directly from site (GOAL Accomplished)
2) Created function to convert chords to roman numerical analysis
3) Need to adjust function to account for borrowed chords and mixolydian and dorian modes

Goals

1) Edit convert_to_roman() function

- Allow mixolydian and dorian modes
- Detect Borrowed chords (from IV, V, and minor scales)

2) Find years for each song

- Merge with original Hot100 csv
- Find other dates by building a dataset from Spotify API with song titles and dates for each artist

3) Extract Features

- Create bigrams
- Create one set with additional features (e.g. 9th, maj7th, min7th, sus, add9) and one plain set
- May consider using indictator variable for presence of ‘complex’ chords

4) Start Analysis

- Start with classifier for early/late Beatles songs (This could be a more realistic topic for my main analysis)

4/9/19

What I’ve done so far

1) Got scraper to work. **Note this took forever to figure out!!!**
2) Got links of top 30 artists from each decade
3) Started compiling list of songs
4) Ordered Real Book as supplement.

- May contain more older songs (50s, 60s, 70s).
- Can scan and use OCR to extract chords time permitting.

Goal for Next Week

1) Write algorithm to guess key

- Guessing a song’s key based on the chords can be tricky.
- Will most likely start with Beatles songs and take easy to predict subset.

2) Convert Selected Beatles Songs to Scale Degree
3) Write code to convert from long to wide format so that each row represents a single song instead of a song part

Other Goals

1) Extract Features

- bigrams
- trigrams

2) May use two or three sets of features to compare

- One with scale degree plus major or minor
- One with scale degree plus major or minor plus other features (e.g. 7th, 6th, 9th, diminished)
- One with scale degree plus major or minor plus indicator variable for presence of more complicated chords

3) Extract features from all Beatles songs
4) Create visualizations based on features
5) Created indicator for early versus later Beatles. May use 1962–1966 and 1967-1970 as defininining years based on *The Red Album* and *The Blue Album*, which were compilations albums released in 1973.

6) Train classification algorithms to identify early from later Beatles.

- Plan use 80/20 train/test split randomly
- Use cross validations to select best model for each set of features using training data.
- Use out-of-sample error on test data to compare each set of features.

Another Intermediate Step

1) Find song years for each song. May be able to use Spotify API. This may seem trivial, but it can actually be pretty complicated because:

- Differences in Capitalization (easier)
- Differences in spelling (feelin' versus feeling)
- Differences in song title (Sometimes)
- Ampersands versus ands
- Same song recorded by different artists
- Same song recorded by same artist at different times
- Same song but remastered in a later year and rereleased on different album

