Statistical Language Models

Week 11-ish

- #https://genius.com song lyrics:
- library(genius)
- library(tidytext)
- library(tidyverse)
- library(textdata)
- **library(janitor)**

Beatles

- Albums span about 10 years. Started as lever-pullers who wanted to make music. Eventually were called to Pepperland to battle the Blue Meanies restoring psychedelic music to all.

Is there a difference in the sentiment allocation in Beatles songs between the start and end of their recording career?

- Ho? Ha?

Bieber data

- Artist = "Justin Bieber"
- Album = "My world ep"
- tracklist = genius_tracklist(artist=Artist, album = Album)
- #All lyrics from an album:
- LyricsJB1 = NULL
- for(songNumber in dim(tracklist)[1]:1){
- LyricsJB1 = rbind(LyricsJB1,genius_lyrics(artist=Artist ,song= tracklist$track_title[songNumber]))
- }
- Album = "Changes"
- tracklist = genius_tracklist(artist=Artist, album = Album)
- tracklist$track_title[15] = "Thats What Love Is"
- tracklist$track_title[12] = "eta"
- LyricsJB2 = NULL
- for(songNumber in (dim(tracklist)[1]-1):1){
- LyricsJB2 = rbind(LyricsJB2,genius_lyrics(artist=Artist ,song= tracklist$track_title[songNumber]))
- }
- LyricsJB2

Plan

- Obtain first and last album lyrics
- Convert to sentiment categories
- Make a cross-tab table counting the sentiment mentions within each album
- Make a plot
- Do a test!

Sentiments

- get_sentiments("nrc")
- table(get_sentiments("nrc")$sentiment)

- Artist = "The beatles"
- Album = "Please please me"
- tracklist = genius_tracklist(artist=Artist, album = Album)
- # fix a punctuation problem
- #
- tracklist$track_title[9] = "ps i love you"
- #All lyrics from album
- Lyrics1 = NULL
- for(songNumber in dim(tracklist)[1]:1){
- Lyrics1 = rbind(Lyrics1,genius_lyrics(artist=Artist ,song= tracklist$track_title[songNumber]))
- }

- Album = "Let it be"
- tracklist = genius_tracklist(artist=Artist, album = Album)

- #All lyrics from an album:
- Lyrics2 = NULL
- for(songNumber in dim(tracklist)[1]:1){
- Lyrics2 = rbind(Lyrics2,genius_lyrics(artist=Artist ,song= tracklist$track_title[songNumber]))
- }

- Lyrics1 = Lyrics1 %>% mutate(album= "Please please me")
- Lyrics2 = Lyrics2 %>% mutate(album= "Let it be")
- Lyrics = rbind(Lyrics1,Lyrics2) %>%
    - unnest_tokens(output = word,input = lyric, token = "words") %>%
    - inner_join(get_sentiments("nrc"))
- #Count the occurrence with each album
- Sents = Lyrics %>%group_by(album)%>%count(sentiment)
-

Distribution of sentiment within album

- # Stacked percent
- Sents %>% ggplot(aes(fill=sentiment, y=n, x=album)) +
- geom_bar(position="fill", stat="identity")

Cross-Tab table

- #Cross-tab table:
- MyTable = Sents %>% spread(album, n)
- #OR (better for later)
- MyTable = tabyl(Lyrics %>%group_by(album),sentiment,album)
- MyTable %>% adorn_totals("row")
- MyTable %>% adorn_totals("col")
- MyTable %>% adorn_percentages("row")
- MyTable %>% adorn_percentages("col")
- MyTable %>% adorn_percentages("all")

Easy formatting for presentations

- MyTable %>% adorn_percentages("row")%>% adorn_ns()
- MyTable %>% adorn_percentages("col")%>% adorn_ns()
- MyTable %>% adorn_percentages("all")%>% adorn_ns()

Chi Square Test for relationship between categorical variables

- Observed counts are the data
- Ho: There is no relationship between variables (album and sentiment); aka. The difference in counts are due to different numbers of words within each album.
- Expected counts are the data assuming the only difference is the total count; E_{ij}= (row i total*column j total) / N_{total}
- $$X = \sum_{i=1}^{N_r} \sum_{j=1}^{N_c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(Nr-1)*(N_c-1)}$$

Chi-Square Test

- MyTable = tabyl(Lyrics %>%group_by(album),sentiment,album)
- chisq.test(MyTable)

2 x 2 table

- PNTable = MyTable%>% filter(sentiment == "positive" | sentiment=="negative")

- fisher.test(PNTable[,2:3])

- Fisher's Exact test is better with a 2x2 table; instead of χ^2 it uses a hyper-geometric distribution for the counts given the row and column totals.

- fisher.test(PNTable[,2:3])

Project

- 10 minute video due end of day Apri 1.
- We will show videos in class starting on April 2 (probably spilling over to April 7). You must be present for answering questions about your work.
- 2-3 page report due April 15th.