Thematically Sequenced Playlist

Georgetown Analytics
Cohort 19

By: Nick Merkling, Adam Goldstein, Patricia Merino, and Navneet Sandhu

Introduction

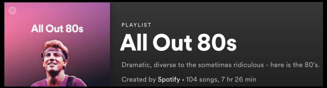
- 286 Million Monthly active Spotify users (130 million premium users)
- 36% of global streaming market
- Average of 25 listening hours a month per user
- Over 50 million tracks available on Spotify (Iqbar)

PLAYLIST

70s Road Trip

Hit the road to the soundtrack of these seventies classics.

Created by Spotify • 100 songs, 7 hr 10 min

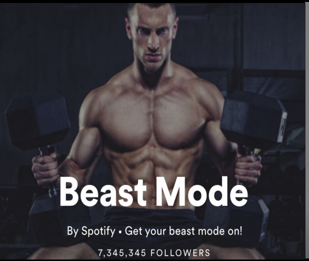


PLAYLIST

All Out 80s

Dramatic, diverse to the sometimes ridiculous - here is the 80's.

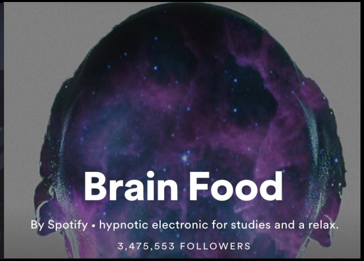
Created by Spotify • 104 songs, 7 hr 26 min



Beast Mode

By Spotify • Get your beast mode on!

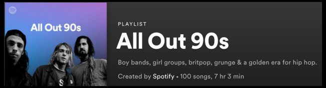
7,345,345 FOLLOWERS



Brain Food

By Spotify • hypnotic electronic for studies and a relax.

3,475,553 FOLLOWERS

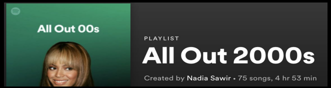


PLAYLIST

All Out 90s

Boy bands, girl groups, britpop, grunge & a golden era for hip hop.

Created by Spotify • 100 songs, 7 hr 3 min



PLAYLIST

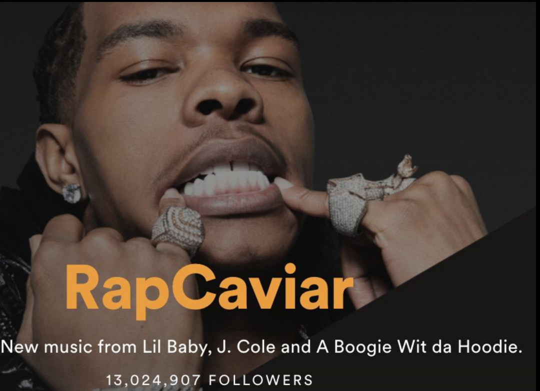
All Out 2000s

Created by Nadia Sawir • 75 songs, 4 hr 53 min

POLLEN

By Spotify • Genre-less. Quality first always. Cover: Noname

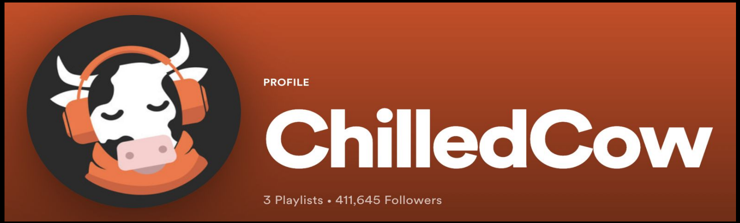
1,099,615 FOLLOWERS



RapCaviar

New music from Lil Baby, J. Cole and A Boogie Wit da Hoodie.

13,024,907 FOLLOWERS



PROFILE

ChilledCow

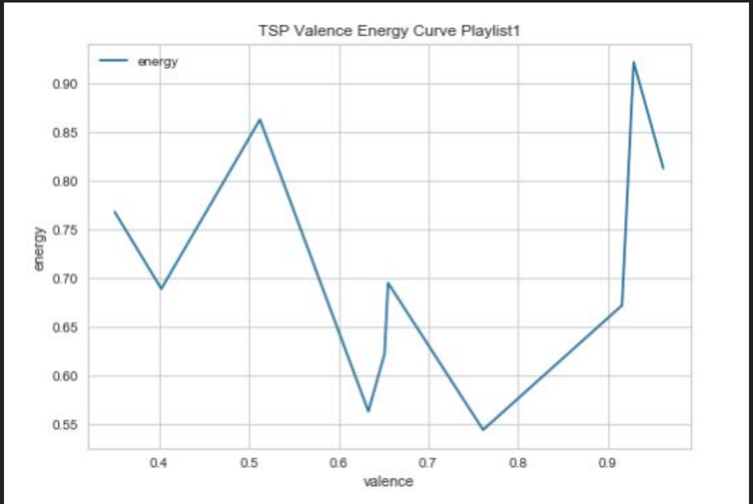
3 Playlists • 411,645 Followers

Spotify Audio Features

1. Valence - Music Positivity or Reflectiveness
2. Energy - Intensity or Activity
3. Danceability - How Suitable for Dancing
4. Acousticness - How confident an acoustic instrument is present
5. Liveness - Detects presence of an audience
6. Speechiness - Detects presence of speech
7. Instrumentalness - Predicts whether a track has no vocals or vocals
8. Key - Scale the track is played in
9. Mode - Indicates major or minor scale
10. Tempo - Beats Per Minute
11. Time Signature - How many beats per minute

