Hip Hop Lyric Comparison

IST 736 Final Project

Team Members:

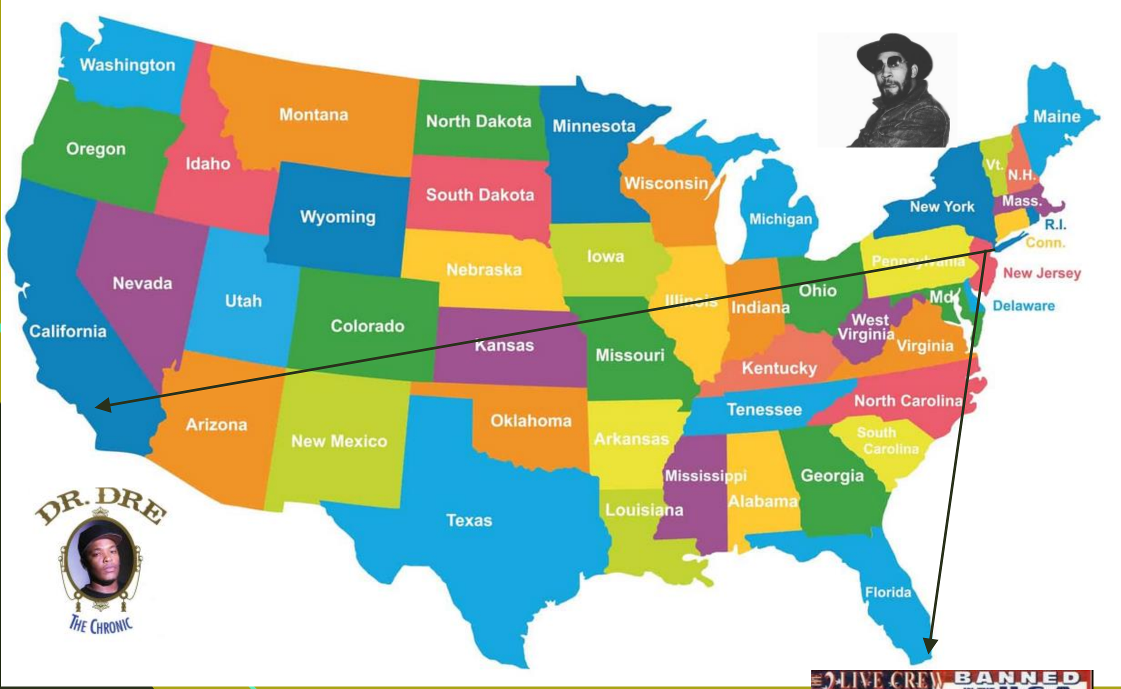
Charles Vanleuvan

Eric Gillis

Jeffrey Chao

Introduction to Hip Hop

- Hip Hop has grown from an underground sound in the 1970s into the most streamed music genre in the US in 2018[1]
- Cultural influences, past musical influences, and regional influences are prevalent in the beats and lyrics (sample culture)
- Started in the late 1970s in the Bronx, quickly gained footholds in neighborhoods across the US
    - The sound has evolved into distinct sub genres over 5 decades and across regions/cities
- The sound has evolved so much that different regions and decades have produced very different sounds
- What we really want to know: **How different is the lyrical composition of Hip Hop songs across decades and regions?**



https://newsroom.spotify.com/2018-12-04/the-top-songs-artists-playlists-and-podcasts-of-2018/

Hip Hop Subgenres

West Coast

G Funk
Gangsta rap

East Coast

Free-flow
Aggressive
Lyrical Consciousness

South

Dirty South
Trap

Midwest

Chopper

Why This Topic Matters

- The big question: **Can a subgenre of hip hop be predicted from the lyrics?**
- Sample culture in hip hop has produced radically different beats across regions (G funk on West Coast, East Coast rhyming patterns) and is well known
  - **The differences in diction, word frequency, and lyrical relations across regions/decades can offer great insight in identifying differences between regions/decades**
- Prominent influences can be identified in songs produced by collaborations
- Big Tech: This type of text modeling is likely very active at Spotify, Apple Music, Amazon Prime
  - **Can provide recommended songs to listeners even if they are listening to a new artist that isn't in an established class yet**
  - Provides a pathway to recommending songs to users building a playlist

Obtain Data

- The Genius lyrics API enables you to get album, artist, and songs information for the most popular songs in a genre

```python
import lyricsgenius
genius = lyricsgenius.Genius(token)
```

- Sample JSON response:


```
    'name': 'Eminem',
    'slug': 'Eminem',
    'url': 'https://genius.com/artists/Eminem',
    'iq': 231909}},
  '_client': <lyricsgenius.genius.Genius at 0x2ae13e09278>,
  'artist': 'Eminem',
  'lyrics': 'Obie Trice! Real name, no gimmicks (*record scratch*)\n\nTw
the outside, \'round the outside\n\nGuess who\'s back, back again\nShady'
who\'s back\nI\'ve created a monster, \'cause nobody wants to\nSee Marsha
```


