Final Project Notebook

DS 5001 Exploratory Text Analytics | Spring 2024

Metadata

- Full Name: Ryan Lipps
- Userid: rhl8pk
- GitHub Repo URL: https://github.com/rlipps/Exploratory_Text_Analytics_Final
- UVA Box URL: https://virginia.app.box.com/folder/262013200623

Overview

The goal of the final project is for you to create a **digital analytical edition** of a corpus using the tools, practices, and perspectives you've learning in this course. You will select a corpus that has already been digitized and transcribed, parse that into an F-compliant set of tables, and then generate and visualize the results of a series of fitted models. You will also draw some tentative conclusions regarding the linguistic, cultural, psychological, or historical features represented by your corpus. The point of the exercise is to have you work with a corpus through the entire pipeline from ingestion to interpretation.

Specifically, you will acquire a collection of long-form texts and perform the following operations:

- **Convert** the collection from their source formats (F0) into a set of tables that conform to the Standard Text Analytic Data Model (F2).
- **Annotate** these tables with statistical and linguistic features using NLP libraries such as NLTK (F3).
- **Produce** a vector representation of the corpus to generate TFIDF values to add to the TOKEN (aka CORPUS) and VOCAB tables (F4).
- **Model** the annotated and vectorized model with tables and features derived from the application of unsupervised methods, including PCA, LDA, and word2vec (F5).
- **Explore** your results using statistical and visual methods.

- **Present** conclusions about patterns observed in the corpus by means of these operations.

When you are finished, you will make the results of your work available in GitHub (for code) and UVA Box (for data). You will submit to Gradescope (via Canvas) a PDF version of a Jupyter notebook that contains the information listed below.

Some Details

- Please fill out your answers in each task below by editing the markdown cell.
- Replace text that asks you to insert something with the thing, i.e. replace `(INSERT IMAGE HERE)` with an image element, e.g. `![](image.png)`.
- For URLs, just paste the raw URL directly into the text area. Don't worry about providing link labels using `[label](link)`.
- Please do not alter the structure of the document or cell, i.e. the bulleted lists.
- You may add explanatory paragraphs below the bulleted lists.
- Please name your tables as they are named in each task below.
- Tasks are indicated by headers with point values in parentheses.

Raw Data

Source Description (1)

Provide a brief description of your source material, including its provenance and content. Tell us where you found it and what kind of content it contains.

The source material is a collection of some of my favorite musicians' albums. It consists of 953 songs from 89 albums from 18 artists. The data is from both the Spotify API and lyricsgenius API. I used the Spotify API to get all of the artist, album, and song metadata, and used this information to query the lyricsgenius API to get the lyrics. It involved heavy cleaning, as the code to get the lyrics was automated as initially the full song list consisted of over 1000 songs. There were some plays, song lists, and many other interesting things that the API returned.

Source Features (1)

Add values for the following items. (Do this for all following bulleted lists.)

- Source URL: https://developer.spotify.com/documentation/web-api and https://docs.genius.com/
- UVA Box URL: https://virginia.box.com/s/66ye5cvhcpazvqi19qpbzwfqqibk8scf
- Number of raw documents: 89
- Total size of raw documents (e.g. in MB): 1.2 MB
- File format(s), e.g. XML, plaintext, etc.: .txt

Source Document Structure (1)

Provide a brief description of the internal structure of each document. That, describe the typical elements found in document and their relation to each other. For example, a corpus of letters might be described as having a date, an addressee, a salutation, a set of content paragraphs, and closing. If they are various structures, state that.

Each document is an album, with lyrics for each song, in the order of the tracklist. If a song is instrumental it was skipped. For the most part the genius API returns lyrics following some structural information such as [VERSE 1]...[PRE-CHORUS 1]... [CHORUS 1], though this is not consistent. Overall for this project, bags were either ALBUM or SONG

Parsed and Annotated Data

Parse the raw data into the three core tables of your addition: the `LIB`, `CORPUS`, and `VOCAB` tables.

These tables will be stored as CSV files with header rows.

You may consider using `|` as a delimitter.

Provide the following information for each.

LIB (2)

The source documents the corpus comprises. These may be books, plays, newspaper articles, abstracts, blog posts, etc.

Note that these are *not* documents in the sense used to describe a bag-of-words representation of a text, e.g. chapter.

- UVA Box URL: https://virginia.box.com/s/oopf095w1m89gh6h85dblfu88nnfk6qf
- GitHub URL for notebook used to create: https://github.com/rlipps/Exploratory_Text_Analytics_Final/blob/main/notebooks/Build_Tables.ipynb
- Delimitter: |
- Number of observations: 89
- List of features, including at least three that may be used for model summarization (e.g. date, author, etc.): 'album_name', 'album_title', 'artist', 'source_file_path', 'song_regex', 'genres', 'release_date', 'label', 'mean_danceability', 'mean_energy', 'mean_loudness', 'mean_speechiness', 'mean_acousticness', 'mean_instrumentalness', 'mean_liveness', 'mean_valence', 'mean_tempo', 'album_term_count', 'album_character_count', 'genre'
- Average length of each document in characters: 8749.236

CORPUS (2)

The sequence of word tokens in the corpus, indexed by their location in the corpus and document structures.

- UVA Box URL: https://virginia.box.com/s/vfhwbfbga0d7xl59x5trif85gjfem2gv
- GitHub URL for notebook used to create: https://github.com/rlipps/Exploratory_Text_Analytics_Final/blob/main/notebooks/Build_Tables.ipynb
- Delimitter: |
- Number of observations Between (should be >= 500,000 and <= 2,000,000 observations.): 192607
- OHCO Structure (as delimitted column names): album_id, song_num, stanza_num, line_num, token_num
- Columns (as delimitted column names, including `token_str`, `term_str`, `pos`, and `pos_group`): ['pos_tuple', 'pos', 'token_str', 'term_str', 'pos_group']

VOCAB (2)

The unique word types (terms) in the corpus.

- UVA Box URL: https://virginia.box.com/s/y9fzwn51o4539p4p2o5hvupuw30ujw04
- GitHub URL for notebook used to create: https://github.com/rlipps/Exploratory_Text_Analytics_Final/blob/main/notebooks/Build_Tables.ipynb
- Delimitter: |
- Number of observations: 9110
- Columns (as delimitted names, including `n`, `p`', `i`, `dfidf`, `porter_stem`, `max_pos` and `max_pos_group`, `stop`): n,n_chars, p, i, max_pos, max_pos_group, stop, porter_stem, song_dfidf, album_dfidf
- Note: Your VOCAB may contain ngrams. If so, add a feature for `ngram_length`.
- List the top 20 significant words in the corpus by DFIDF.

['him', 'bad', 'true', 'own', 'living', 'fire', 'first', 'found', 'feet', 'someone', 'ask', 'kind', 'play', 'took', 'everyone', 'myself', 'seen', 'save', 'once', 'sky']

Derived Tables

BOW (3)

A bag-of-words representation of the CORPUS.

- UVA Box URL: https://virginia.box.com/s/vjmoudem87ohv1h9u0c8e9161syf7cbk
- GitHub URL for notebook used to create: https://github.com/rlipps/Exploratory_Text_Analytics_Final/blob/main/notebooks/Build_Tables.ipynb
- Delimitter: |
- Bag (expressed in terms of OHCO levels): album_id, OHCO[:1]
- Number of observations: 43158
- Columns (as delimitted names, including `n`, `tfidf`): n, tfidf

DTM (3)

A represenation of the BOW as a sparse count matrix.

- UVA Box URL: https://virginia.box.com/s/2vj1kgimc8r31oiovr0up5jxqh02mpca
- UVA Box URL of BOW used to generate (if applicable): https://virginia.box.com/s/vjmoudem87ohv1h9u0c8e9161syf7cbk
- GitHub URL for notebook used to create: https://github.com/rlipps/Exploratory_Text_Analytics_Final/blob/main/notebooks/Build_Tables.ipynb
- Delimitter: |
- Bag (expressed in terms of OHCO levels): album_id, OHCO[:1]

TFIDF (3)

A Document-Term matrix with TFIDF values.