Hypothesis

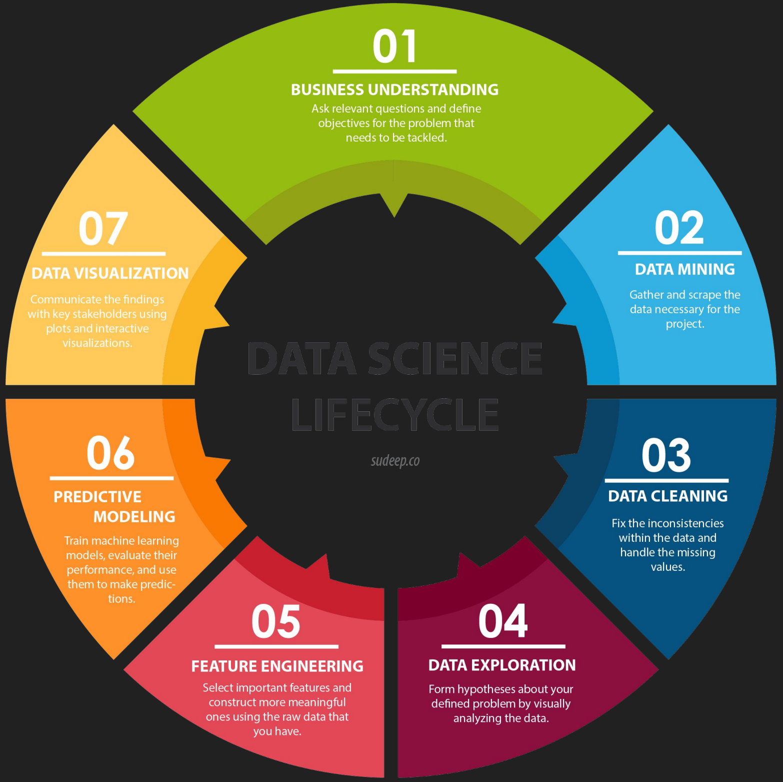
- We believe that creating playlists driven by lyrical content can give the user a glimpse into thematic sequences that exist within their “liked” tracks.
- Additionally, we believe that the sonic variation of a playlist, ordered according to a valence-energy curve, provides the user with a captivating listening experience.



Our Data Science Pipeline

Tools that we used:

- Spotify API
- AWS S3 Storage
- Python 3
  - Jupyter Notebook
  - Yellowbrick
  - NLTK



DATA SCIENCE LIFECYCLE

sudeep.co

01 BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.

02 DATA MINING
Gather and scrape the data necessary for the project.

03 DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.

04 DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.

05 FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.

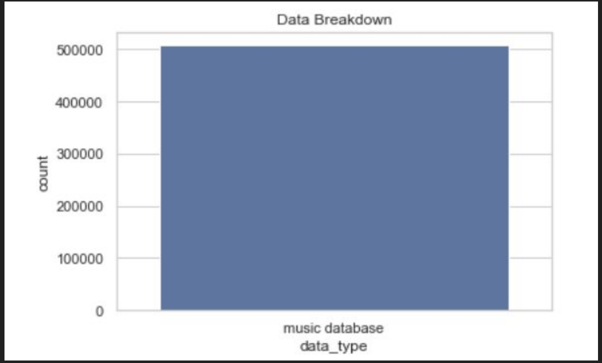
06 PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predictions.

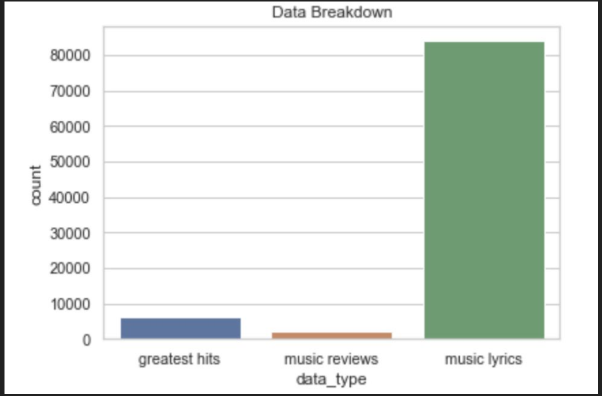
07 DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.

Image Source: Sudeep Agarwal - http://sudeep.co/data-science/Understanding-the-Data-Science-Lifecycle/

Ingestion Phase

- Retrieved data from 4 different sources including:
  - Personal Music database
  - Kaggle file which includes song lyrics
  - Spotify's Greatest hits per Decade (1960s-2010s)
  - Critically Acclaimed Albums over the last 50 years
- Why?
  - Personal touch, legal acquisition of song lyrics to bypass web scraping methods, notable music based off popular Spotify playlists, and to introduce expert perspectives of target parameters

