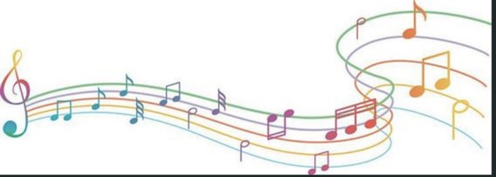


Artist Classification Based on Song Lyrics

Content

Part 1: Introduction and Data Summary

That's what people say,
But I keep cruisin'
Can't stop, won't stop movin'
It's like I got this music in my mind
Sayin' it's gonna be alright
'Cause the players gonna play,
play, play, play, play
And the haters gonna hate, hate,
hate, hate, hate
Baby, I'm just gonna shake, shake,
shake, shake, shake
I shake it off, I shake it off

“I used to rule the world
Seas would rise when I gave the word
Now in the morning, I sleep alone
Sweep the streets I used to own

I used to roll the dice
Feel the fear in my enemies' eyes
Listen as the crowd would sing
"Now the old king is dead, long live the king"
One minute, I held the key
Next, the walls were closed on me

Taylor Swift
- Shake It Off

Coldplay
- Viva La Vida

Task: Build Classification Models

1. Text mining (Lyrics to Vectors)

- Bag-of-Words/Term-doc count matrix
- IF-IDF Weight Matrix

	Lyric 1	Lyric 2	Lyric 3
	Antony and Cleopatra	Julius Caesar	The Tempest
Antony	5.25	3.18	0
Brutus	1.21	6.1	0
Caesar	8.59	2.54	0
Calpurnia	0	1.54	0
Cleopatra	2.85	0	0
mercy	1.51	0	1.9
worser	1.37	0	0.11

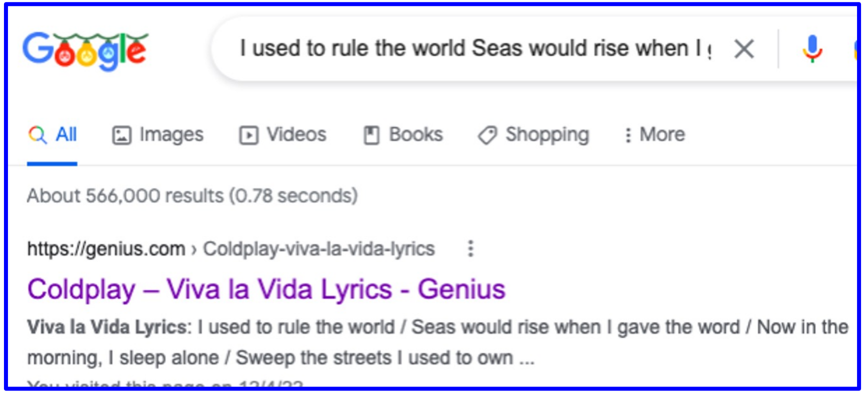
2. ML Prediction Models

- Naive Bayes (NB)
- Support Vector Machine (SVM)

