kaggle_lyrics_genre_clustering

May 18, 2021

[42]:
```
!pip3 install langdetect
from langdetect import DetectorFactory, detect
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
from scipy.cluster.hierarchy import linkage, dendrogram
from matplotlib import pyplot as plt
import json
import re
import numpy as np
import pandas as pd
```

```
Requirement already satisfied: langdetect in /usr/local/lib/python3.7/dist-
packages (1.0.9)
Requirement already satisfied: six in /usr/local/lib/python3.7/dist-packages
(from langdetect) (1.15.0)
```

[43]:
```
with open("artist_lyrics_kaggle.json", "r", encoding='utf-8') as json_file:
  artist_lyrics = json.load(json_file)

with open("artist_genre_kaggle.json", "r", encoding='utf-8') as json_file:
  artist_genre = json.load(json_file)
```

[44]:
```
def extract_artists_lyrics_of_file (json_file):
  json_file_content = {}
  for k,v in json_file.items():
    json_file_content[k] = ' '.join(re.findall('[A-Za-z]+', v))
  return json_file_content

def extract_genre_by_artist (json_file, artist):
  return json_file[artist]

def load_stopwords_file(txt_file):
  stopwords_file = open(txt_file)
  stopwords_file_content = stopwords_file.read()
  return stopwords_file_content.splitlines()
```

```
def detect_language (content):
  # check lg iso-codes here: https://en.wikipedia.org/wiki/
 ↪List_of_ISO_639-1_codes
  return detect(content)

def evaluate_tfidf_matrix (stop_words, max_features, lyrics):
  vectorizer = TfidfVectorizer(stop_words=stop_words, max_features=max_features)
  lyrics_tfidf_matrix = vectorizer.fit_transform(lyrics)
  feature_names = vectorizer.get_feature_names()
  return lyrics_tfidf_matrix, feature_names

def extract_features_by_artist␣
 ↪(json_content_artist_lyrics,stop_words,max_features):
  artists, lyrics =␣
 ↪zip(*extract_artists_lyrics_of_file(json_content_artist_lyrics).items())
  lyrics_tfidf_matrix, feature_names =␣
 ↪evaluate_tfidf_matrix(stop_words,max_features,lyrics)
  return zip(artists,lyrics_tfidf_matrix.toarray())

def extract_features_by_genre␣
 ↪(json_content_artist_lyrics,json_content_artist_genre,stop_words,max_features):
  words_per_genre = {}
  for artist,lyrics in␣
 ↪extract_artists_lyrics_of_file(json_content_artist_lyrics).items():
    genre = extract_genre_by_artist(json_content_artist_genre, artist)
    words_per_genre[genre] = lyrics if not genre in words_per_genre else␣
 ↪words_per_genre[genre]+" "+lyrics
  genres, lyrics = zip(*words_per_genre.items())
  lyrics_tfidf_matrix, feature_names =␣
 ↪evaluate_tfidf_matrix(stop_words,max_features,lyrics)
  return zip(genres,lyrics_tfidf_matrix.toarray())

def apply_kmeans (json_content_artist_lyrics,stop_words,max_features):
  artists, lyrics =␣
 ↪zip(*extract_artists_lyrics_of_file(json_content_artist_lyrics).items())
  lyrics_tfidf_matrix, feature_names =␣
 ↪evaluate_tfidf_matrix(stop_words,max_features,lyrics)
  kmeans = KMeans(n_clusters=10)
  kmeans.fit(lyrics_tfidf_matrix.toarray())
  y_kmeans = kmeans.predict(lyrics_tfidf_matrix.toarray())
  scatter_kmeans(kmeans,y_kmeans,lyrics_tfidf_matrix.toarray())

def scatter_kmeans(kmeans,y_kmeans,lyrics_tfidf_matrix):
  plt.scatter(lyrics_tfidf_matrix[:, 0], lyrics_tfidf_matrix[:, 1], c=y_kmeans,␣
 ↪s=50, cmap='viridis')
  centers = kmeans.cluster_centers_
```

```
  plt.scatter(centers[:, 0], centers[:, 1], c='black', s=200, alpha=0.5);

def apply_hierarchical_clustering␣
 ↪(json_content_artist_lyrics,stop_words,max_features):
  artists, lyrics =␣
 ↪zip(*extract_artists_lyrics_of_file(json_content_artist_lyrics).items())
  lyrics_tfidf_matrix, feature_names =␣
 ↪evaluate_tfidf_matrix(stop_words,max_features,lyrics)
  clustered = linkage(lyrics_tfidf_matrix.toarray(), method='ward')