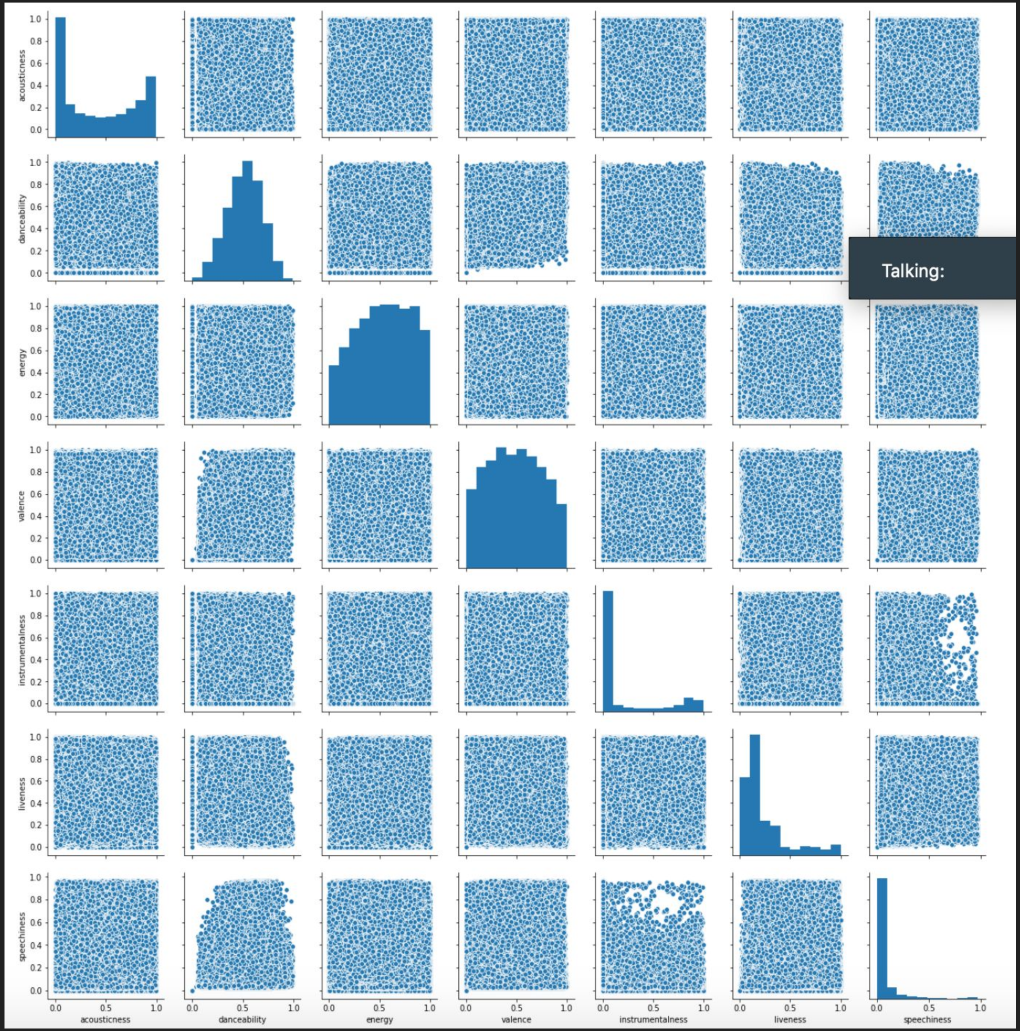




Wrangling Phase

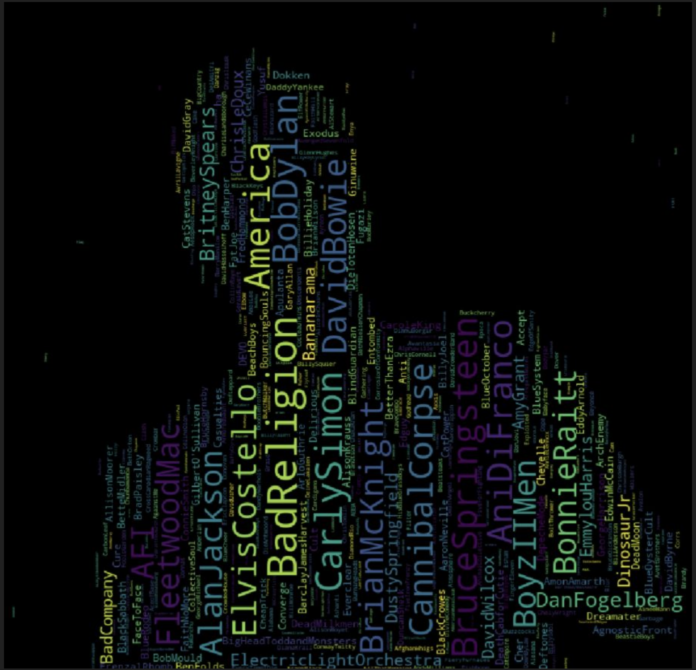
- Navigated Spotify API file’s JSON tree to locate useful data fields
    - Music database, Spotify Greatest Hits playlists, and critically acclaimed albums
- Mapped Kaggle data to track uri from Spotify API
    - Validated the match using Python library, fuzzymatch
    - Set threshold at 96% probability (4% error)
    - Dataset is dramatically reduced using a 1-to-1 match
- 3 phases of lyric cleaning:
    - Spacy Lemmatization, NLTK tokenization, and removal of stop words using NLTK, as well as other .txt file resources