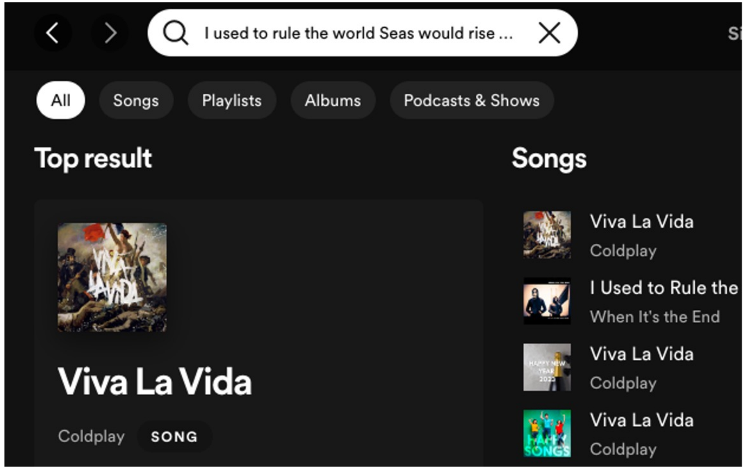
Vector ➡ **Artist name**

Applications

- **Search engines**
  e.g. Google



- **Music streaming platforms**
  e.g. Spotify



Data Summary

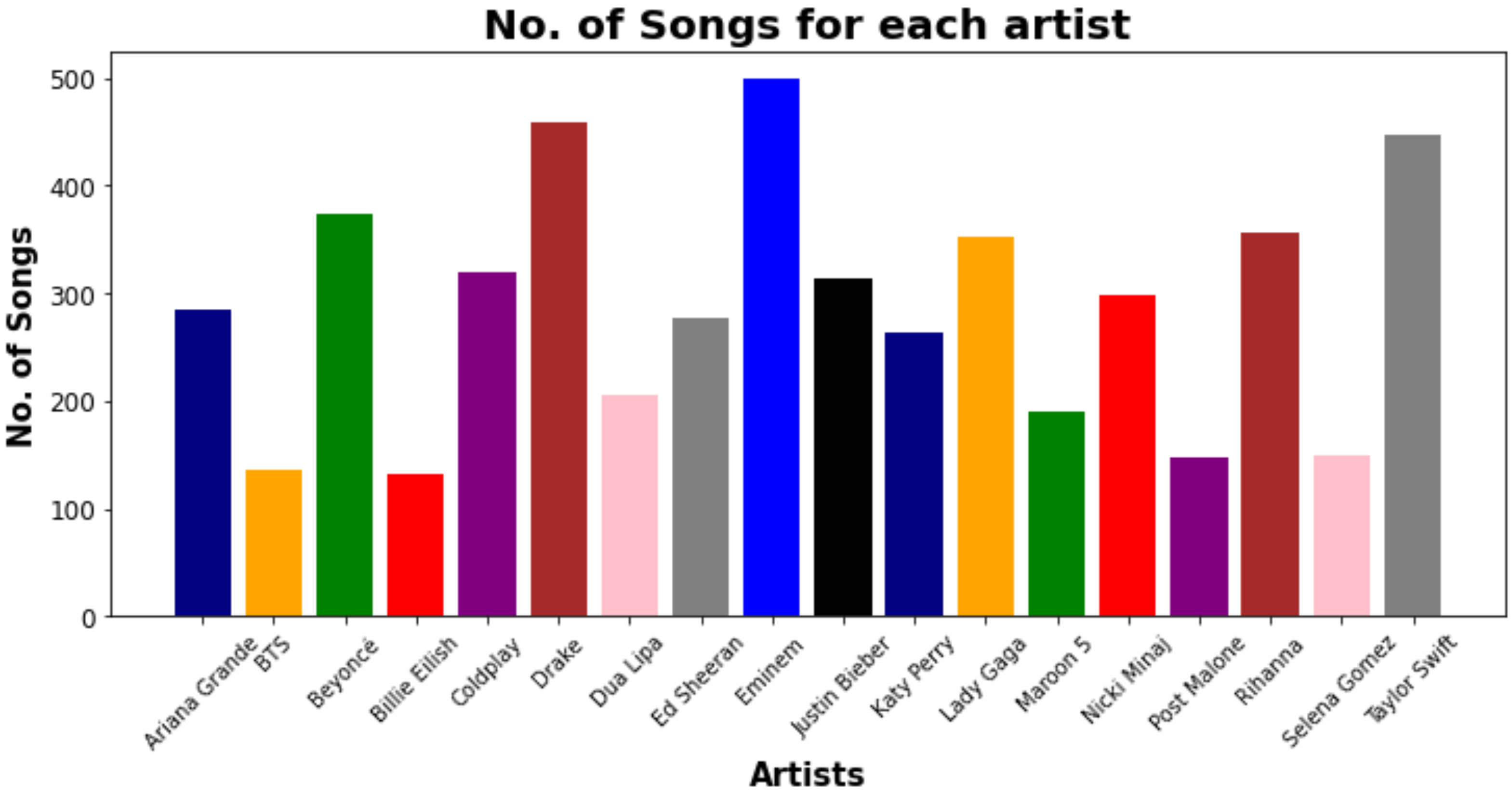
Description	Combined dataset from different files
Variables	• Index • **Artist** • Title • Album • Year • Date • **Lyric**
Original Size	6027 * 7 (21 artists)
Size after cleaning	5203 * 7 (18 artists)

Average: 289 songs/artist

Training set (70%)

Testing set (30%)

Data Summary



Part 2: Data Preprocessing

Artists

Select artists: number of songs > 100

Remove:

Khalid (64 songs)
CardiB (75 songs)
CharliePuth (75 songs)

Lyrics

- **Songs with no lyrics (NaN)**
- **Songs with duplicated lyrics**

```
Out[130]:
      Artist                                        Title  \
277   Coldplay                Viva La Vida (Thin White Duke mix)
334   Coldplay  Viva La Vida (Grant's Uplifting original mix)

                                               Lyric
277   i used to rule the world seas would rise when ...
334   i used to rule the world seas would rise when ...
```

- **Songs with non-English lyrics**

(Lady Gaga)

```
Out[160]: "des yeux qui font baisser les miens un rire qui se perd
le portrait sans retouches de l'homme auquel j'appartiens  quand il
bras il me parle tout bas je vois la vie en rose il me dit des mots
de tous les jours et ça m'fait quelque chose  il est entré dans mon
bonheur dont je connais la cause c'est lui pour moi moi pour lui da
dit l'a juré pour la vie  et dès que je l'aperçois alors je sens en
```