- Used Billboard to get list of hip hop artists and #1 songs

Title(from Genius)	Searched Song	Artist(from Genius)	Searched Artist	Lyrics	Date	State
Hotline Bling	Hotline Bling	Drake	Drake	you used to call me on my you used to you used to yeah you used to	10/10/2015	CAN
Summer Sixteen	Summer Sixteen	Drake	Drake	looking looking looking looking looking looking looking looking look	2/20/2016	CAN
Godâ€™s Plan	God's Plan	Drake	Drake	and they wishin and wishin and wishin and wishin they wishin on m	2/3/2018	CAN

LABEL	act	actin	acting	action	actions	activate	activated
GA, CAN	0	0	0.021845	0	0	0	0
CA	0	0	0	0	0	0	0
CA	0	0	0	0	0	0	0
CA	0	0	0	0	0	0	0
CA	0	0	0	0	0	0	0
CA	0	0	0	0	0	0	0
CA	0	0	0	0	0	0	0
CA, NC	0	0.027432	0	0	0	0.0402331	0
CA, NC	0	0.027432	0	0	0	0.0402331	0
NY	0	0	0	0	0	0	0
NY	0.026804	0	0	0	0	0	0
NY	0	0	0	0	0	0	0
NY	0	0	0	0	0	0	0
NY, CA	0	0	0	0	0	0	0
NY	0	0	0	0	0	0	0

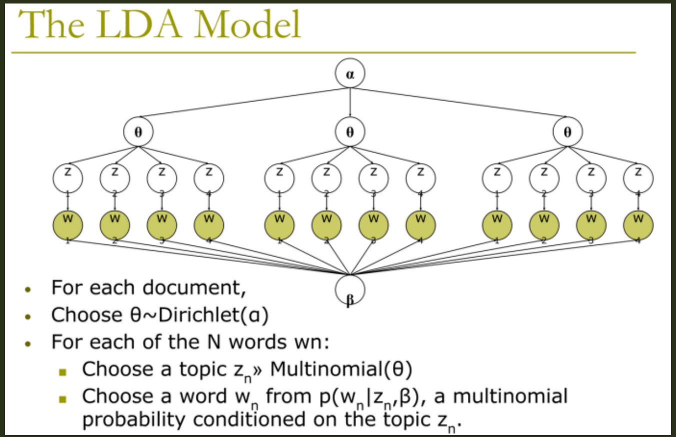
Scrub Data

- Manually annotated corpus for supervised learning
    - Decade labels (80s, 90s, 00s, 10s)
    - Region labels (East, West, South, MW, Can)
- String formatting for "feat." Vs "ft."
- Vectorize to create term document matrices
    - Add label for State/region in vectorization steps
- Remove non ASCII characters, line breaks, dashes,
- Keep numbers (e.g., 6 -> Drake)