- UVA Box URL: https://virginia.box.com/s/u07hfehp69gy8kdy6p63l6dh7tclivqx
- UVA Box URL of DTM or BOW used to create: https://virginia.box.com/s/2vj1kgimc8r31oiovr0up5jxqh02mpca
- GitHub URL for notebook used to create: https://github.com/rlipps/Exploratory_Text_Analytics_Final/blob/main/notebooks/Build_Tables.ipynb
- Delimitter: |
- Description of TFIDIF formula ($\LaTeX$ OK): $\frac{TF}{TF_{max}} * log_2 \frac{N}{DF}$

Reduced and Normalized TFIDF_L2 (3)

A Document-Term matrix with L2 normalized TFIDF values.

- UVA Box URL: https://virginia.box.com/s/tu3n2sgmvgbs7275229t4qwwfokuiu2d
- UVA Box URL of source TFIDF table: https://virginia.box.com/s/u07hfehp69gy8kdy6p63l6dh7tclivqx
- GitHub URL for notebook used to create: https://github.com/rlipps/Exploratory_Text_Analytics_Final/blob/main/notebooks/Build_Tables.ipynb
- Delimitter: |
- Number of features (i.e. significant words): 2000
- Principle of significant word selection: I used only singular and plural nouns and removed stopwords. The inclusion of stopwords is that there are many stopwords that were getting marked as significant because it appears that they're used

as ad-libs and vocalizations in songs, which was adding a lot of noise. I used only nouns as using multiple parts of speech was very noisy and not very interprable.

Models

PCA Components (4)

- UVA Box URL: https://virginia.box.com/s/2unm4woxrwt99s22mjujiuvhhtiarsmj
- UVA Box URL of the source TFIDF_L2 table: https://virginia.box.com/s/tu3n2sgmvgbs7275229t4qwwfokuiu2d
- GitHub URL for notebook used to create: https://github.com/rlipps/Exploratory_Text_Analytics_Final/blob/main/notebooks/PCA.ipynb
- Delimitter: |
- Number of components: 10
- Library used to generate: by hand, with scipy linalg
- Top 5 positive terms for first component: mystery step hes lie everything
- Top 5 negative terms for second component: someones denial hes round hope

PCA DCM (4)

The document-component matrix generated.

- UVA Box URL: https://virginia.box.com/s/rh8bcr8e2tjdw4yngk921c8w2gcgrrdc
- GitHub URL for notebook used to create: https://github.com/rlipps/Exploratory_Text_Analytics_Final/blob/main/notebooks/PCA.ipynb
- Delimitter: |

PCA Loadings (4)

The component-term matrix generated.

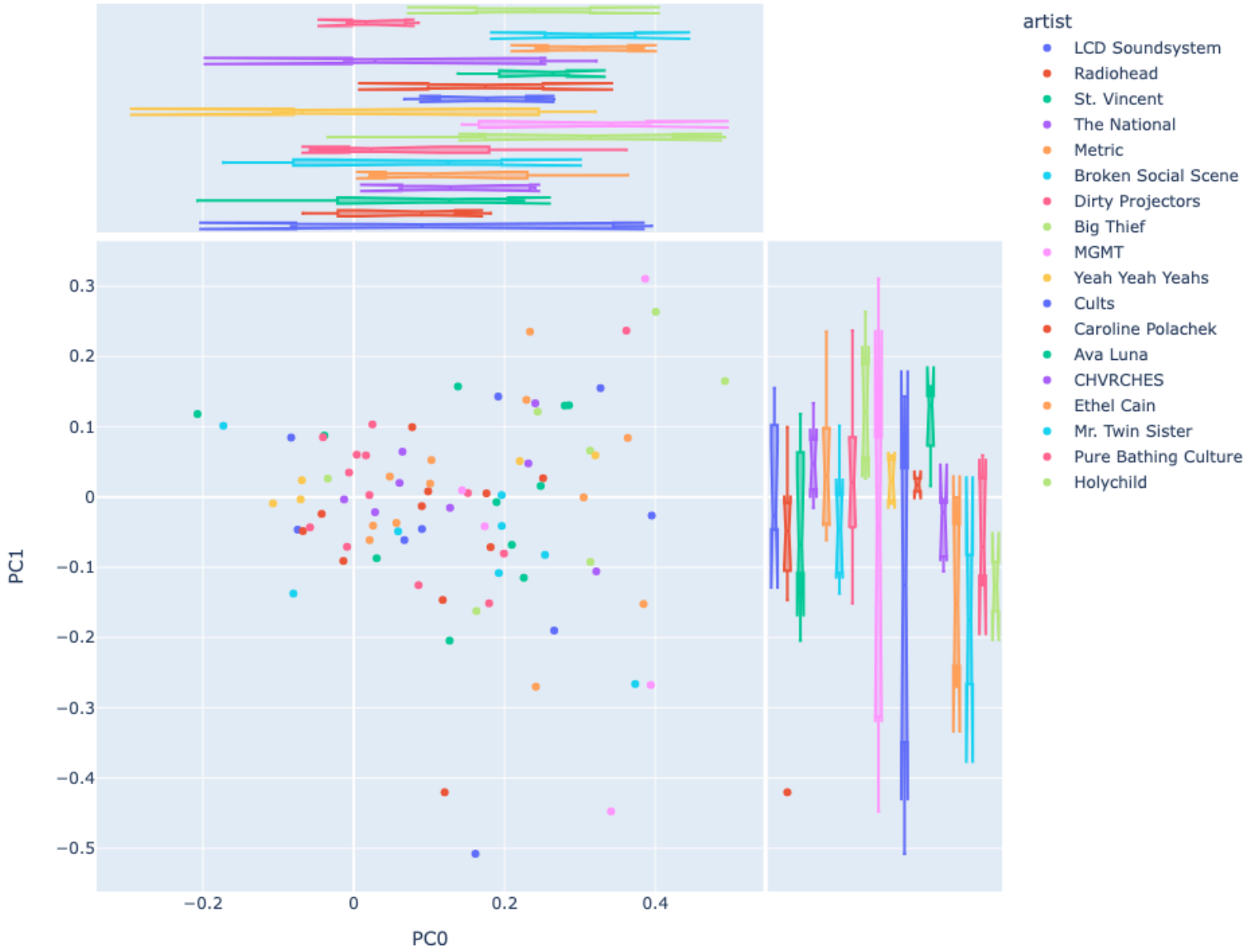
- UVA Box URL: https://virginia.box.com/s/k8q2qjmq4vnjwqh3agi4ulugrqjovks0
- GitHub URL for notebook used to create: https://github.com/rlipps/Exploratory_Text_Analytics_Final/blob/main/notebooks/PCA.ipynb
- Delimitter: |