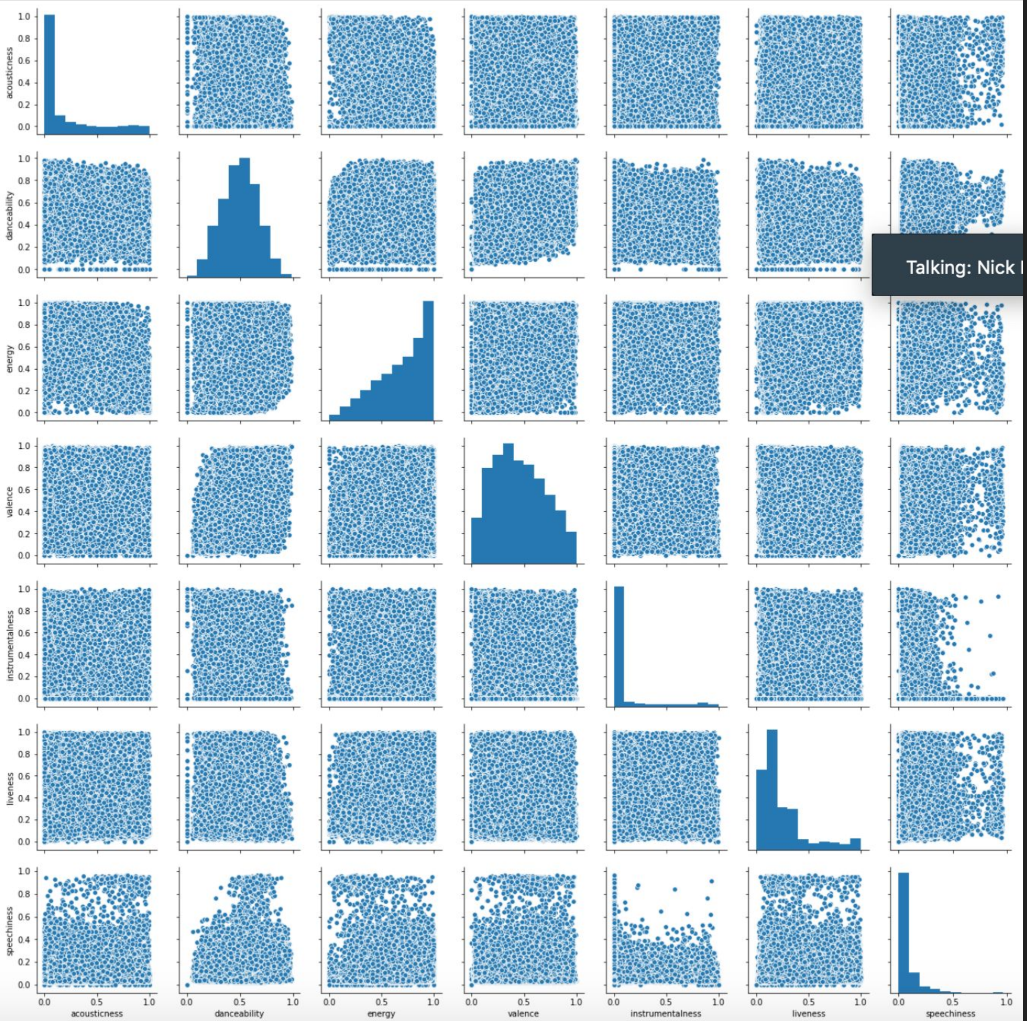
EDA Music Database



Talking:

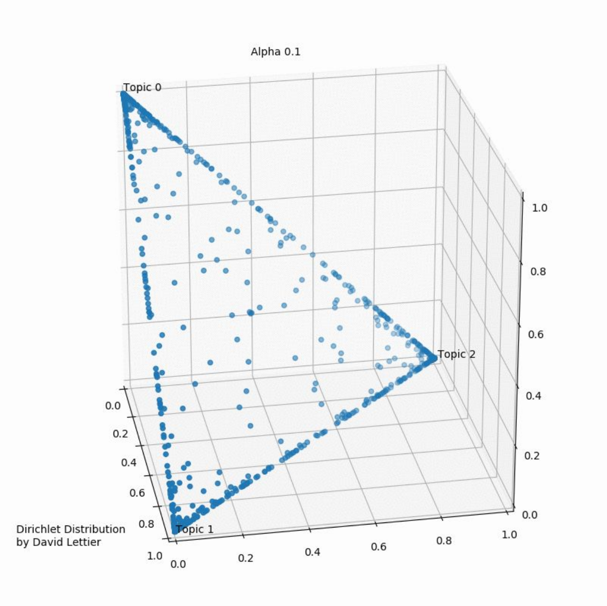
EDA Lyric Dataset





Topic Modeling

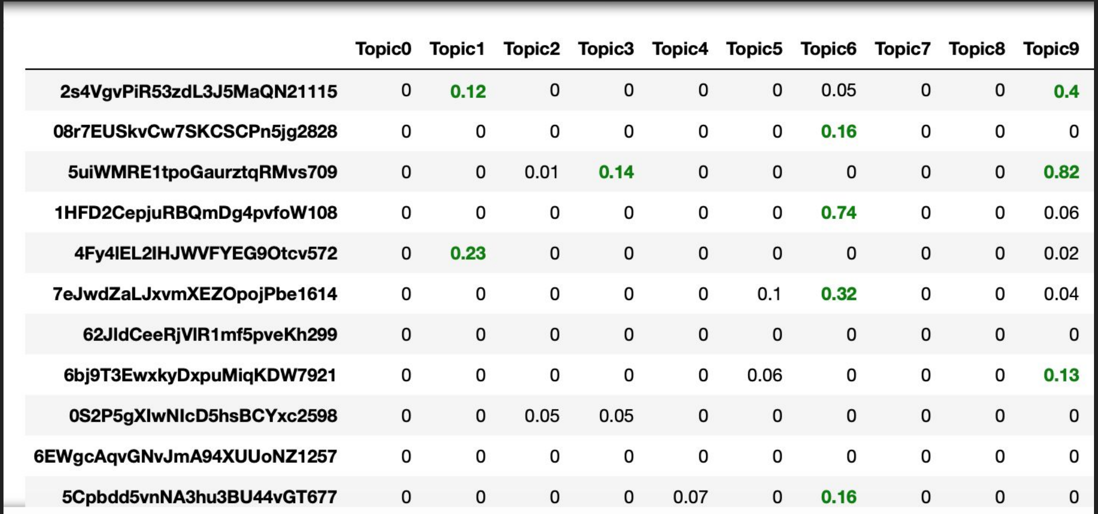
- Sklearn LatentDirichletAllocation/LDA Topic Modeling
- CountVectorizer to convert lyrics to tokens/terms
- Fitted the LDA model on our vectorized data
    - Output n_topics and n_words
- Fine tuned hyperparameters to produce coherent topics
- Determined probability of each document fitting with in a topic



Logic to Creating Playlists Using Topic Analyses

Return all playlist that meet the following criteria:

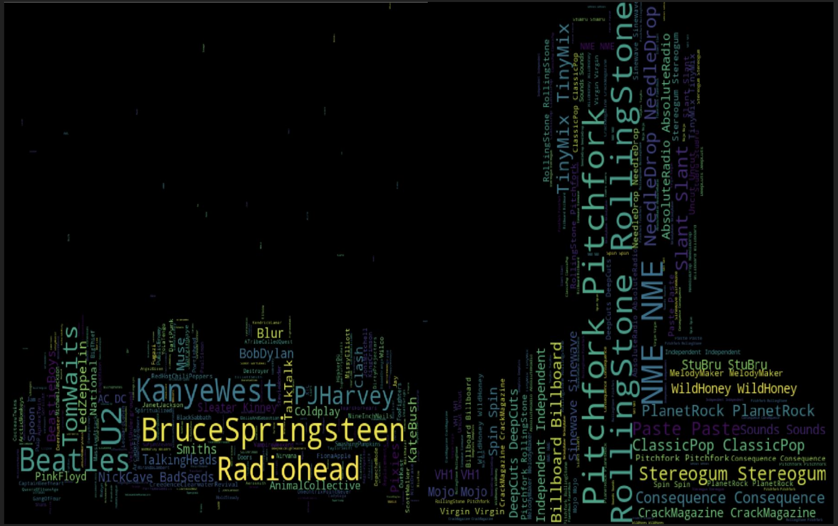
- 10 =< Playlist length =< 20
- Only return track_uris that have a probability of 0.65 or more for a topic.



	Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9
2s4VgvPiR53zdL3J5MaQN21115	0	0.12	0	0	0	0	0.05	0	0	0.4
08r7EUSkvCw7SKCSCPn5jg2828	0	0	0	0	0	0	0.16	0	0	0
5uiWMRE1tpoGaurztqRMvs709	0	0	0.01	0.14	0	0	0	0	0	0.82
1HFD2CepjuRBQmDg4pvfoW108	0	0	0	0	0	0	0.74	0	0	0.06
4Fy4lEL2lHJWVFYEG9Otcv572	0	0.23	0	0	0	0	0	0	0	0.02
7eJwdZaLJxvmXEZOpojPbe1614	0	0	0	0	0	0.1	0.32	0	0	0.04
62JldCeeRjVlR1mf5pveKh299	0	0	0	0	0	0	0	0	0	0
6bj9T3EwxkyDxpuMiqKDW7921	0	0	0	0	0	0.06	0	0	0	0.13
0S2P5gXlwNlcD5hsBCYxc2598	0	0	0.05	0.05	0	0	0	0	0	0
6EWgcAqvGNvJmA94XUUoNZ1257	0	0	0	0	0	0	0	0	0	0
5Cpbdd5vnNA3hu3BU44vGT677	0	0	0	0	0.07	0	0.16	0	0	0

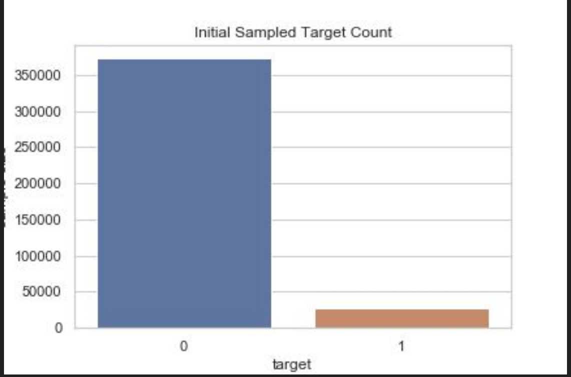
Binary Classification: Setting Target Parameter

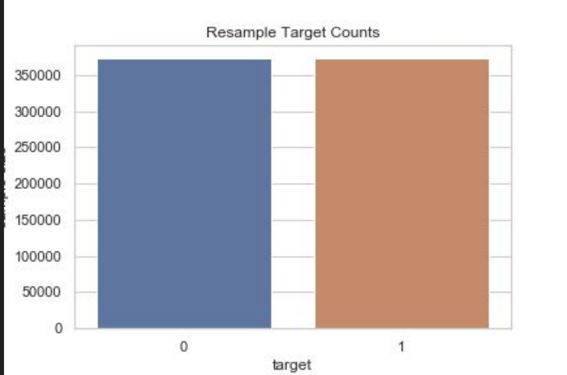
- Set our target as good or bad playlist
    - (1 or 0) using binary classification
    - Set thresholds for Valence, Energy, Danceability feature values
- Parameters (mean)
    - Valence >= 0.45
    - Energy >= 0.65
    - Danceability >= 0.52



Binary Classification: Class Imbalance

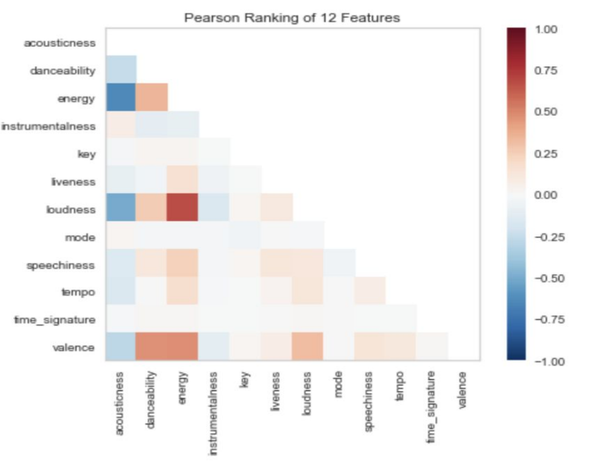
- 1 = “Good Playlist”
- 0 = “Bad Playlist”
- Initially had massive class imbalance, but used Sci-kit learn’s resample utility to create a 50/50 split between our binary classification