(Taylor Swift)

```
Out[156]: 'zwrotka  siedzę i patrzę jak czytasz z głową pochyloną bu
patrzę jakl oddychasz z zamkniętymi oczyma siedzę i oglądam ciebie z
wszystko co robisz i czego nie robisz jesteś tyle starszy i mądrzejs
refren  czekam przy drzwiach jak małe dziecko używam najlepsze farby
portret nakrywam stół wykwintnymi pierdołami i patrzę jak ty to jedy
znosisz jeśli to wszystko dzieje się w mojej głowie to powiedz mi te
```

Tokenization
Case Folding

Already applied

```
Out[163]: "i used to rule the world seas would rise when
word now in the morning i sleep alone sweep the streets
i used to roll the dice feel the fear in my enemy's eyes
the crowd would sing now the old king is dead long live
minute i held the key next the walls were closed on me a
discovered that my castles stand upon pillars of salt an
```

Stop words

e.g. the, a, and, to, be.

- Remove or not? Tried both

Stemming

Reduce terms to their "roots"
e.g. accepted - accept

- Applied or not? Tried both
- Used Snowball stemmer

Part 3: Word Embedding

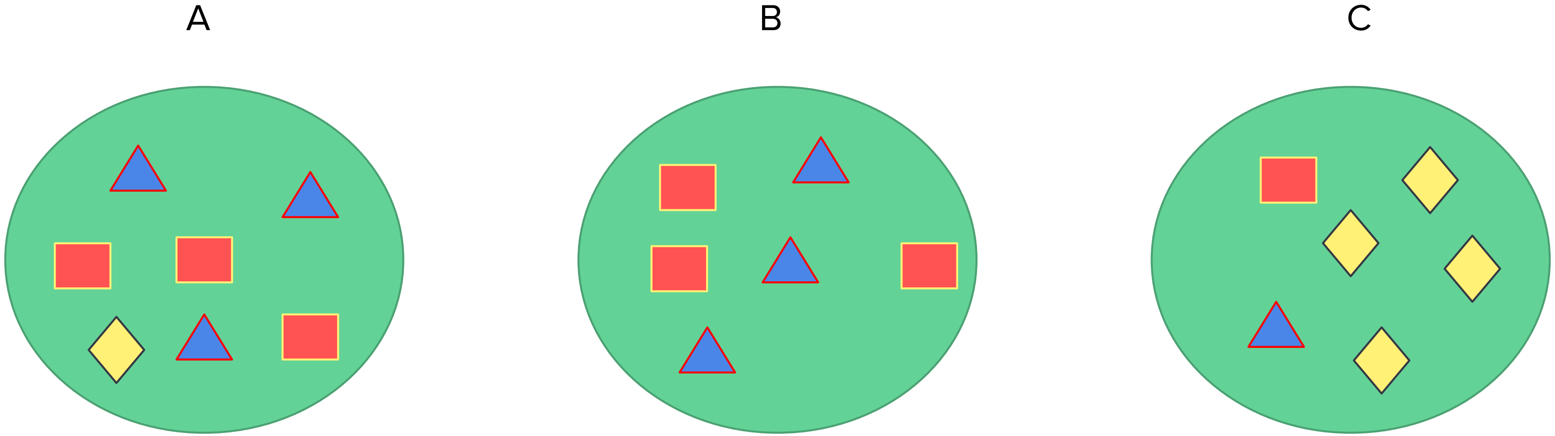
Bag-of-Words

Intuition:

- Documents with similar content are similar
- Measures vocabulary and strength of presence
- Limitation: Does not consider order of words