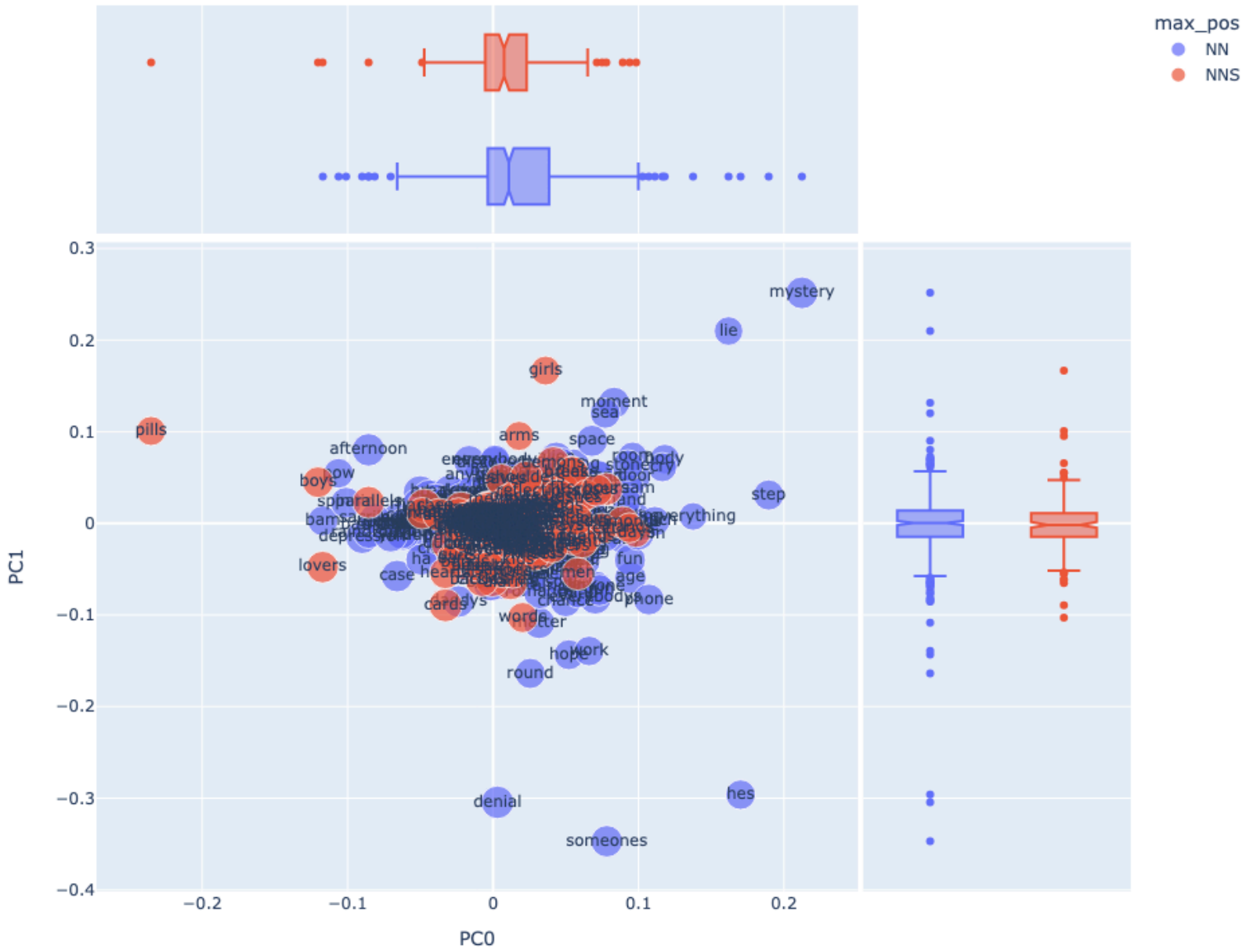
PCA Visualization 1 (4)

Include a scatterplot of documents in the space created by the first two components.

Color the points based on a metadata feature associated with the documents.

Also include a scatterplot of the loadings for the same two components. (This does not need a feature mapped onto color.)


artist
LCD Soundsystem
Radiohead
St. Vincent
The National
Metric
Broken Social Scene
Dirty Projectors
Big Thief
MGMT
Yeah Yeah Yeahs
Cults
Caroline Polachek
Ava Luna
CHVRCHES
Ethel Cain
Mr. Twin Sister
Pure Bathing Culture
Holychild
PC1
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.2
0
0.2
0.4
PC0


max_pos
NN
NNS
PC0
PC1
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
−0.2
−0.1
0
0.1
0.2
mystery
lie
girls
moment
sea
pills
arms
space
afternoon
step
everything
fun
age
phone
work
round
hes
denial
someones
lovers
cards
words

Briefly describe the nature of the polarity you see in the first component:

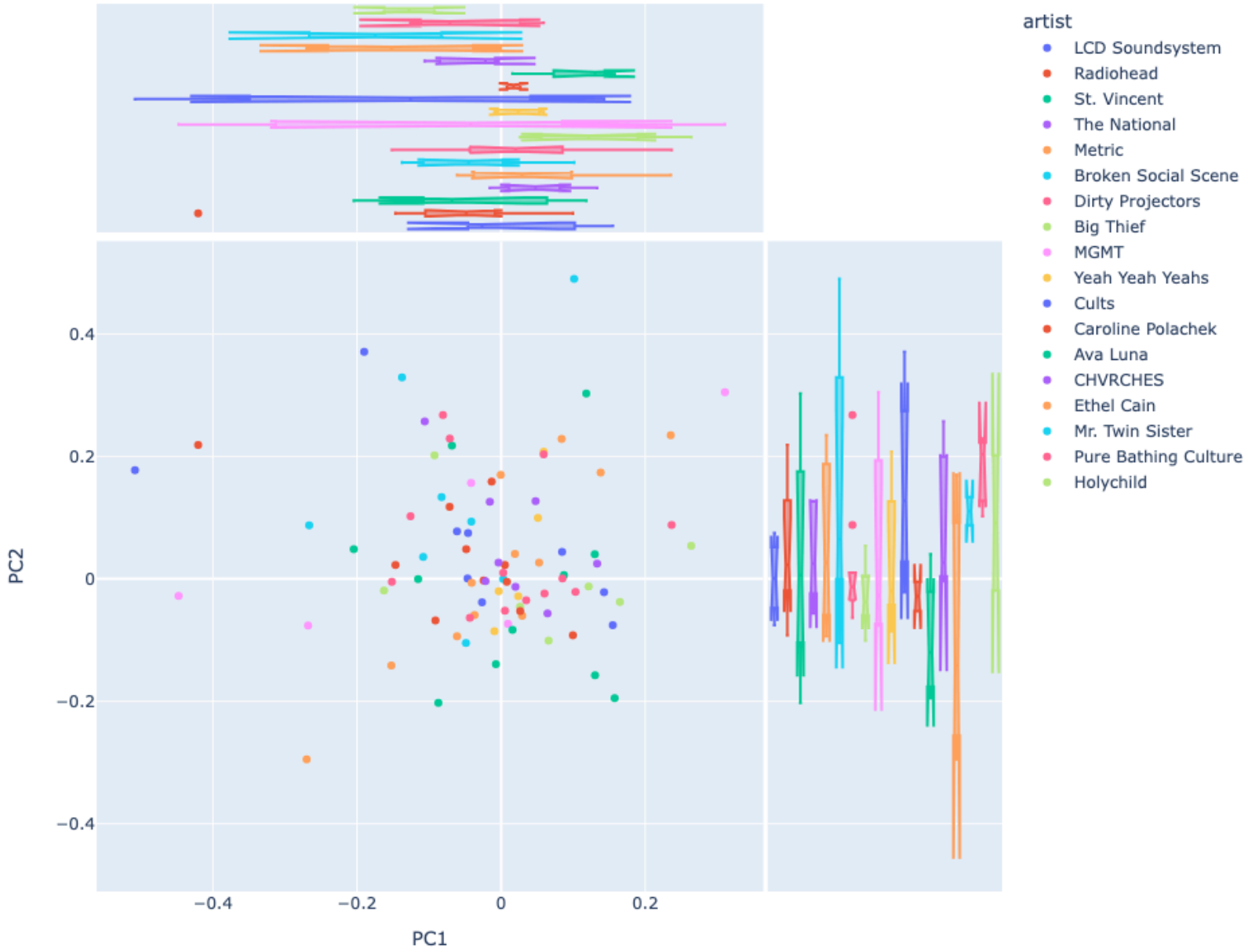
The polarity in the first component looks like teenage angst on the negative axis and more mature concepts on the positive axis