  plot_dendrogram(clustered, artists)

def plot_dendrogram(clustered, artists):
  plt.figure(figsize=(10, 25))
  plt.title('Hierarchical Clustering Dendrogram')
  plt.xlabel('Distance')
  plt.ylabel('Artists')
  plt.tight_layout()
  dendrogram(clustered,leaf_font_size=8.,labels = artists,orientation = 'left')
  plt.show()

def apply_pca (json_content_artist_lyrics,stop_words,max_features):
  artists, lyrics =␣
 ↪zip(*extract_artists_lyrics_of_file(json_content_artist_lyrics).items())
  genre_targets_df = pd.DataFrame ([extract_genre_by_artist(artist_genre,artist)␣
 ↪for artist in artists],columns=['target'])
  lyrics_tfidf_matrix, feature_names =␣
 ↪evaluate_tfidf_matrix(stop_words,max_features,lyrics)
  pca = PCA(n_components=2)
  principal_components = pca.fit_transform(lyrics_tfidf_matrix.toarray())
  principal_components_df = pd.DataFrame(data = principal_components,columns =␣
 ↪['PC1', 'PC2'])
  principal_components_targets_df = pd.concat([principal_components_df,␣
 ↪genre_targets_df], axis = 1)
  scatter_pca(principal_components_targets_df)

def scatter_pca(dataframe):
  fig = plt.figure(figsize = (8,8))
  ax = fig.add_subplot(1,1,1)
  ax.set_xlabel('PC1', fontsize = 15)
  ax.set_ylabel('PC2', fontsize = 15)
  ax.set_title('Principle Component Analysis', fontsize = 20)
  targets =␣
 ↪['Metal','Hip-Hop','Electronic','R&B','Country','Folk','Pop','Indie','Rock','Jazz']
  colors =␣
 ↪["#c3618c","#b45ac2","#c8ac42","#7178ca","#cb7140","#4bafd0","#d0454e","#52a674","#877f3a","#
  for target, color in zip(targets,colors):
```

```
    indicesToKeep = dataframe['target'] == target
    ax.scatter(dataframe.loc[indicesToKeep, 'PC1']
                , dataframe.loc[indicesToKeep, 'PC2']
                , c = color
                , s = 50)
  ax.legend(targets)
  ax.grid()
```

[45]:
```
artist_features_collection =␣
 ↪extract_features_by_artist(artist_lyrics,load_stopwords_file("stopwords.
 ↪txt"),20)
for artist_features in artist_features_collection:
  print(artist_features)
```

```
/usr/local/lib/python3.7/dist-packages/sklearn/feature_extraction/text.py:385:
UserWarning: Your stop_words may be inconsistent with your preprocessing.
Tokenizing the stop words generated tokens ['aren', 'couldn', 'didn', 'doesn',
'hadn', 'hasn', 'haven', 'isn', 'quelqu', 'shouldn', 'wasn', 'weren', 'won',
'wouldn'] not in stop_words.
  'stop_words.' % sorted(inconsistent))

('annihilator', array([0.0273178 , 0.31716876, 0.17854181, 0.18718156, 0.0216386
,
       0.19954673, 0.1697119 , 0.07279283, 0.37436313, 0.03119693,
       0.16989261, 0.24957542, 0.17854181, 0.        , 0.15598464,
       0.07984799, 0.61353957, 0.27037337, 0.00622117, 0.11370048]))
('busta-rhymes', array([0.14378449, 0.12256078, 0.0293295 , 0.04774456,
0.05044255,
       0.09668792, 0.03765435, 0.00984424, 0.03470094, 0.01747352,
       0.14718349, 0.19934585, 0.05045668, 0.70845626, 0.01353583,
       0.55221583, 0.07014021, 0.02928661, 0.17932904, 0.19320499]))
('crduan-xshadows', array([0.01813367, 0.40554446, 0.09585919, 0.19414362,
0.05835292,
       0.29193479, 0.        , 0.40985876, 0.37103004, 0.10354327,
       0.43572329, 0.17257211, 0.04357236, 0.        , 0.12942908,
       0.        , 0.29768689, 0.23728665, 0.00516204, 0.00943436]))
('fats-domino', array([0.53722901, 0.36924551, 0.11793196, 0.03471539,
0.11163924,
       0.07649641, 0.128763  , 0.1420175 , 0.03155945, 0.15464128,
       0.6343448 , 0.15464128, 0.10836991, 0.        , 0.13254967,
       0.        , 0.1420175 , 0.05049511, 0.03776074, 0.01380261]))
('game', array([0.04699422, 0.12001572, 0.05584047, 0.03245708, 0.03043136,
       0.04002218, 0.03580105, 0.01830429, 0.04962914, 0.01849299,