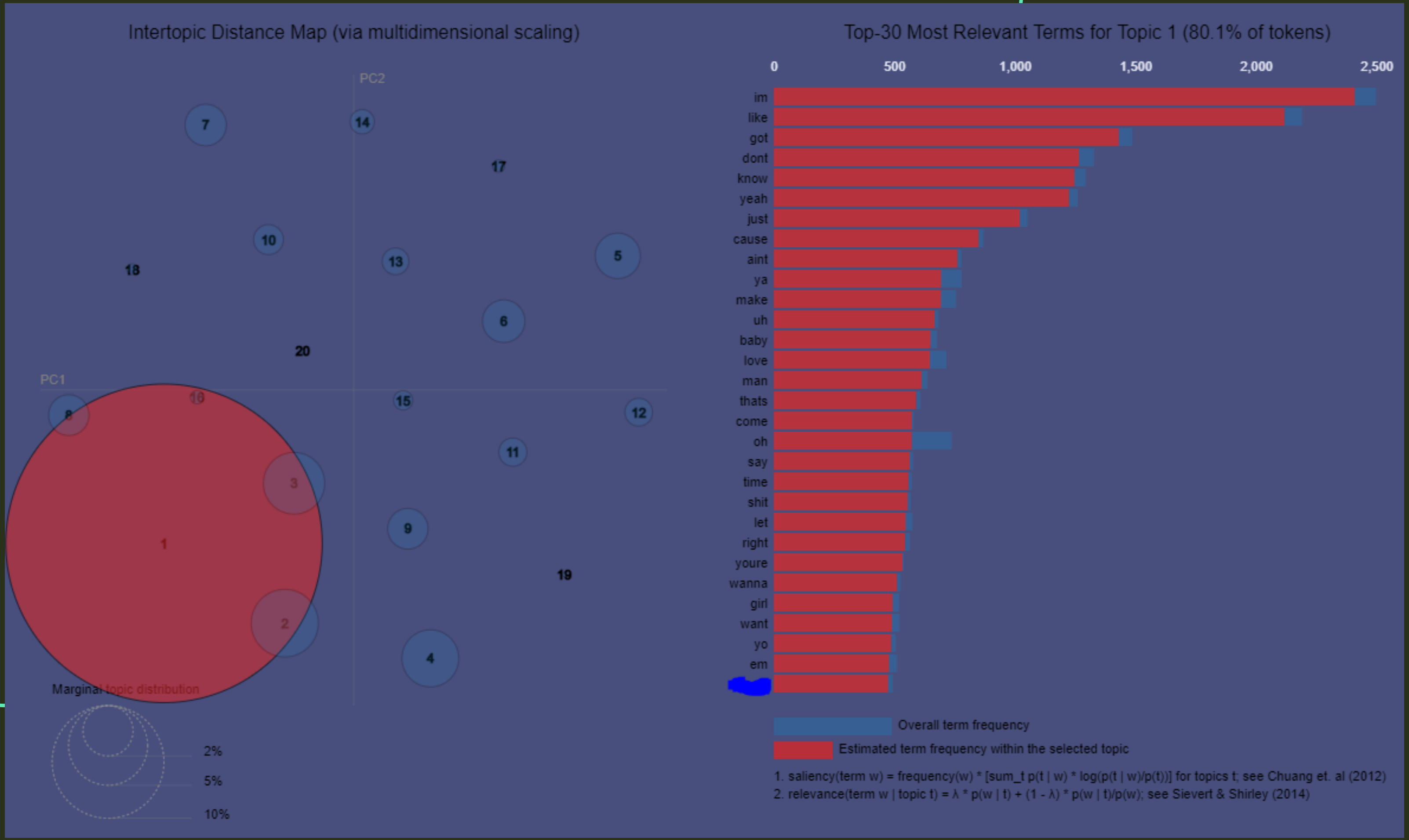
*image from https://lab.getapp.com/importance-of-data-cleaning-and-governance/

Topic Modeling and Clustering

- Clustering via KMeans to inspect natural patterns across the lyrics
  - Used entire lyrics dictionary to build 10 clusters to look for elbow point
  - Depending on interpretation, 3 or 5 clusters is ideal.
    - Using regional labels, 4 is the ideal number of clusters we expected to represent potentially 4 distinct vocabularies across the US.