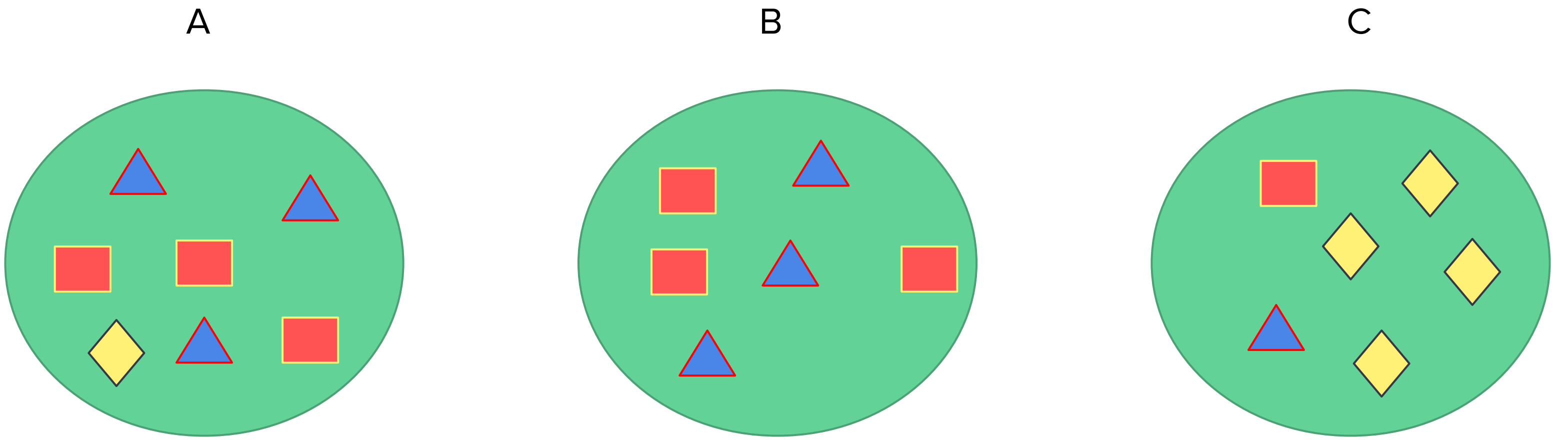
Bag-of-Words

Which documents are the most similar?



Bag-of-Words

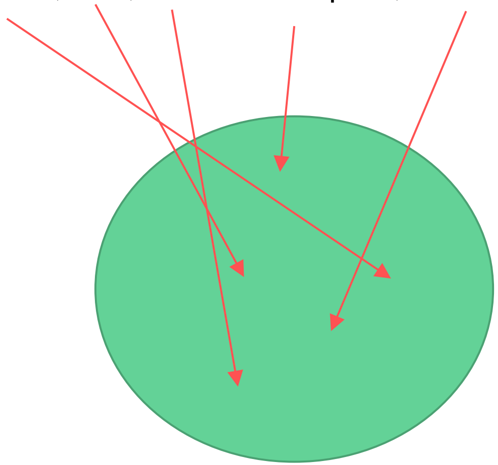
Documents A and B both have 3 squares and 3 triangles



Bag-of-Words

Why is it called Bag-of-Words?

[“this”, “is”, “an” “example”, “sentence”]



Bag-of-Words Example

Consider the following three “documents”:

[“max”, “is”, “from”, “wisconsin”]

[“michael”, “studies”, “at”, “university”, “georgia”]

[“yan”, “lives”, “in”, “georgia”]

Bag-of-Words Example

Consider the following three “documents”:

[“max”, “is”, “from”, “wisconsin”]

[“michael”, “studies”, “at”, “university”, “georgia”]

[“yan”, “lives”, “in”, “georgia”]

Vocabulary:

[“max”, “is”, “from”, “wisconsin”, “michael”, “studies”, “at”, “university”, “georgia”, “yan”, “lives”, “in”]

Count Matrix:

Vocab	D1	D2	D3
max	1	0	0
is	1	0	0
from	1	0	0
wisconsin	1	0	0
michael	0	1	0
studies	0	1	0
at	0	1	0
university	0	1	0
georgia	0	1	1
yan	0	0	1
lives	0	0	1
in	0	0	1

TF-IDF

Intuition:

- Stands for “term frequency-inverse document frequency
- Penalizes common vocabulary

Ex: “I”, “love”, “you” in love songs

TF-IDF

IDF:

- df_t is the document frequency of word t

- Inverse document frequency:

$$\mathrm{idf}_t = \log_{10}(N/\mathrm{df}_t)$$

TF-IDF Example

Document A	
Term	**Count**
this	1
is	1
a	2
sample	1

Document B	
Term	**Count**
this	1
is	1
another	2
example	3

$$\mathrm{tf}_{\text{“this”},d1} = \frac{1}{5} \qquad \mathrm{tf}_{\text{“this”},d2} = \frac{1}{7}$$

$$\mathrm{idf}_{\text{“this”},D} = \log \frac{2}{2} = 0$$

$$\text{tf-idf}_{\text{“this”},d1,D} = 0.2 \times \log \frac{2}{2} = 0$$

$$\text{tf-idf}_{\text{“this”},d2,D} = 0.14 \times \log \frac{2}{2} = 0$$

TF-IDF Example

Document 1	
Term	**Count**
this	1
is	1
a	2
sample	1

Document 2	
Term	**Count**
this	1
is	1
another	2
example	3

$$\mathrm{tf}_{\text{“example”},d1} = \frac{0}{5} \quad \mathrm{tf}_{\text{“example”},d2} = \frac{3}{7}$$

$$\mathrm{idf}_{\text{“example”},D} = \log\frac{2}{1} = 0.301$$

$$\text{tf-idf}_{\text{“example”},d1,D} = 0 \times \log\frac{2}{1} = 0$$

$$\text{tf-idf}_{\text{“example”},d2,D} = 0.429 \times \log\frac{2}{1} = 0.129$$

Part 4: Predictive Modeling

Naive Bayes Classifier

Bayes Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$\propto P(B|A)P(A)$$

We call this formulation the posterior, which is the product of the sampling likelihood and the prior distribution