       0.10332518, 0.08510549, 0.06346375, 0.91842568, 0.03604246,
       0.29623051, 0.05340322, 0.03094745, 0.08647508, 0.08046692]))
('american-idol', array([0.39633939, 0.13026864, 0.09114959, 0.18178009,
0.10142909,
       0.11952635, 0.11680796, 0.17496866, 0.16560294, 0.0911029 ,
```

```
       0.68581341, 0.18518581, 0.09931866, 0.        , 0.12771435,
       0.01089609, 0.22350011, 0.12132863, 0.05959577, 0.28440116]))
('dar-williams', array([0.04157767, 0.24620146, 0.29749375, 0.14068655,
0.03201911,
       0.12876595, 0.1031412 , 0.17585819, 0.20663337, 0.10551491,
       0.60508639, 0.21982273, 0.0888041 , 0.        , 0.18904755,
       0.        , 0.40887028, 0.22861564, 0.        , 0.19708709]))
('50-cent', array([0.09659613, 0.11464702, 0.02047469, 0.03565243, 0.07473247,
       0.06089455, 0.0440309 , 0.01765145, 0.03372999, 0.01922435,
       0.09621299, 0.11761805, 0.12796682, 0.85813725, 0.01363181,
       0.36172677, 0.06903288, 0.01625331, 0.17941443, 0.13662705]))
('amos-lee', array([0.36329588, 0.34573663, 0.06285189, 0.26275984, 0.12949604,
       0.11872023, 0.09874273, 0.19361251, 0.06223259, 0.08989152,
       0.48610906, 0.13137992, 0.15363795, 0.        , 0.16595358,
       0.        , 0.33882189, 0.15212412, 0.07446103, 0.37046159]))
('david-bowie', array([0.31168932, 0.21405056, 0.23480295, 0.14270038,
0.19934788,
       0.12711746, 0.13328904, 0.10822781, 0.20042188, 0.19721513,
       0.5500958 , 0.166751  , 0.26233157, 0.        , 0.16354425,
       0.00820762, 0.307848  , 0.15472569, 0.06043059, 0.26208854]))
('bobby-bare', array([0.18811546, 0.21482789, 0.20935288, 0.07914712,
0.09803122,
       0.13322456, 0.19798157, 0.12814295, 0.14133414, 0.28266827,
       0.56909613, 0.14321859, 0.36541593, 0.        , 0.24874808,
       0.        , 0.33920193, 0.13944968, 0.01352846, 0.10508185]))
('chamillionaire', array([0.09501377, 0.19877773, 0.05584937, 0.08593774,
0.08591262,
       0.11207609, 0.0689821 , 0.01382477, 0.10125708, 0.02540768,
       0.14872437, 0.20550329, 0.20905776, 0.53930734, 0.02167126,
       0.06056764, 0.16066621, 0.05567271, 0.62767432, 0.28147968]))
('e-40', array([0.07537628, 0.17209048, 0.0532903 , 0.03517681, 0.05740965,
       0.05259369, 0.06191141, 0.01448457, 0.05586905, 0.03759091,
       0.09595389, 0.1058753 , 0.14977012, 0.82715044, 0.02689991,
       0.39073769, 0.10173686, 0.02069224, 0.14607326, 0.18288149]))
('beirut', array([0.        , 0.08321148, 0.29413842, 0.1456201 , 0.06493116,
       0.02100989, 0.06365701, 0.24963445, 0.1456201 , 0.56167752,
       0.08239166, 0.06240861, 0.02100989, 0.        , 0.37445168,
       0.        , 0.56167752, 0.10401436, 0.        , 0.        ]))
('clannad', array([0.03892769, 0.1926402 , 0.26190395, 0.04445543, 0.0847959 ,
       0.11224455, 0.15870645, 0.28896029, 0.07409238, 0.31118801,
       0.57956298, 0.05927391, 0.30680177, 0.        , 0.21486791,
       0.        , 0.32600648, 0.22968639, 0.        , 0.14582023]))
('chumbawamba', array([0.18702674, 0.32538275, 0.20784559, 0.06118308,
0.15045985,
       0.20503687, 0.29785143, 0.05562098, 0.17798714, 0.09455567,
       0.41580106, 0.29201016, 0.20784559, 0.        , 0.1529577 ,
       0.23252287, 0.27532386, 0.37822268, 0.03327513, 0.1277114 ]))
('doris-day', array([0.14361149, 0.05994311, 0.18404046, 0.06473856, 0.062366  ,
```

```
       0.08717706, 0.06358788, 0.28293148, 0.0983067 , 0.06713628,
       0.85942468, 0.1174885 , 0.02663743, 0.        , 0.13906802,
       0.        , 0.2157952 , 0.09111353, 0.0172132 , 0.01835139]))
('andre-rieu', array([0.        , 0.07456574, 0.15061554, 0.16777292,
0.01939492,
       0.33888495, 0.03802866, 0.48467731, 0.29826296, 0.03728287,