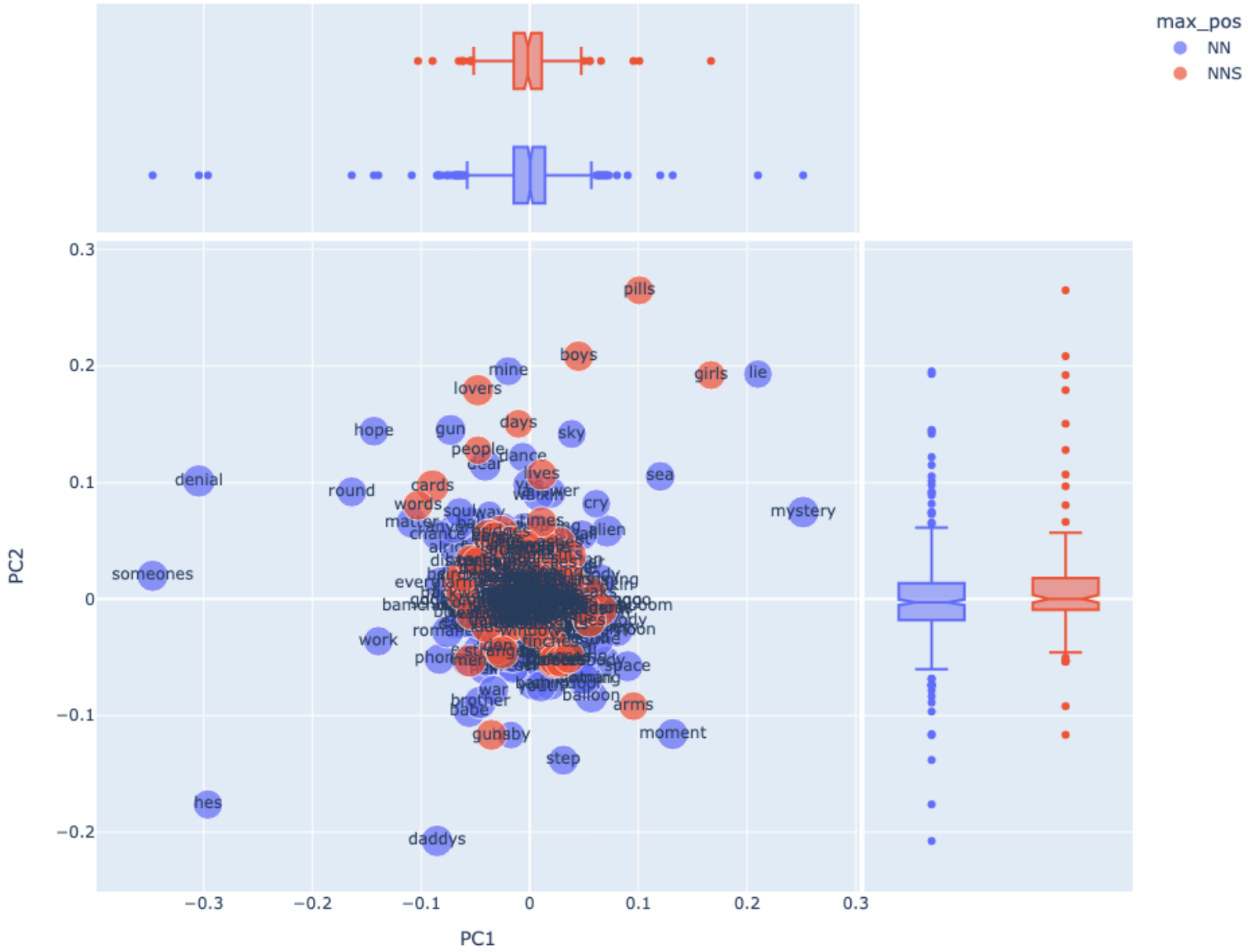
PCA Visualization 2 (4)

Include a scatterplot of documents in the space created by the second two components.

Color the points based on a metadata feature associated with the documents.

Also include a scatterplot of the loadings for the same two components. (This does not need a feature mapped onto color.)


artist
LCD Soundsystem
Radiohead
St. Vincent
The National
Metric
Broken Social Scene
Dirty Projectors
Big Thief
MGMT
Yeah Yeah Yeahs
Cults
Caroline Polachek
Ava Luna
CHVRCHES
Ethel Cain
Mr. Twin Sister
Pure Bathing Culture
Holychild
PC2
0.4
0.2
0
−0.2
−0.4
PC1
−0.4
−0.2
0
0.2


max_pos
NN
NNS
PC1
PC2
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
pills
boys
girls
lie
mine
lovers
hope
gun
days
sky
people
dance
sea
denial
round
cards
words
cry
mystery
someones
alien
work
space
balloon
arms
moment
brother
babe
gunsby
step
hes
daddys

Briefly describe the nature of the polarity you see in the second component:

The polarity in the second component seems to related to general feelings and teenage angst again on the positive axis

LDA TOPIC (4)

- UVA Box URL: https://virginia.box.com/s/1jcx53mxamzg1twz37uv9gv6l01knoj7
- UVA Box URL of count matrix used to create: https://virginia.box.com/s/v2rq9jzsq00lleq9sndfqu9r4kpuqhz1
- GitHub URL for notebook used to create: https://github.com/rlipps/Exploratory_Text_Analytics_Final/blob/main/notebooks/LDA.ipynb
- Delimitter: |
- Libary used to compute: sklearn
- A description of any filtering, e.g. POS (Nouns and Verbs only): Nouns only
- Number of components: 10
- Any other parameters used: removed stopwords for previous description
- Top 5 words and best-guess labels for topic five topics by mean document weight:
  - T00: way time something eyes heart: this topic seems to indicate some waywardness and being lost
  - T01: time war house love arms: this topic seems to be about conflict, maybe relating to domestic relationships
  - T02: love life sacrilege eyes look: this topic seems to be about infidelity or lust
  - T03: time raindrops guns duh days: this topic seems to be about malaise
  - T04: dreamt baby love everything lie: this topic seems to be about uncertainty in relationships

LDA THETA (4)

- UVA Box URL: https://virginia.box.com/s/5b0t68l18xxefb3g5q0u72rjwn9qtw1m
- GitHub URL for notebook used to create: https://github.com/rlipps/Exploratory_Text_Analytics_Final/blob/main/notebooks/LDA.ipynb
- Delimitter: |

LDA PHI (4)

- UVA Box URL: https://virginia.box.com/s/qc69sz3q84euwkehr4z1hsfgdzoy68ki
- GitHub URL for notebook used to create: https://github.com/rlipps/Exploratory_Text_Analytics_Final/blob/main/notebooks/LDA.ipynb
- Delimitter: |

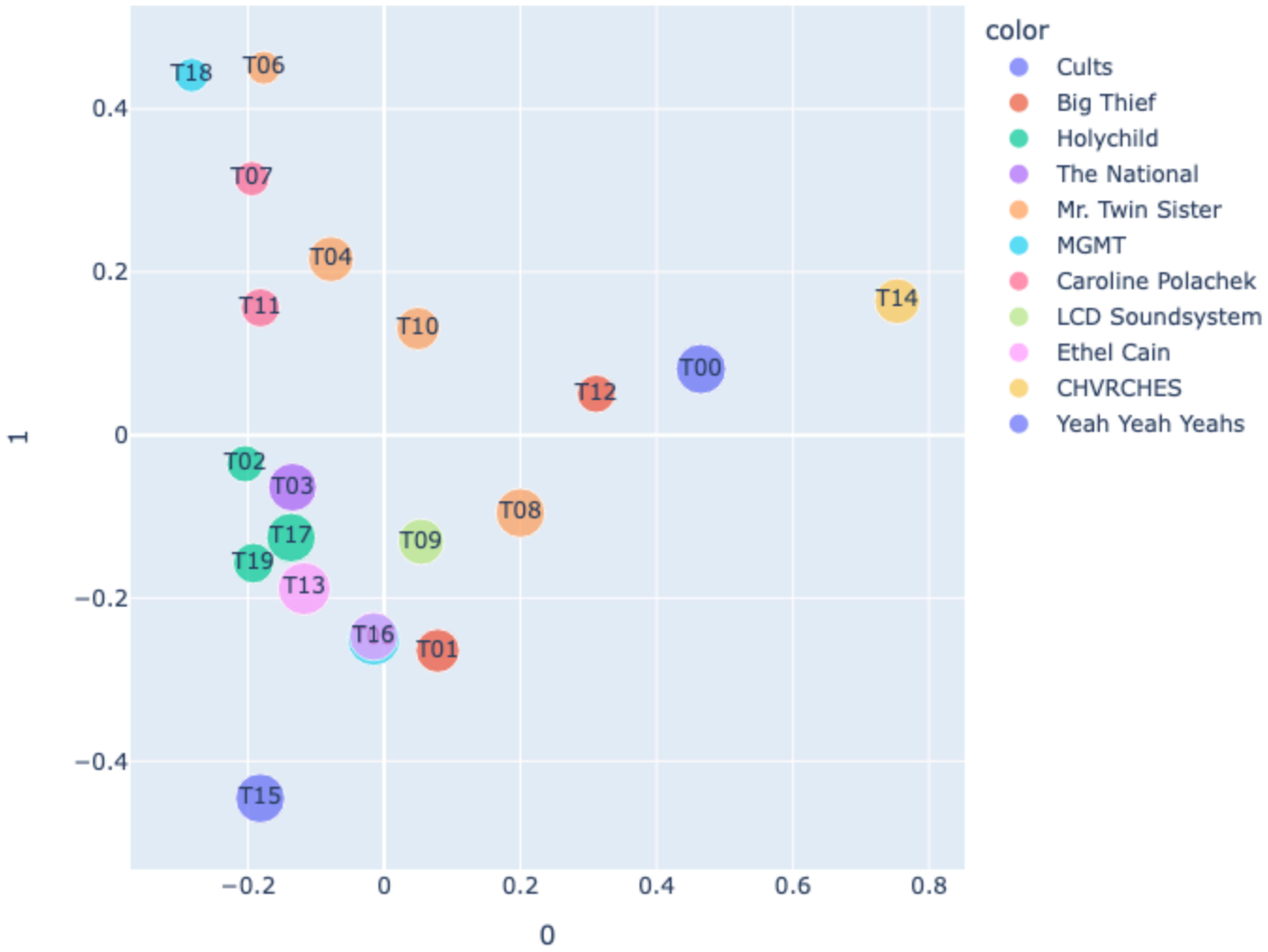
LDA + PCA Visualization (4)

Apply PCA to the PHI table and plot the topics in the space opened by the first two components.