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
## 1                         Tom Petty               Free Fallin'
## 2  Tom Petty And the Heartbreakers            I Won't Back Down
## 3  Tom Petty And The Heartbreakers            I Won't Back Down
## 4  Tom Petty And The Heartbreakers            I Won't Back Down
## 5                         Tom Petty     Into The Great Wide Open
## 6                         Tom Petty                  Zombie Zoo
## 7                         Tom Petty                     Refugee
## 8                         Tom Petty                     Refugee
## 9  Tom Petty and the Heartbreakers       Mary Jane's Last Dance
## 10 Tom Petty and the Heartbreakers       Mary Jane's Last Dance
## 11 Tom Petty and the Heartbreakers       Mary Jane's Last Dance
## 12 Tom Petty and the Heartbreakers                 Wildflowers
## 13                     Led Zeppelin                   Tangerine
## 14                     Led Zeppelin                   Tangerine
## 15                     Led Zeppelin          Stairway to Heaven
## 16                     Led Zeppelin     Communication Breakdown
## 17                     Led Zeppelin     Communication Breakdown
## 18                     Led Zeppelin                   Thank You
## 19                    Led Zeppelin             Ten Years Gone
## 20                   Elvis Presley   Can't Help Falling In Love
## 21                   Elvis Presley I Can't Help Falling In Love
## 22                   Elvis Presley               Love Me Tender
## 23                   Elvis Presley               Love Me Tender
## 24                   Elvis Presley                    Hound Dog
## 25                   Elvis Presley               Don't Be Cruel
## 26                   Elvis Presley               That's Alright
## 27                   Elvis Presley               That's Alright
## 28                   Elvis Presley             Suspicious Minds
## 29                   Elvis Presley   You Were Always On My Mind
## 30                   Elvis Presley   You Were Always On My Mind
## 31                   Elvis Presley   You Were Always On My Mind
## 32                   Elvis Presley   You Were Always On My Mind
## 33                   Elvis Presley               If I Can Dream
## 34                     Ray Charles            Hit The Road Jack
## 35                     Ray Charles                      Georgia
## 36                     Ray Charles                      Georgia
## 37                     Ray Charles                      Georgia
## 38                     The Beatles                     Hey Jude
## 39                     The Beatles                     Hey Jude
## 40                     The Beatles                     Hey Jude
## 41                     The Beatles            A Day in the Life
## 42                     The Beatles            A Day in the Life
## 43                     The Beatles                    Let It Be
## 44                     The Beatles                    Let It Be
## 45                     The Beatles                    Blackbird
## 46                     The Beatles                    Blackbird
##      song_parts
## 1        Chorus
## 2         Verse
## 3         Verse
## 4    Pre-Chorus
## 5         Verse
## 6        Chorus
## 7        Chorus
## 8        Bridge
## 9         Intro
## 10       Chorus
## 11       Bridge
## 12       Chorus
## 13        Verse
## 14       Chorus
## 15        Intro
## 16        Verse
## 17       Chorus
## 18        Intro
## 19 Instrumental
## 20       Chorus
## 21       Chorus
## 22        Verse
## 23       Chorus
## 24        Verse
## 25        Verse
## 26      Intro
## 27      Verse
## 28     Chorus
## 29      Verse
## 30 Pre-Chorus
## 31     Chorus
## 32     Bridge
## 33      Verse
## 34      Intro
## 35      Intro
## 36      Verse
## 37     Bridge
## 38      Verse
## 39     Chorus
## 40      Outro
## 41      Verse
## 42     Bridge
## 43      Verse
## 44     Chorus
## 45      Verse
## 46     Chorus
##                                                                                   chords
## 1                              F-B-b(add9)-B-b(add9)-F-Csus4-F-B-b(add9)-B-b(add9)-F-Csus4
## 2                                                              em-D-G-em-D-G-em-D-C-em-D-G
## 3                                                              em-D-G-em-D-G-em-D-C-em-D-G
## 4                                                G-C-G-C-G-D-G-C-G-D-G-C-G-D-em-D-G-em-D-G
## 5                                                   G-C-Dsus4-G-em-D-am-G-C-Dsus4-G-F-em-A
## 6                                                                     em7-Asus2-A-em7-A-G6
## 7                                                        Bb-F-C-F-Bb-F-C-Bb-F-C-F-Bb-F-C-F
## 8                                                                   Bb-dm7-dm-G-C-Bb-dm7-C
## 9                                                         am-F-am-F-am-F-A-E-bm-E-A-E-bm-D
## 10                                                         E-f#m-D-E-f#m-D-E-f#m-D-E-A-D-A
## 11                                                                           E-f#m-E-f#m-D
## 12                                                           f#m-A-E-f#m-A-E-f#m-A-E-f#m-A
## 13 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
## 14                                                                         G-D-C-D-G-D-C-D
## 15       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
## 16                     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
## 17                                                                   A7-B7-E-D-A-D-E-D-A-D
## 18                                                                           D-C-G-D-C-G-D
## 19                                                                               A-G-D-F-A
## 20                                                        D-f#m-bm-G-D-A7-G-A-bm-em-D-A7-D
## 21                                                         D-f#m-bm-G-D-A7-G-A-bm-em-D-A-D
## 22                                                                         D-E7-A7-D-E7-A7
## 23                                                          D-F#7-bm-D7-G-gm-D-bm7-E7-A7-D
## 24                                                                             C-F-C-G-F-C
## 25                                                                            D-G-D-em-A-D
## 26                                                                 C-am7-C-am7-C-am7-C-am7
## 27                               C-am-F-G-C-am-F-G-C-F-C-F-E-am-C-am-F-am7-dm7-G-dm7-G
## 28                                                                                   A-D-E
## 29
## 30                                                                   G-C-D-C-G-C-D-C-D-C-G
## 31                                                                    C-G-bm-C-D-em-bm-C-D
## 32                                                                    C-G-bm-C-D-em-bm-C-D
## 33                                         G-D-em-D-C-D-G-D-em-D-A-C-G-C-G-am-G-em
## 34                        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
## 35                                                              G-bm-em-cm-c#o-G-am7-D7
## 36                                         G-B7-em-G7-C-c#o-G-E7-A7-D7-F7-E7-A7-D7
## 37                            em-am7-A7-am7-em-am7-em7-A7-em-am7-em-F#-F#7-bm7-E7-A7-D7
## 38                                                         F-C-C7-C7sus4-C7-F-Bb-F-C-F
## 39                            F7-Bb-dm7-gm7-dm-C7-F-Fmaj7-F7-Bb-dm7-gm7-dm-C7-F-F7-C7-F
## 40                                                                            F-Eb-Bb-F
## 41                                      G-em-C-am7-f#m7(b5)-G-bm-em-C-F-em-C-F-em-C
## 42                  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
## 43                                                        C-G-am-Fmaj7-F6-C-G-F-C-dm-C
## 44                                                                   am-G-F-C-G-F-C-dm-C
## 45                        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
## 46                                                      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.