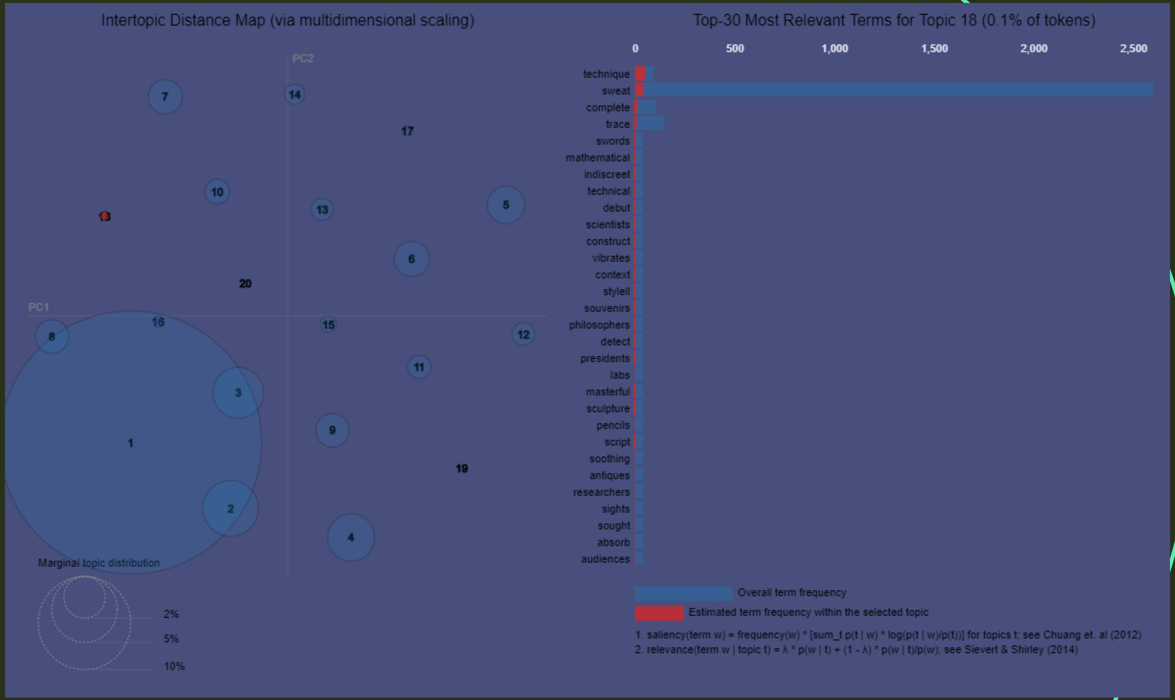
Topic Modeling and Clustering

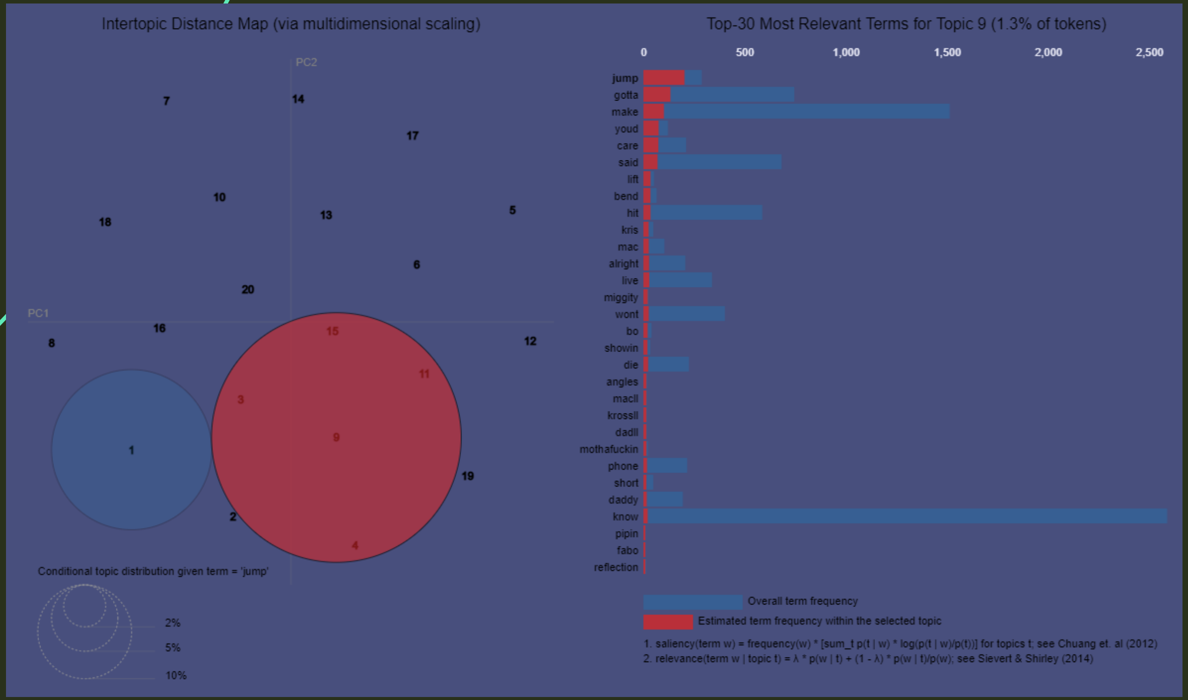


- LDA Topic Modeling
  - Largest topic includes the common "song" words like "love", "baby"

Topic Modeling and Clustering

- LDA Topic Modeling
  - Other topics are extremely specific to artist or song
    - Rakim with larger, more complex words
    - Kriss Kross "Jump" becomes a separate topic