Size the points based on the mean document weight of each topic (using the THETA table).

Color the points basd on a metadata feature from the LIB table.

Provide a brief interpretation of what you see.



- A few artists are not max assigned to a topic.
- HOLYCHILD and Caroline Polachek are separated along PC1, which is interesting given they're relatively similar in genre.
- Using hover info, max artist assignment and max album assignment don't always line up for topics

- Topics 1 and 12 have agreement between Big Thief and Big Thief albums, and in my opinion those two albums are definitely their most focused.

Sentiment VOCAB_SENT (4)

Sentiment values associated with a subset of the VOCAB from a curated sentiment lexicon.

- UVA Box URL: https://virginia.box.com/s/0816t62hj7ztuj0v6sks2aisnk32o0ou
- UVA Box URL for source lexicon: https://virginia.box.com/s/0r5ateb6mu8fq5nw6uaabgqfslq8dslq
- GitHub URL for notebook used to create: https://github.com/rlipps/Exploratory_Text_Analytics_Final/blob/main/notebooks/Sentiment_Analysis.ipynb
- Delimitter: |

Sentiment BOW_SENT (4)

Sentiment values from VOCAB_SENT mapped onto BOW.

- UVA Box URL: https://virginia.box.com/s/apkph60vbm2k9tpn44bwj0cdh0ow2zr9
- GitHub URL for notebook used to create: https://github.com/rlipps/Exploratory_Text_Analytics_Final/blob/main/notebooks/Sentiment_Analysis.ipynb
- Delimitter: |

Sentiment DOC_SENT (4)

Computed sentiment per bag computed from BOW_SENT.

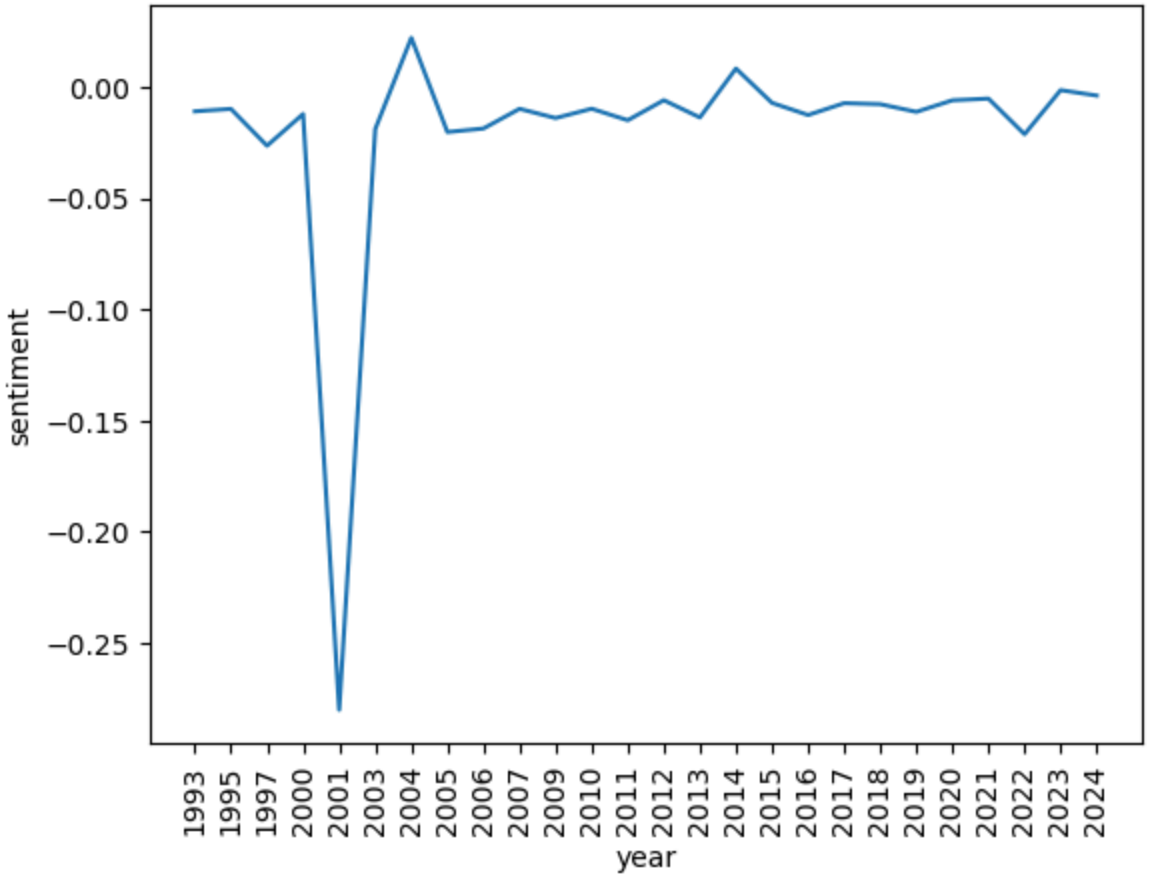
- UVA Box URL: https://virginia.box.com/s/j9i03fymod7kse758ov0qf0uu5t43jtf
- GitHub URL for notebook used to create: https://github.com/rlipps/Exploratory_Text_Analytics_Final/blob/main/notebooks/Sentiment_Analysis.ipynb
- Delimitter: |
- Document bag expressed in terms of OHCO levels: album_id, OHCO[:1]

Sentiment Plot (4)

Plot sentiment over some metric space, such as time.

If you don't have a metric metadata features, plot sentiment over a feature of your choice.

You may use a bar chart or a line graph.



VOCAB_W2V (4)