Naive Bayes Classifier

Consider with label y and p features: x1, ..., xp:

$$P(y|x_1, \ldots, x_p) \propto P(x_1, \ldots, x_p|y)P(y)$$

The “naive” assumption:

$$P(y|x_1, \ldots, x_p) \propto P(x_1|y) \ldots P(x_p|y)P(y)$$

Naive Bayes Classifier

Consider $y = \{1, \ldots, k\}$

Then the NBC model is:

$$\hat{y} = \underset{j \in 1,\ldots,k}{\arg\max} P(y_j) \prod_{i=1}^{p} P(x_i | y_j)$$

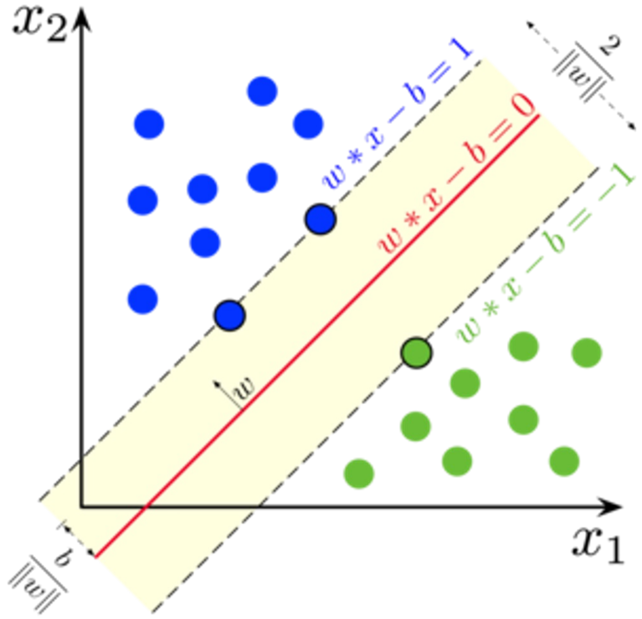
Naive Bayes Classifier

Why use this model?

- Calculation of probabilities is easy and intuitive given term frequence features

Support Vector Machine (SVM)

- Attempts to find the hyperplane with largest margin between two classes
- Uses support vectors to define the margin



SVM

Loss function: (hinge loss)

$$\lambda||\boldsymbol{w}||^2 + \left[\frac{1}{n}\sum_{i=1}^{n}\max(0, 1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i - b))\right]$$

Updated using gradient descent

SVM

Why use this model?