Exploratory Data Analysis

US Regions

World

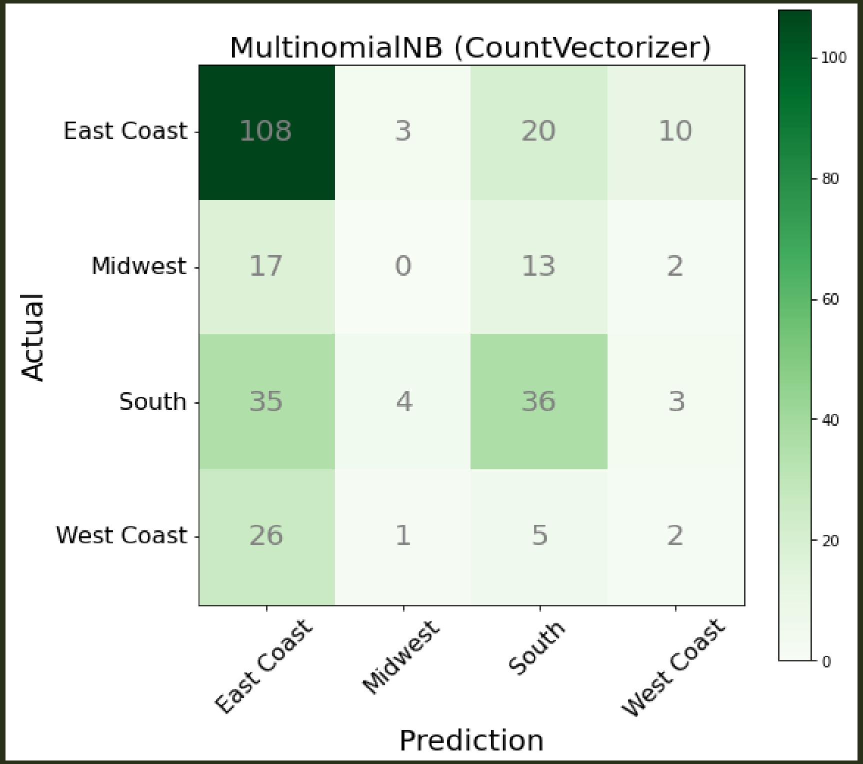
Exploratory Data Analysis

- Across both region and decade, no major visual difference is seen in word frequency
- The lyrics explain the core motivations of the artists:
  - Who/what they are → "I'm"
  - What they have → "Got"
  - What they know and want the listener to know → "Know"

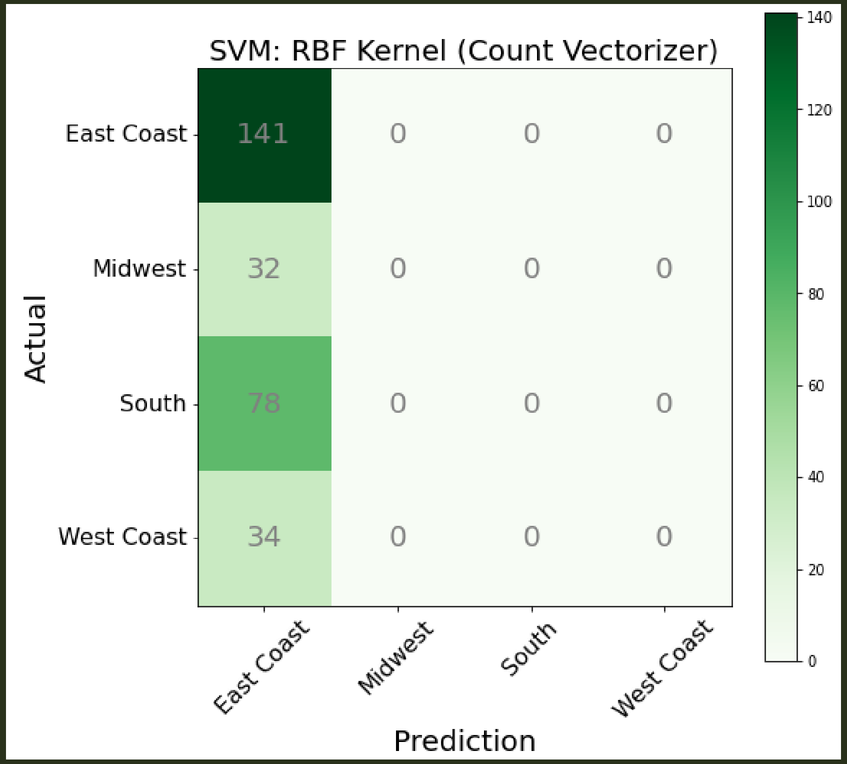
Model Data (Regions)

- Started modeling with the vectorized dataset.
- Results heavily skewed due to imbalanced data.