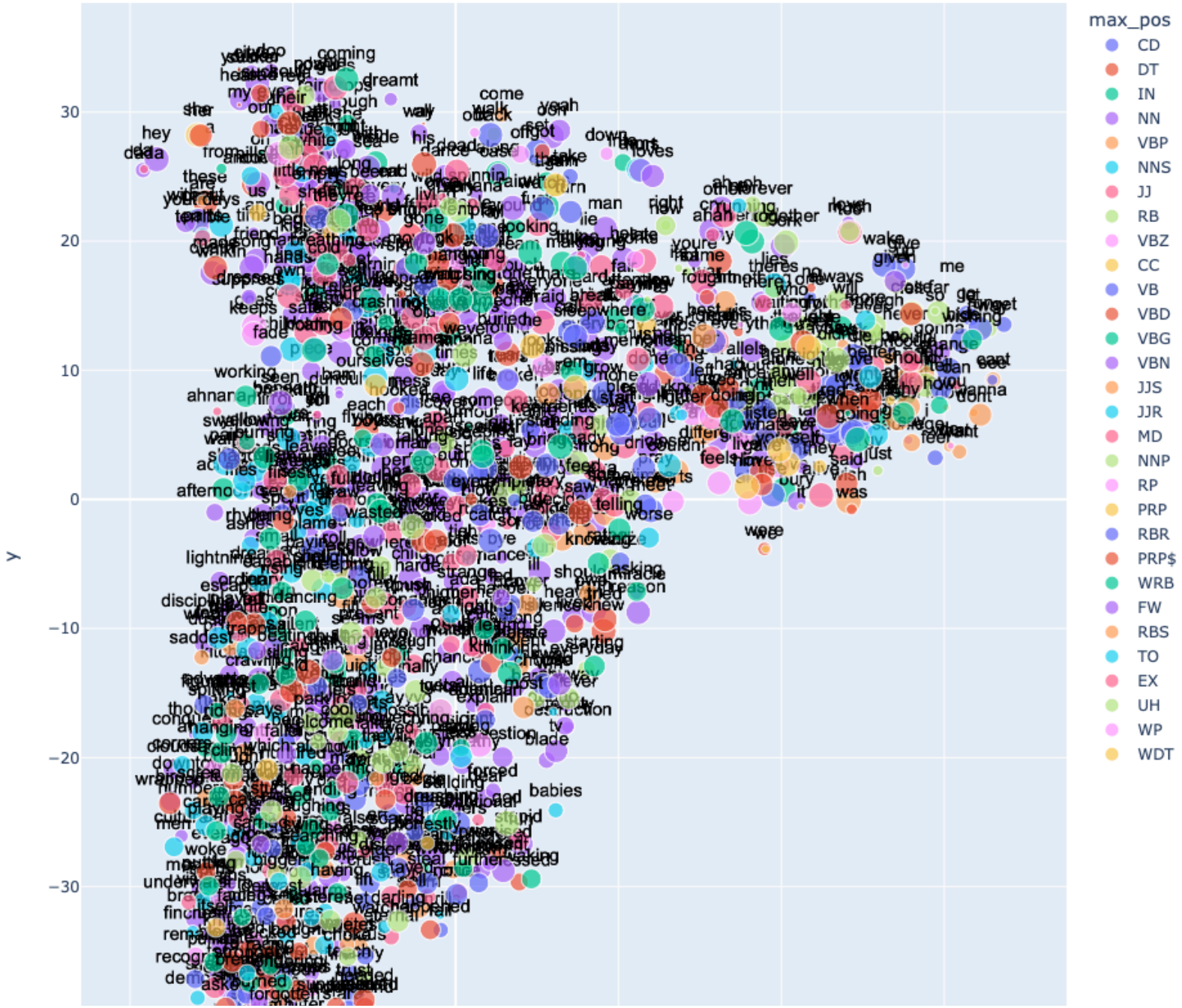
A table of word2vec features associated with terms in the VOCAB table.

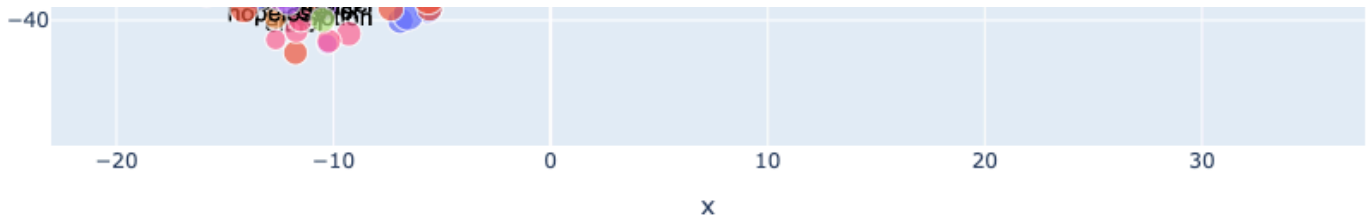
- UVA Box URL: https://virginia.box.com/s/k1337c8mhi3o9640a2v7r0j0se7ke4un
- GitHub URL for notebook used to create: https://github.com/rlipps/Exploratory_Text_Analytics_Final/blob/main/notebooks/Word_Embeddings.ipynb
- Delimitter: |
- Document bag expressed in terms of OHCO levels: album_id, OHCO[:1]
- Number of features generated: 246
- The library used to generate the embeddings: gensim

Word2vec tSNE Plot (4)

Plot word embedding featues in two-dimensions using t-SNE.

Describe a cluster in the plot that captures your attention.


max_pos
CD
DT
IN
NN
VBP
NNS
JJ
RB
VBZ
CC
VB
VBD
VBG
VBN
JJS
JJR
MD
NNP
RP
PRP
RBR
PRP$
WRB
FW
RBS
TO
EX
UH
WP
WDT
y
30
20
10
0
−10
−20
−30



Zooming in on the bottom there is a lot of emotion and relationship related words, such as: hopeless, devotion, understand, believed, trust, needed, loving, regret

Riffs

Provde at least three visualizations that combine the preceding model data in interesting ways.

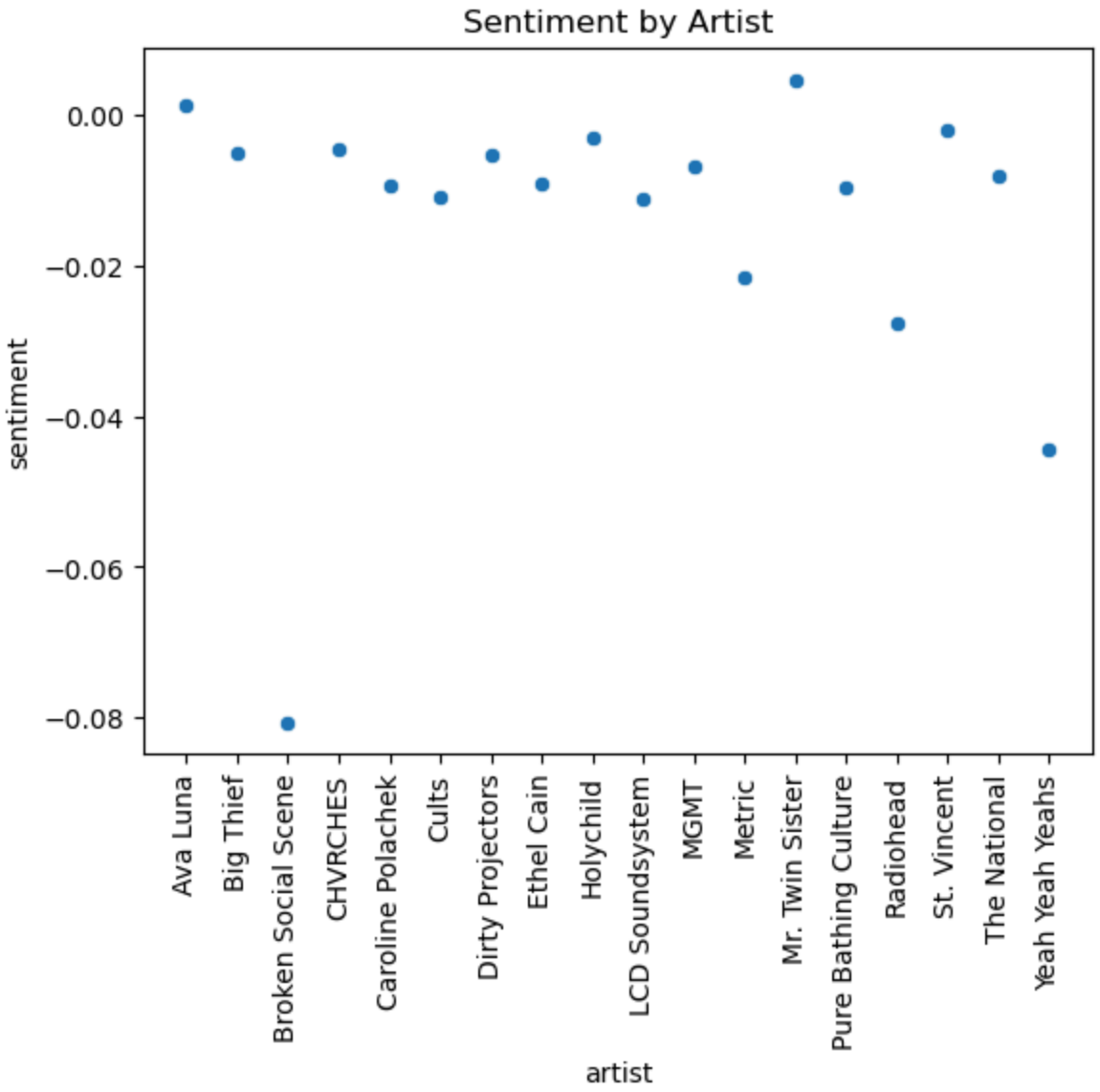
These should provide insight into how features in the LIB table are related.

The nature of this relationship is left open to you -- it may be correlation, or mutual information, or something less well defined.