- Computationally efficient
- Also works well with text classification

Part 5: Results

Effect on Accuracy of Removing Stop Words

Naive Bayes

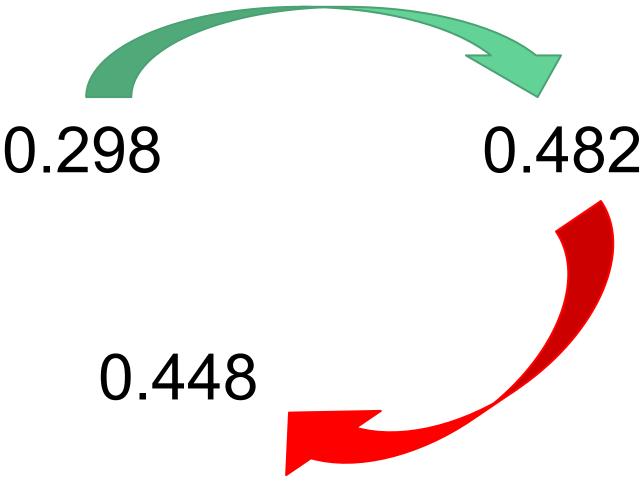
0.298 0.482

Support Vector Machine

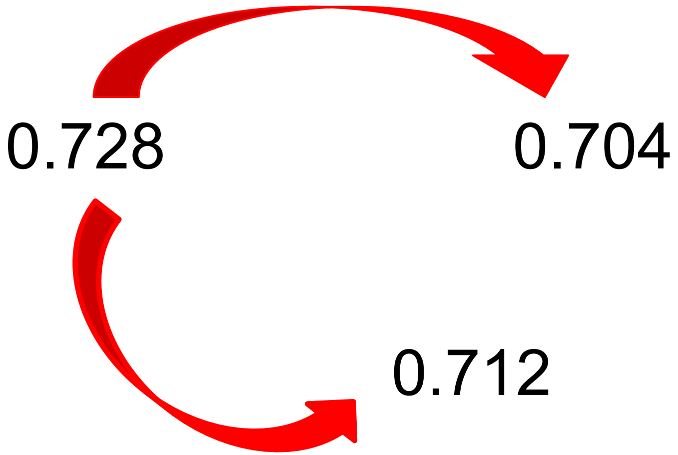
0.728 0.704

Effect on Accuracy of Stemming

Naive Bayes



Support Vector Machine



Accuracy of Best Model

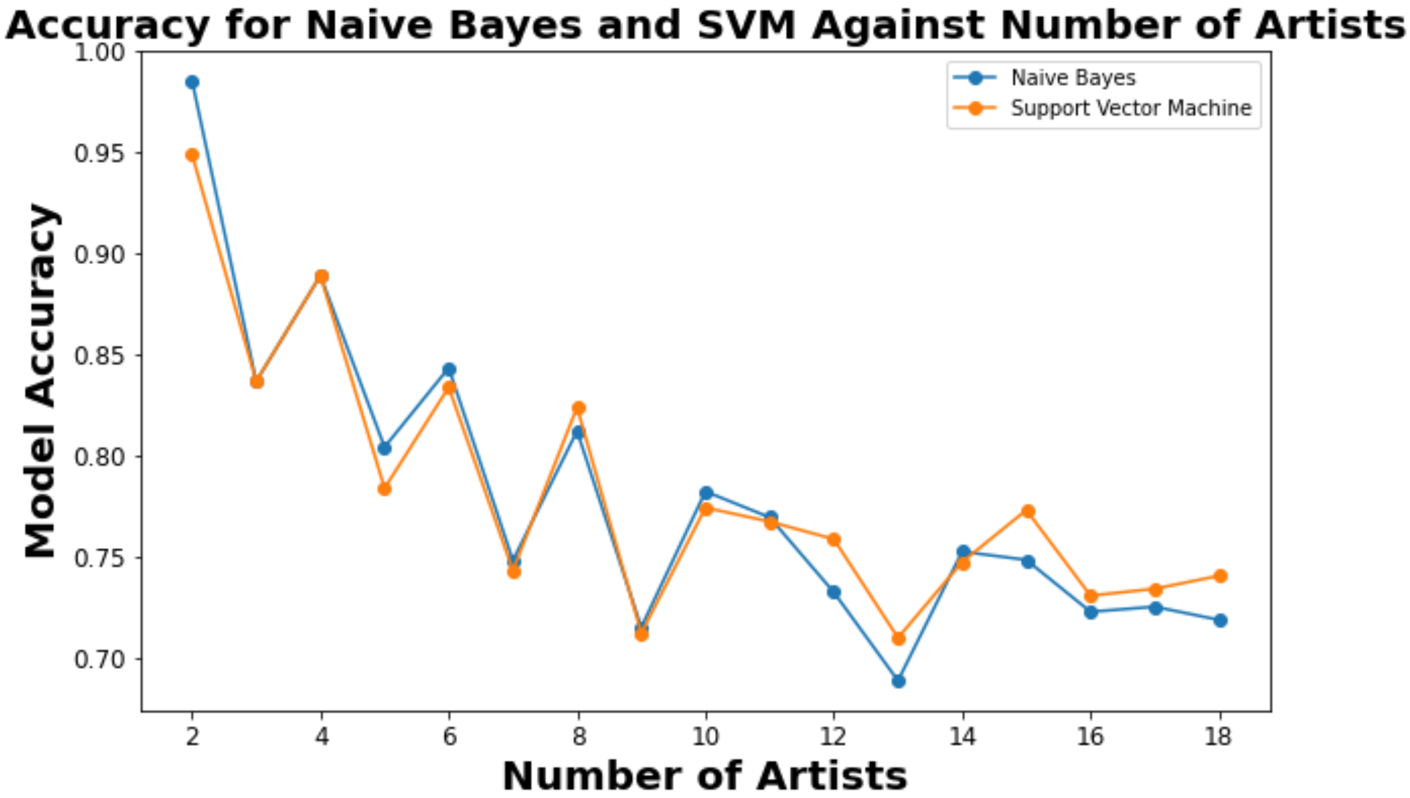
Naive Bayes: Remove stop words. Don’t perform stemming.

Accuracy: 0.729

SVM: Keep stop words. Don’t perform stemming.