Region	Song Count
East Coast	141
South	78
West Coast	34
Midwest	32



Accuracy: 51.23%



Accuracy: 49.47%

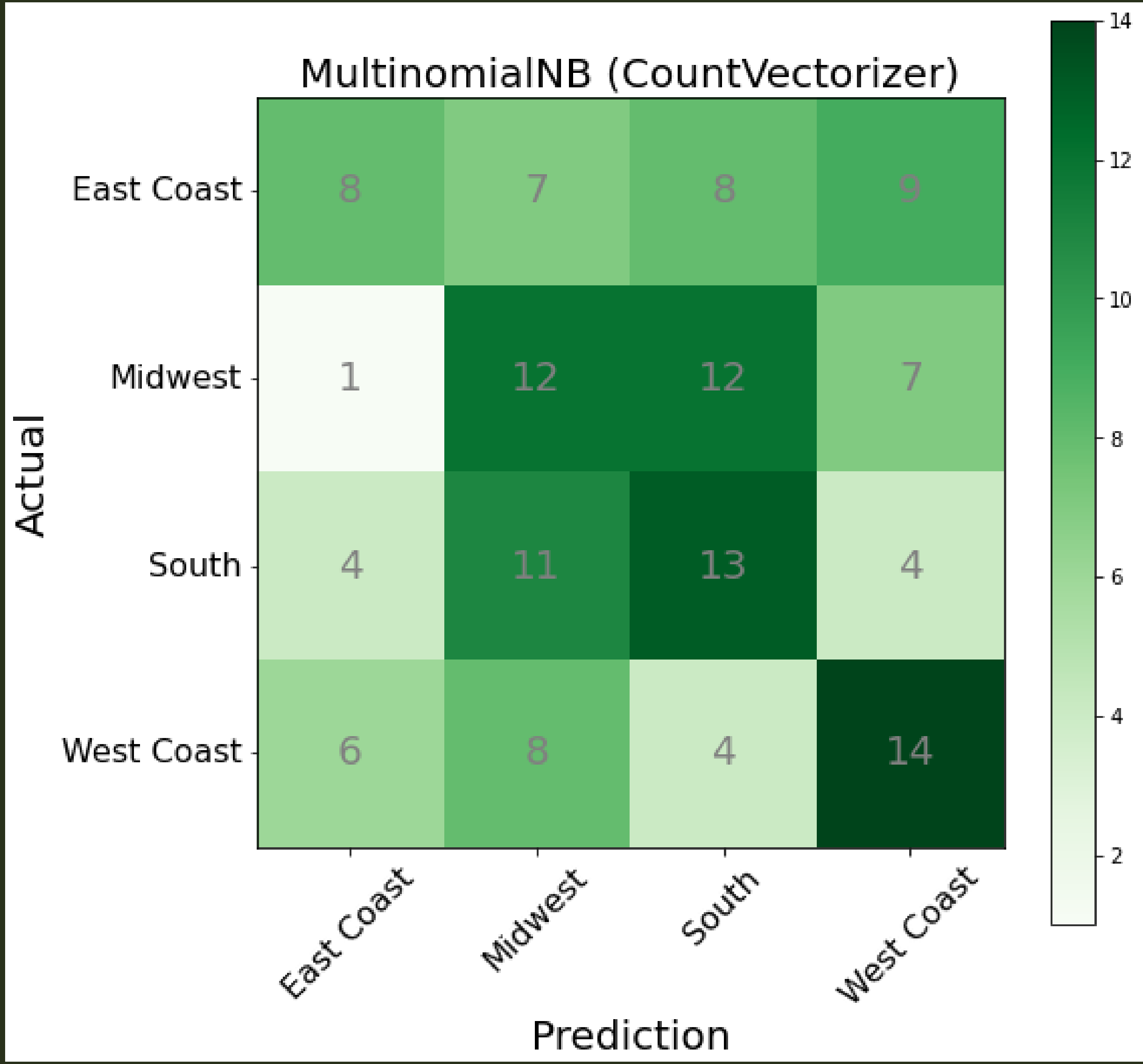
Model Data (Regions)

- Solution: Resample the dataset to equally represent each region.
- Under-sample each region to match the lowest count.