In doing so, consider the following visualization types:

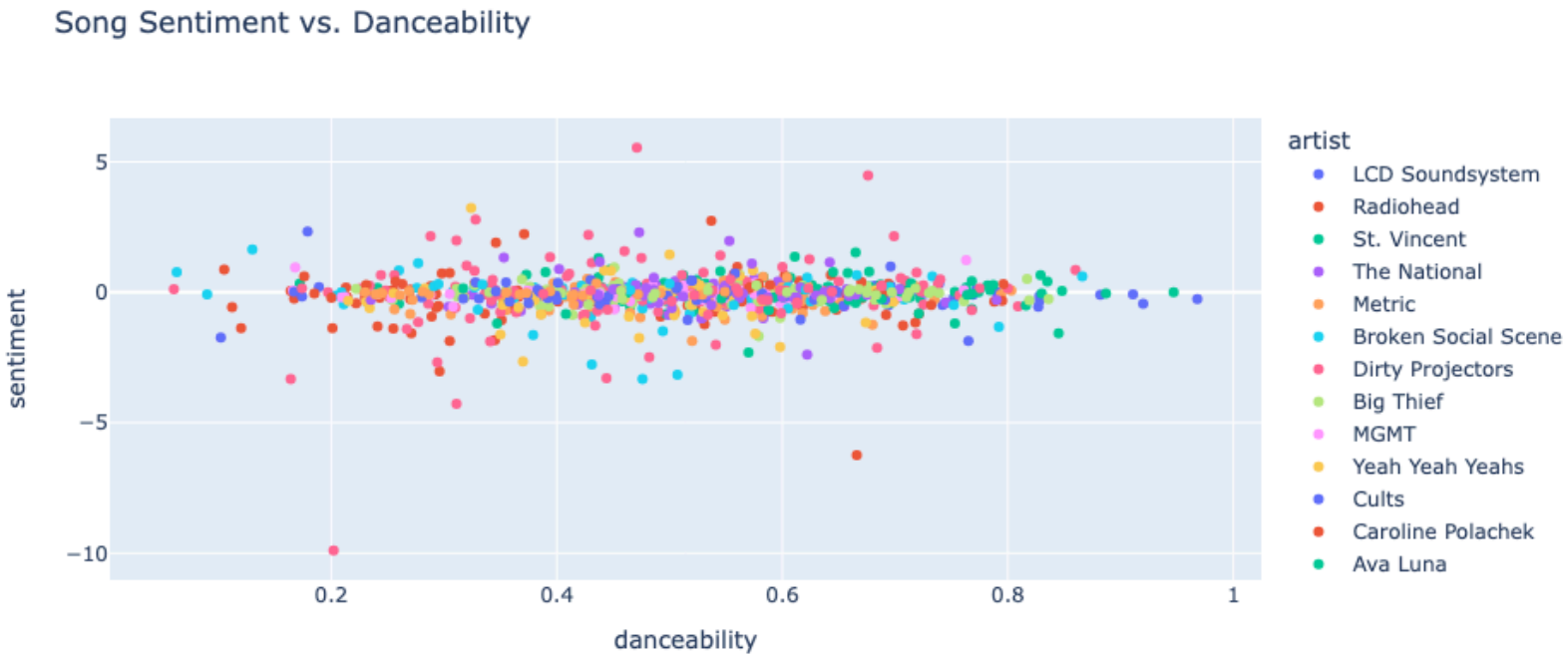
- Hierarchical cluster diagrams
- Heatmaps
- Scatter plots
- KDE plots
- Dispersion plots
- t-SNE plots
- etc.

Riff 1 (5)



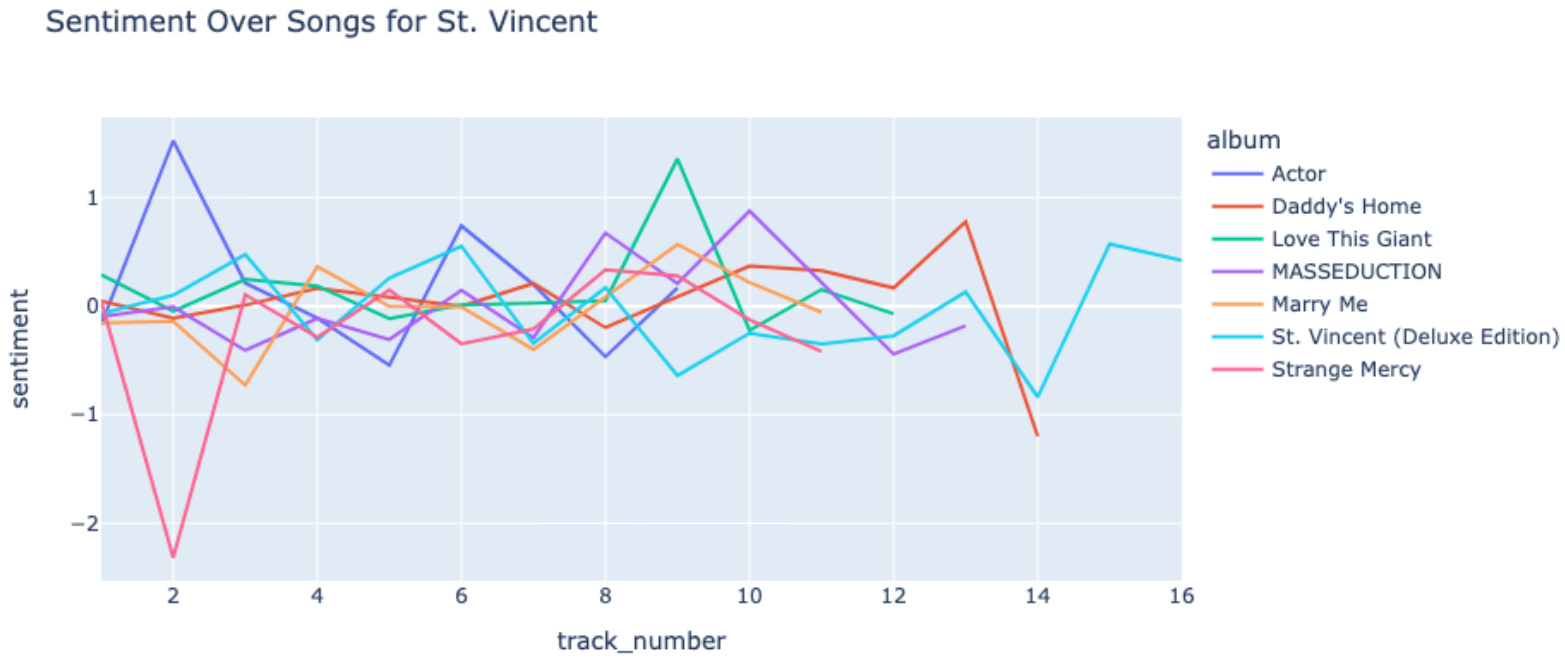
Generally my work indicated that I listen to a lot of sad and emotional music. I was looking to see who is the saddest artist with this. My guess is that it would be the National or Radiohead. I was surprised to see that Yeah Yeah Yeahs was as sad as they were because their music is generally upbeat.