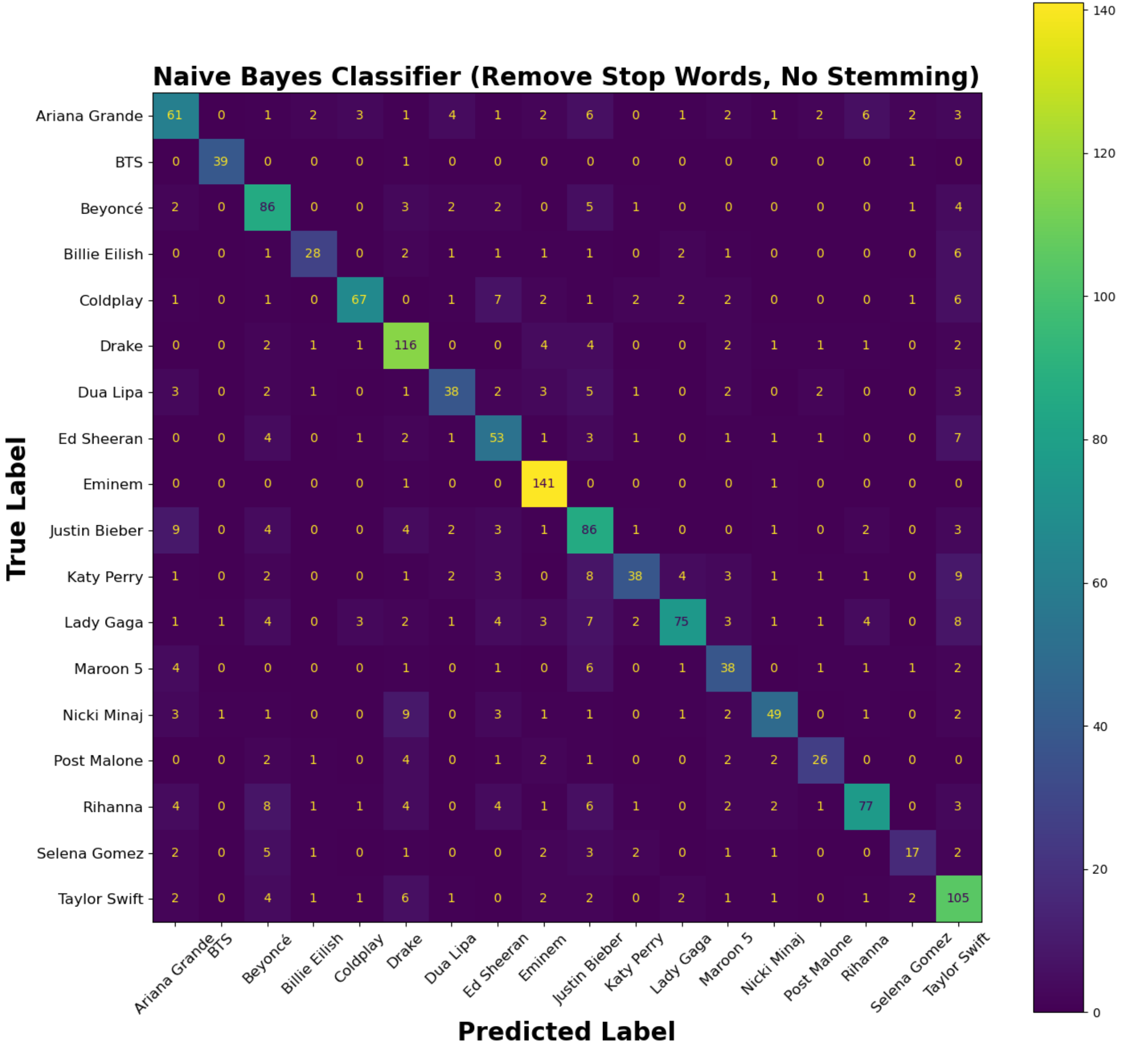
Accuracy: 0.728

Effect of Adding More Artists

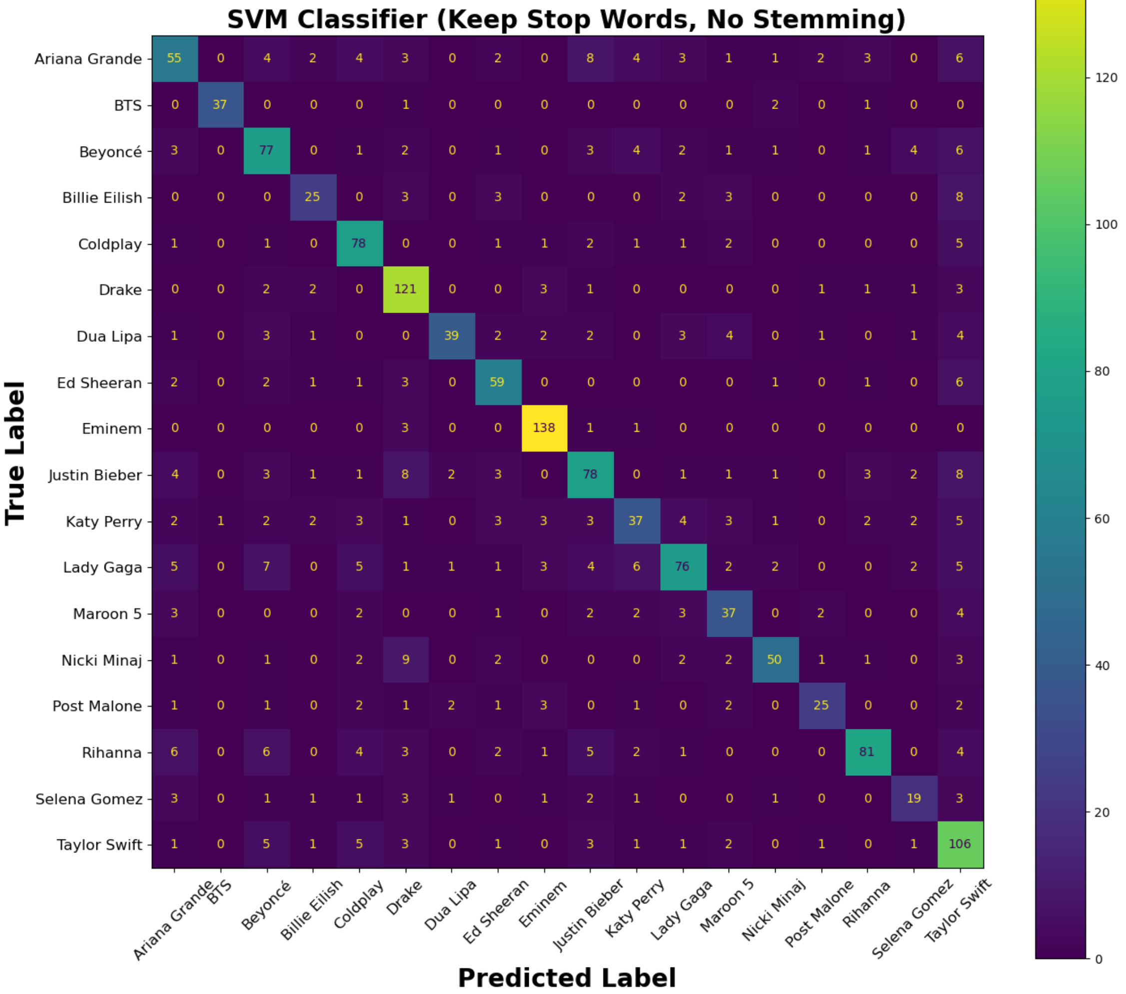


Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP



Multilabel Confusion Matrix for Naive Bayes Classifier



Multilabel Confusion Matrix for SVM Classifier

Artist with Best Precision (Naive Bayes)

BTS	0.95
Coldplay	0.87
Lady Gaga	0.85
Eminem	0.85
Rihanna	0.82

Artist with Best Recall (Naive Bayes)

Eminem	0.99
BTS	0.95
Drake	0.86
Beyoncé	0.81
Taylor Swift	0.80

Artist with Best f1-score (Naive Bayes)

BTS	0.95
Eminem	0.91
Coldplay	0.79
Drake	0.79
Beyoncé	0.74

Artist with Best Precision (SVM)

BTS	0.97
Eminem	0.89
Dua Lipa	0.87
Rihanna	0.86
Lady Gaga	0.77

Artist with Best Recall (SVM)

Eminem	0.97
Drake	0.90
BTS	0.90
Coldplay	0.84
Taylor Swift	0.81

Artist with Best f1-score (SVM)

BTS	0.94
Eminem	0.93
Drake	0.81
Rihanna	0.78
Coldplay	0.77

Performance of Rank Ordering (NB)

Query: "I have this thing where I get older"

Taylor Swift	(0.266)
Selena Gomez	(0.150)
Drake	(0.104)
Eminem	(0.086)
Ed Sheeran	(0.082)

Performance of Rank Ordering

Accuracy of Model Matching Top k Artist

	1	2	3	4	5
NB	0.729	0.793	0.823	0.852	0.875
SVM	0.717	0.789	0.822	0.850	0.872

List of All Artist

Ariana Grande

BTS

Beyoncé

Billie Eilish

Coldplay

Drake

Dua Lipa

Ed Sheeran

Eminem

Justin Bieber

Katy Perry

Lady Gaga

Maroon 5

Nicki Minaj

Post Malone

Rihanna

Selena Gomez

Taylor Swift