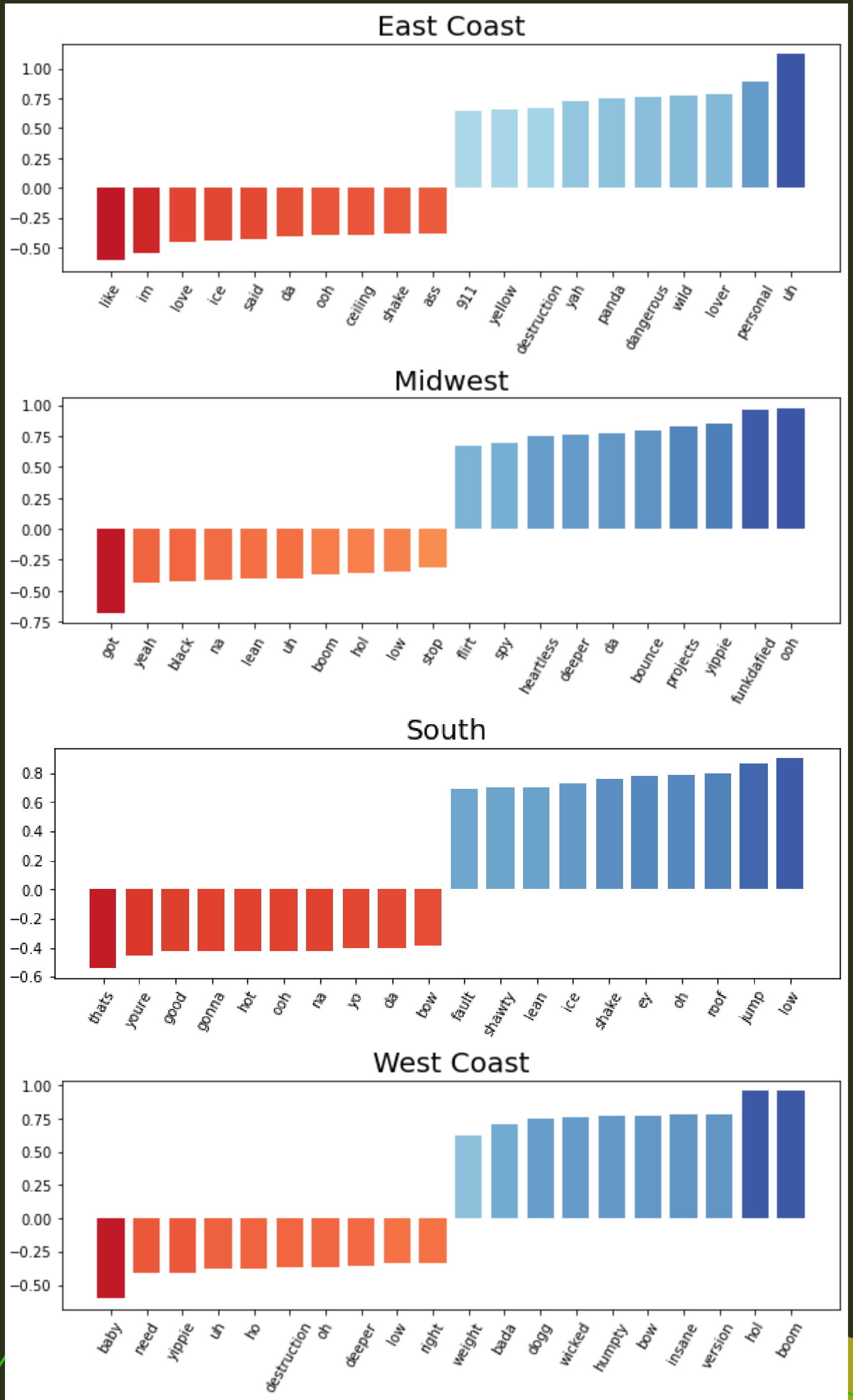
Cross-Validation Accuracy Scores for Different Classifiers	
Model	**5-Fold CV Accuracy**
Bernoulli Naïve Bayes	26.56%
Multinomial NB (Count Vectorizer)	36.72%
Multinomial NB (TF-IDF Vectorizer)	22.66%
SVM: Linear Kernel (Count Vectorizer)	34.38%
SVM: Linear Kernel (TF-IDF Vectorizer)	35.16%
SVM: RBF Kernel (Count Vectorizer)	17.19%
SVM: RBF Kernel (TF-IDF Vectorizer)	14.84%
SVM: Polynomial Kernel (Count Vectorizer)	13.28%
SVM: Polynomial Kernel (TF-IDF Vectorizer)	14.06%

Region	Song Count
East Coast	32
South	32
West Coast	32
Midwest	32

Model Data (Regions)