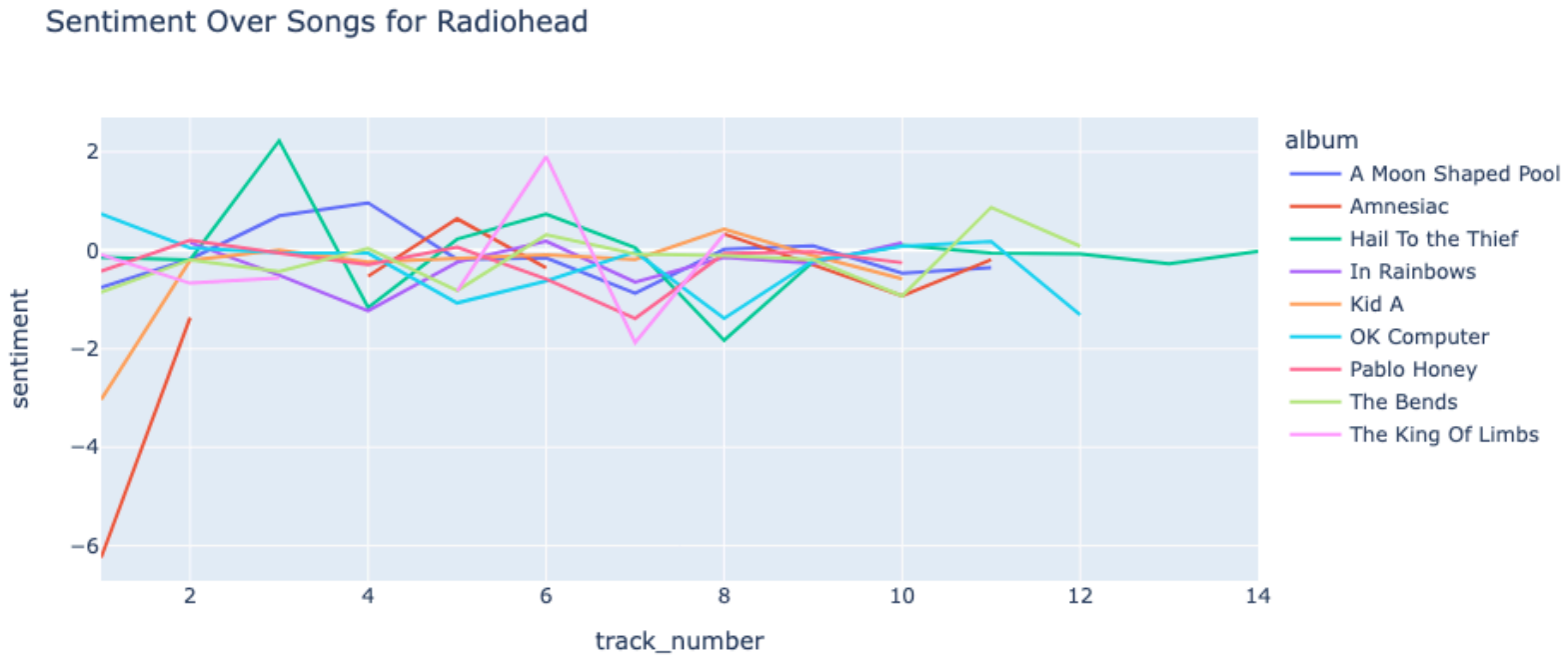
Riff 2 (5)

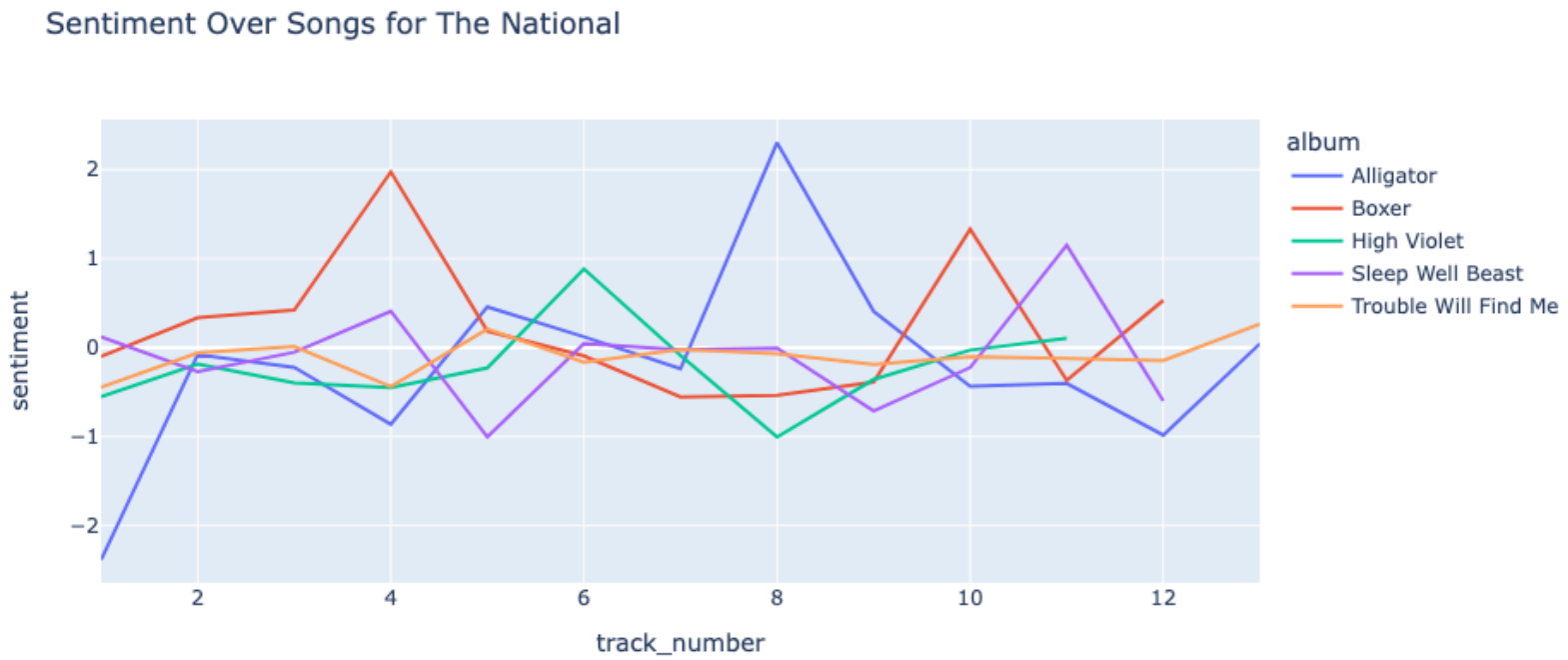


For this riff I was looking to see if song sentiment was related to danceability. One might expect that happier songs are more danceable, however there doesn't appear to be a correlation. For note, this is Spotify's definition of danceability: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.

Riff 3 (5)


Sentiment Over Songs for St. Vincent
sentiment
track_number
album
Actor
Daddy's Home
Love This Giant
MASSEDUCTION
Marry Me
St. Vincent (Deluxe Edition)
Strange Mercy


Sentiment Over Songs for Radiohead
sentiment
track_number
album
A Moon Shaped Pool
Amnesiac
Hail To the Thief
In Rainbows
Kid A
OK Computer
Pablo Honey
The Bends
The King Of Limbs



Generally, albums have narrative arcs, and I was interested to see if there was any trend in sentiment by song across album. Since album narratives told over songs generally discretized into songs rather than having arcs over chapters, I looked at this by graphing sentiment by track number grouped by album for a few artists that fleshed out catalogs. It looks like albums generally bounce around more that I thought - I was expecting more distinct shapes.

Interpretation (4)

Describe something interesting about your corpus that you discovered during the process of completing this assignment.

At a minumum, use 250 words, but you may use more. You may also add images if you'd like.

I learned a lot over the course of this project. Especially building a corpus by hand took a lot of effort, attention, and cleaning, which really makes me appreciate the people that work to build and clean text datasets for popular use.

One of the biggest realizations was that it was very difficult to find meaningful insights, as it seems as though a lot of these artists write about the same things. This makes sense as they're all my favorite bands and I suppose I have a "type", but I was expecting a little more differentiation. This highlights that even though lyrics can carry a lot of emotion, the instrumentation and delivery of the lyrics carry a significant amount of meaning. This seems trivial, but I feel as though the lack of differentiation in the lyrics analysis highlights this; each of my artists write about very similar things and use a similar vocabulary, however, their styles and musical artistic choices can be relatively disparate. With this in mind I will be doing more critical lyric-listening to see if indeed all of my favorite artists are giving more or less the same messages. Furthermore, it has inspired me to expand this project to other genres and seeing if this trend carries or if I really just like one kind of music.

As I mentioned in some of the notes, I also found that using lyrics for this project is particularly difficult. Most songs are very short when you look at the lyrics written down, and it was a lot of work to curate a reasonably sized corpus. I ran into some issues where a few songs have very few lyrics, and they were showing up in some topic modeling. For instance, LCD Soundsystem has a song called "Yeah" and...you guessed it...the only lyric is "Yeah" a bunch of times. This also highlights the different use of language in songs. A lot of stopwords can be used as ad-libs and vocalizations, and there are some people that rely heavily on this. It is an interesting use-case to include these to see who is performing a lot of melisma, but it was frustrating to deal with in getting a good signal to noise ratio for the deeper insights for this work.