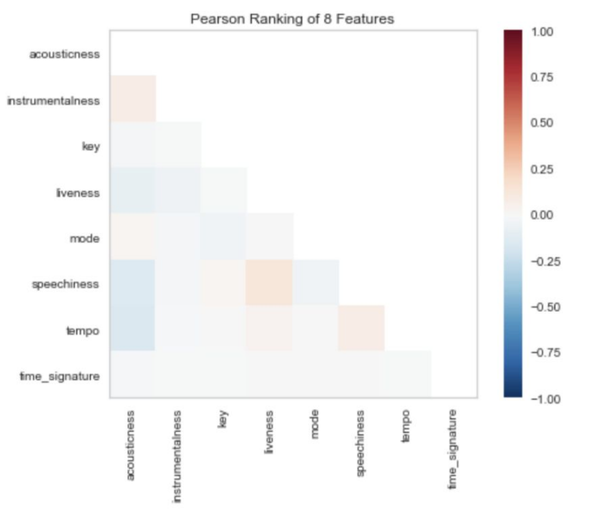




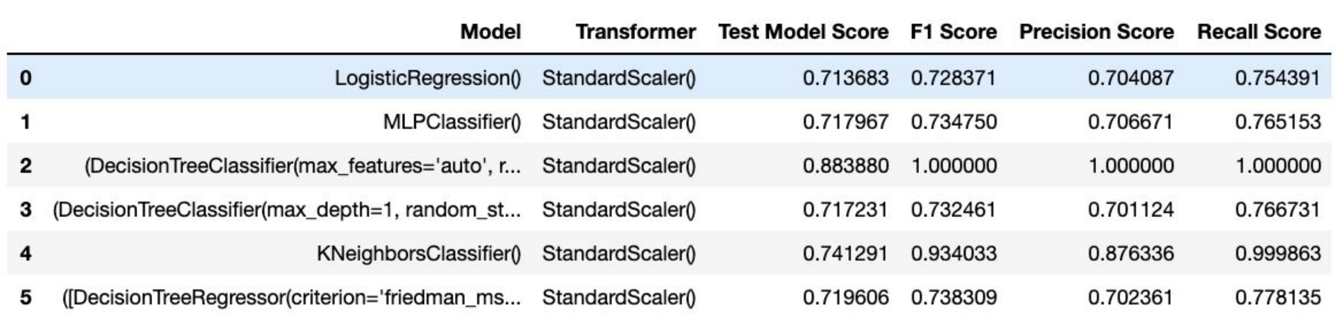
Feature Selection

- With 12 initial features saw recognized leakage between Valence, Energy, and Danceability
- Removed loudness because there was some collinearity found and did not provide

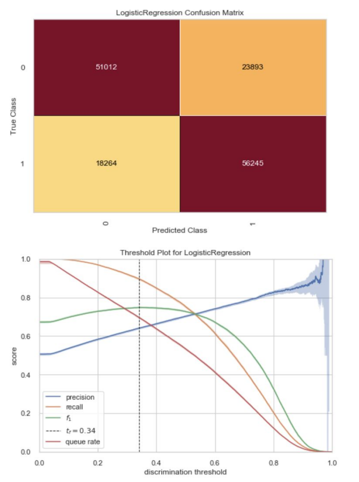


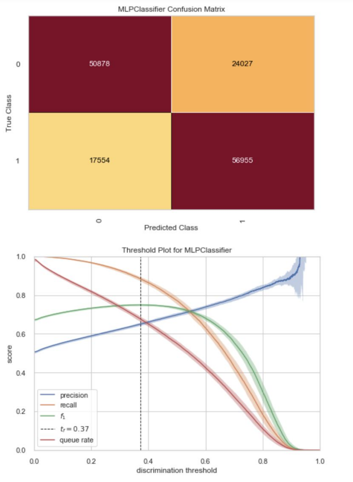


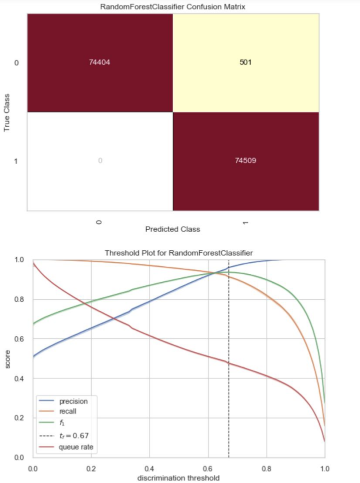
Model Selection

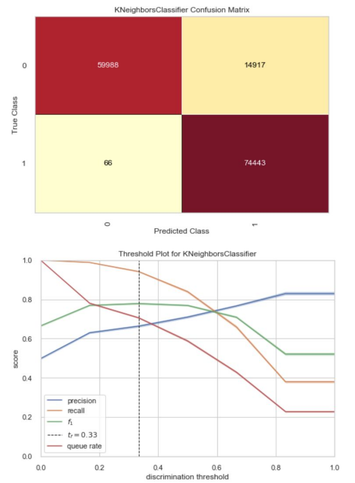


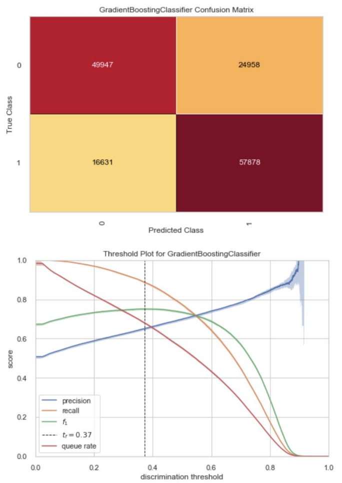
	Model	Transformer	Test Model Score	F1 Score	Precision Score	Recall Score
0	LogisticRegression()	StandardScaler()	0.713683	0.728371	0.704087	0.754391
1	MLPClassifier()	StandardScaler()	0.717967	0.734750	0.706671	0.765153
2	(DecisionTreeClassifier(max_features='auto', r...	StandardScaler()	0.883880	1.000000	1.000000	1.000000
3	(DecisionTreeClassifier(max_depth=1, random_st...	StandardScaler()	0.717231	0.732461	0.701124	0.766731
4	KNeighborsClassifier()	StandardScaler()	0.741291	0.934033	0.876336	0.999863
5	([DecisionTreeRegressor(criterion='friedman_ms...	StandardScaler()	0.719606	0.738309	0.702361	0.778135