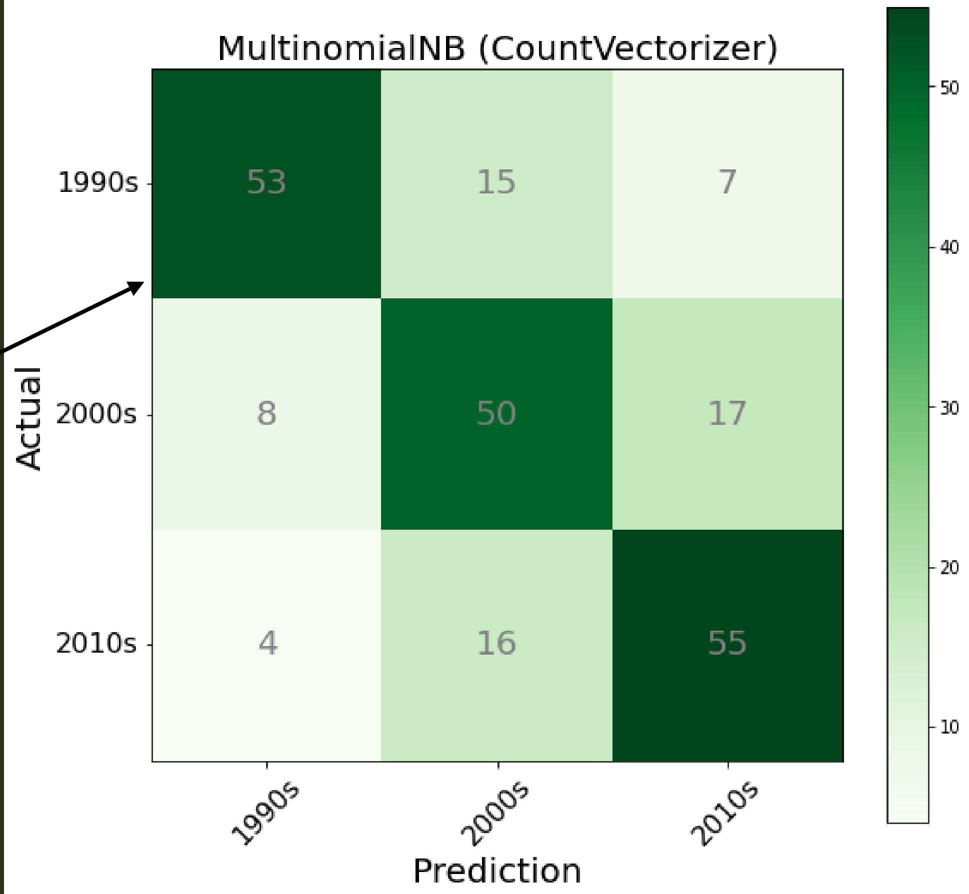
Accuracy: 36.72%

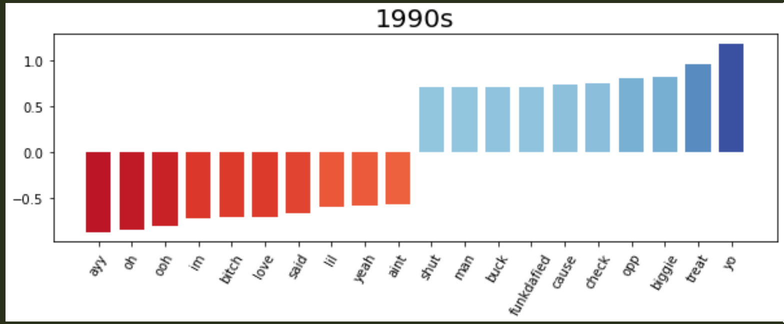


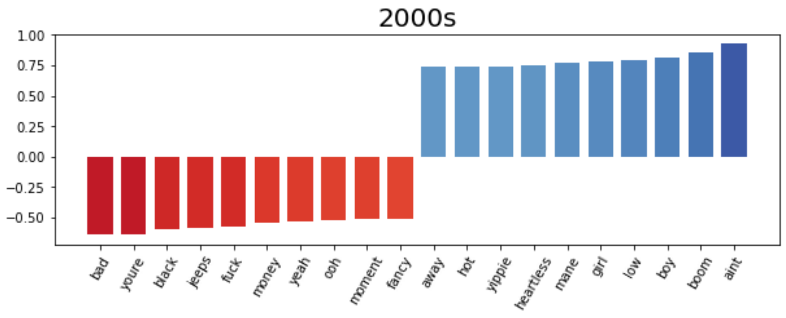
Model Data (Decades)

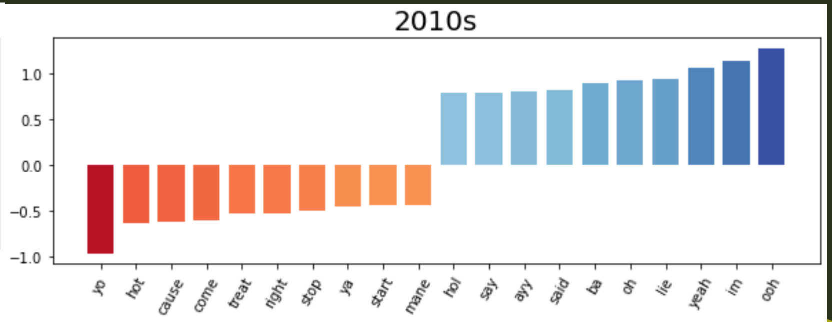
- Repeated process with decade labels

Model	5-Fold CV Accuracy
Bernoulli Naïve Bayes	56.89%
Multinomial NB (Count Vectorizer)	70.22%
Multinomial NB (TF-IDF Vectorizer)	49.33%
SVM: Linear Kernel (Count Vectorizer)	58.22%
SVM: Linear Kernel (TF-IDF Vectorizer)	66.22%
SVM: RBF Kernel (Count Vectorizer)	42.22%
SVM: RBF Kernel (TF-IDF Vectorizer)	37.78%
SVM: Polynomial Kernel (Count Vectorizer)	21.78%
SVM: Polynomial Kernel (TF-IDF Vectorizer)	27.11%









Interpret Results

- The lyrical vocabulary of hip-hop songs between regions does not differ enough to create an accurate prediction model.
- The vocabulary choice between different decades has enough differences to distinguish between the time periods.

1989 **2019**

Interpret Results

- Possible reasons that lyric data could not accurately model the region:
  1) **There is not much difference in word choice across the country**.
  2) **Selection Bias**:
     - Songs that hit *Billboard* number-one are typically nationally known artists.
     - These artists may choose to use vocabulary that will not isolate them from particular regions.
  3) **Insufficient Cleaning**:
     - Dimensionality was mostly reduced by removing stop words from *sci-kit learn*.
     - Subject-matter-expert could help change the stop word vocabulary to improve model accuracy.

Thank You for Listening!