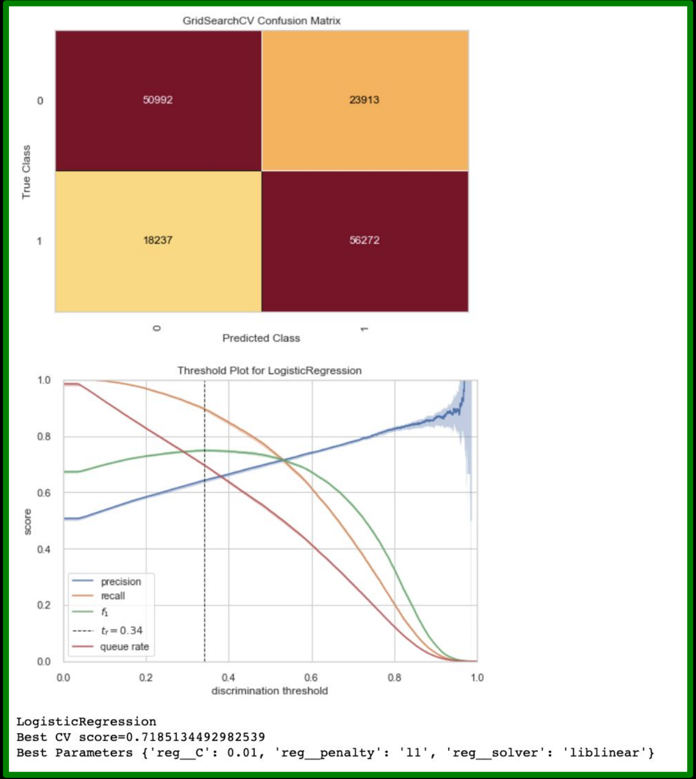




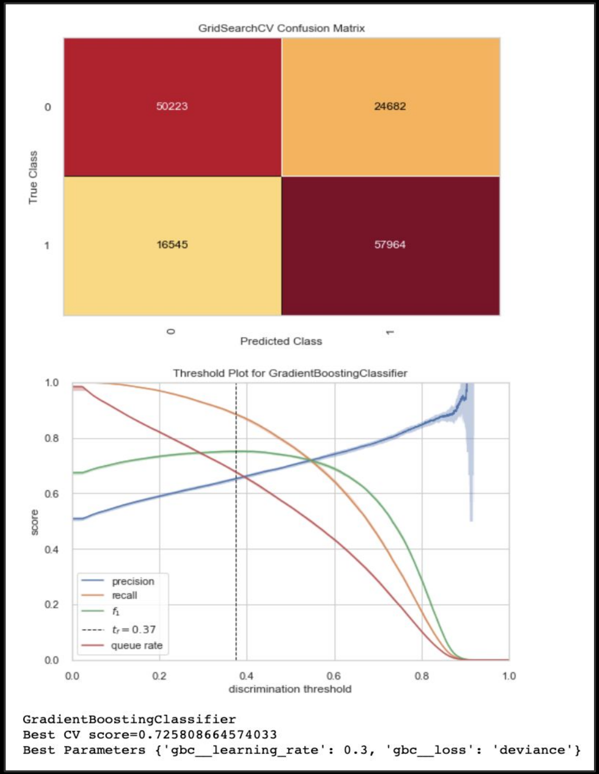


Grid Search Results

- Tested various models, opted for simplicity and favorability towards greater recall than precision



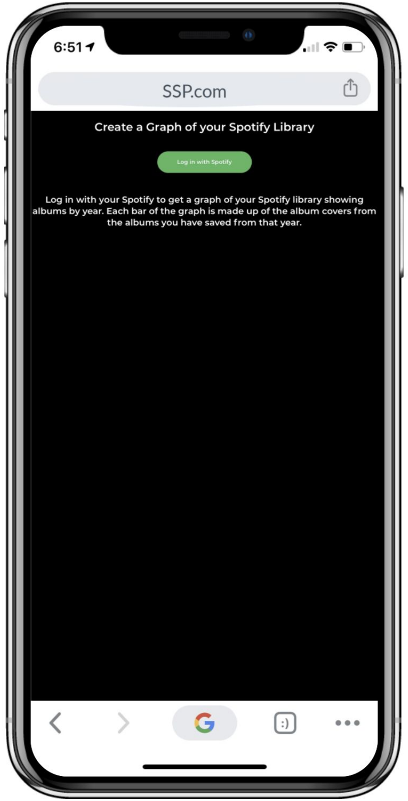
```
LogisticRegression
Best CV score=0.7185134492982539
Best Parameters {'reg__C': 0.01, 'reg__penalty': 'l1', 'reg__solver': 'liblinear'}
```



```
GradientBoostingClassifier
Best CV score=0.725808664574033
Best Parameters {'gbc__learning_rate': 0.3, 'gbc__loss': 'deviance'}
```

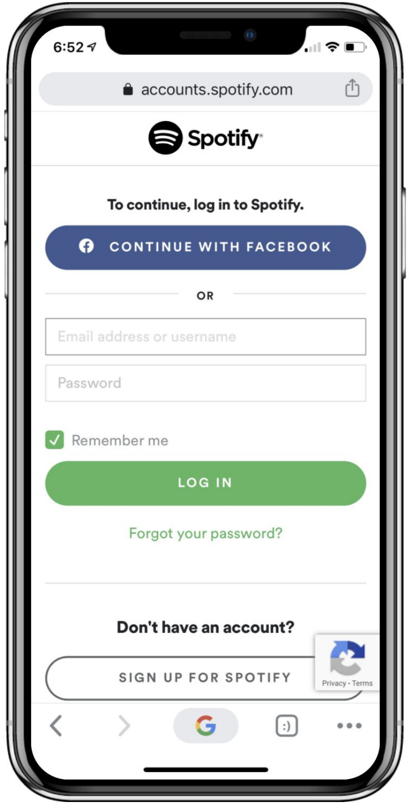
Live Jupyter Notebook Demo

Envisioned User Interface



Create a Graph of your Spotify Library

Log in with your Spotify to get a graph of your Spotify library showing albums by year. Each bar of the graph is made up of the album covers from the albums you have saved from that year.



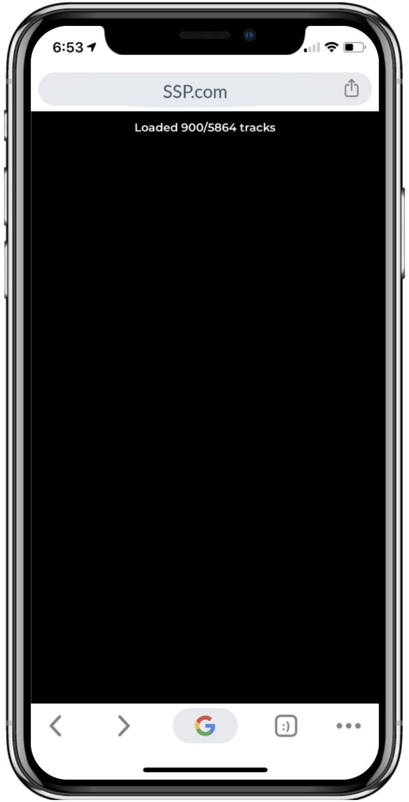
To continue, log in to Spotify.

Email address or username

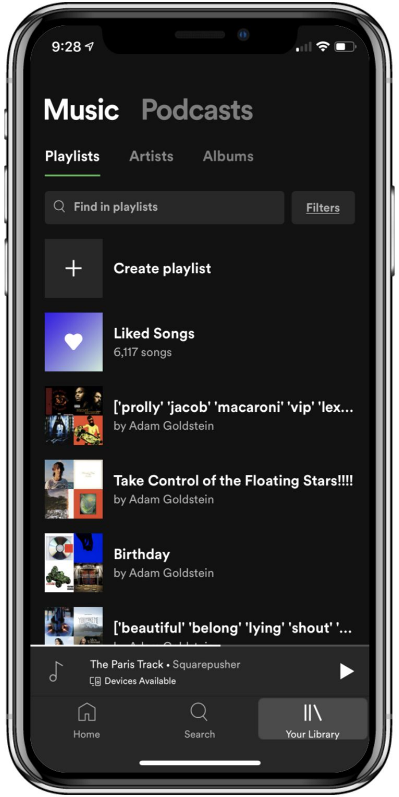
Password

☑ Remember me

Don't have an account?



Loaded 900/5864 tracks



Liked Songs
6,117 songs

['prolly' 'jacob' 'macaroni' 'vip' 'lex...
by Adam Goldstein

Take Control of the Floating Stars!!!!
by Adam Goldstein

Birthday
by Adam Goldstein

['beautiful' 'belong' 'lying' 'shout' '...

The Paris Track • Squarepusher
Devices Available

Thanks for Listening!

Any questions?

References

Iqbal, M. (2020, May 08). Spotify Usage and Revenue Statistics (2020). Retrieved June 19, 2020, from https://www.businessofapps.com/data/spotify-statistics/

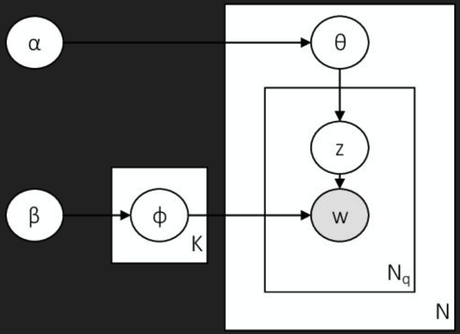
Spotify Audio Features: https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/

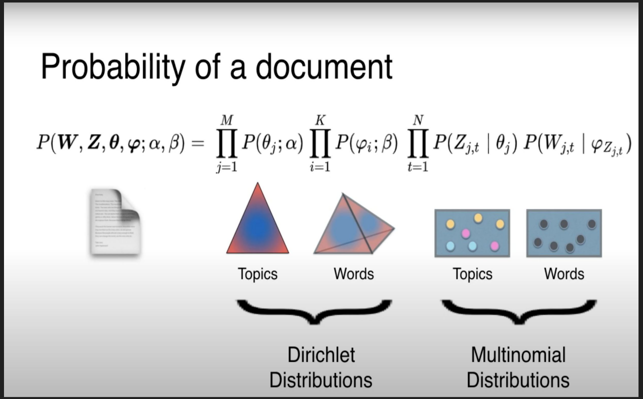
LDA: https://www.youtube.com/watch?v=T05t-SqKArY&t=670s

Q&A: https://www.youtube.com/watch?v=5qap5aO4i9A

Visualization:https://www.scikit-yb.org/en/latest/api/features/rankd.html

Q&A: Topic Modeling



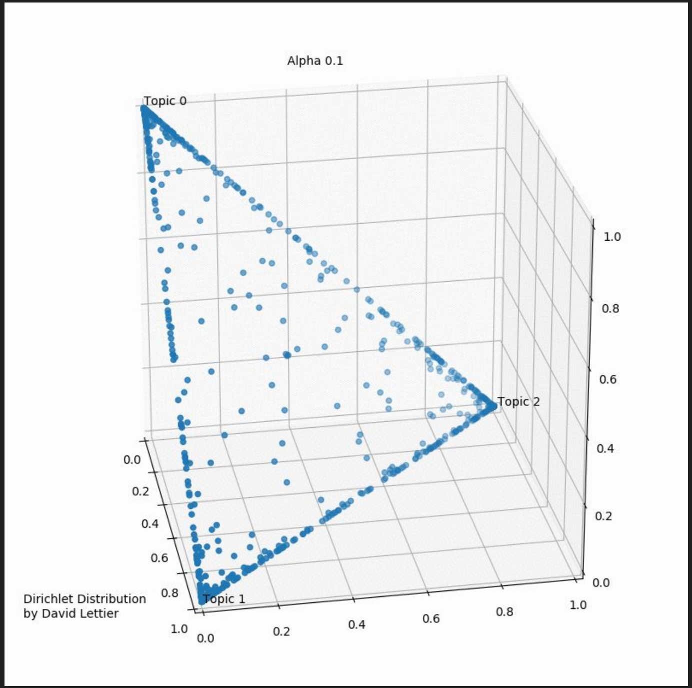


Probability of a document

$$P(\boldsymbol{W},\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\varphi};\alpha,\beta)=\prod_{j=1}^{M}P(\theta_j;\alpha)\prod_{i=1}^{K}P(\varphi_i;\beta)\prod_{t=1}^{N}P(Z_{j,t}\mid\theta_j)\,P(W_{j,t}\mid\varphi_{Z_{j,t}})$$

Topics
Words
Topics
Words
Dirichlet Distributions
Multinomial Distributions

Alpha



Dirichlet Distribution
by David Lettier