       0.60910655, 0.03728287, 0.01882694, 0.        , 0.16777292,
       0.        , 0.24233866, 0.20505579, 0.        , 0.        ]))
('avicii', array([0.18751602, 0.13627302, 0.15073665, 0.24334468, 0.05401187,
       0.07864521, 0.14230848, 0.21738792, 0.19467575, 0.09084868,
       0.62646267, 0.34392715, 0.0196613 , 0.        , 0.06164732,
       0.00553632, 0.19467575, 0.19143115, 0.08928934, 0.37959117]))
('etta-james', array([0.51212012, 0.05757366, 0.10547521, 0.1138084 ,
0.07522423,
       0.07843029, 0.13247345, 0.12184193, 0.06292935, 0.08836887,
       0.53957245, 0.20351619, 0.16767855, 0.        , 0.11113055,
       0.        , 0.11246947, 0.05489581, 0.05607052, 0.51238351]))
('ella-fitzgerald', array([0.19060229, 0.0902568 , 0.22036493, 0.08850424,
0.02735096,
       0.09911997, 0.12245175, 0.3093267 , 0.10427727, 0.10340099,
       0.78608748, 0.1717508 , 0.15044995, 0.        , 0.19979175,
       0.        , 0.18752383, 0.0858754 , 0.03984166, 0.04694729]))
('city-and-colour', array([0.00760615, 0.30401795, 0.20469555, 0.20991715,
0.04518661,
       0.03655278, 0.06644988, 0.24610977, 0.40535726, 0.13029341,
       0.54470769, 0.04343114, 0.27049055, 0.        , 0.26058681,
       0.01235125, 0.33297204, 0.14477045, 0.        , 0.01582893]))
('armin-van-buuren', array([0.01622262, 0.37052462, 0.13513205, 0.28818581,
0.00535418,
       0.08835557, 0.11548059, 0.14923908, 0.14409291, 0.09263115,
       0.69807988, 0.30362434, 0.05197386, 0.        , 0.19555466,
       0.        , 0.14923908, 0.20070083, 0.        , 0.07314753]))
('ernest-tubb', array([0.16069097, 0.21693908, 0.17599668, 0.06045843,
0.07585259,
       0.11493661, 0.12333564, 0.40364893, 0.1138041 , 0.17248435,
       0.725396  , 0.131586  , 0.09159011, 0.        , 0.15470246,
       0.        , 0.25428105, 0.10491316, 0.03404153, 0.00388848]))
('curtis-mayfield', array([0.40792377, 0.13616233, 0.11996193, 0.06083849,
0.04822672,
       0.23992385, 0.17434636, 0.02317657, 0.11877991, 0.0898092 ,
       0.75155437, 0.15644183, 0.13166553, 0.01563204, 0.09850041,
       0.00494334, 0.19120668, 0.16223597, 0.00693267, 0.06651968]))
('ataraxia', array([0.        , 0.13854303, 0.        , 0.20781454, 0.07207146,
       0.06996086, 0.        , 0.06927151, 0.41562909, 0.20781454,
       0.41153418, 0.13854303, 0.        , 0.        , 0.69271514,
       0.        , 0.13854303, 0.13854303, 0.        , 0.        ]))
('britney-spears', array([0.67127893, 0.11005483, 0.04993067, 0.19044644,
0.1435763 ,
```

```
       0.10116388, 0.07805323, 0.09758768, 0.0769524 , 0.04943869,
       0.39927484, 0.23945523, 0.03126094, 0.        , 0.10188669,
       0.0029342 , 0.17496998, 0.06706466, 0.10801886, 0.41082067]))
('cradle-of-filth', array([0.00803427, 0.2064407 , 0.12355263, 0.03440678,
0.06364   ,
       0.02316612, 0.01949725, 0.39376652, 0.3020151 , 0.09939737,
       0.49209041, 0.08028249, 0.11583059, 0.        , 0.44728819,
       0.02609291, 0.2867232 , 0.36700569, 0.00457417, 0.00835994]))
('everything-but-the-girl', array([0.29269627, 0.20588428, 0.12537143,
0.19377344, 0.1795551 ,
       0.05504111, 0.10808964, 0.13927466, 0.15744092, 0.07266504,
       0.70150395, 0.11808069, 0.08256167, 0.        , 0.13624695,
       0.01549874, 0.38451917, 0.18771802, 0.0253585 , 0.08276105]))
('eric-clapton', array([0.48938042, 0.14797032, 0.16093842, 0.15840412,
0.12533211,
       0.10920821, 0.12771023, 0.18022026, 0.11477185, 0.15650707,
       0.60952945, 0.14607326, 0.18009776, 0.        , 0.12615418,
       0.        , 0.32249942, 0.0958013 , 0.01021418, 0.13793465]))
('frank-zappa', array([0.30227641, 0.21898931, 0.09815172, 0.05571917,
0.25075971,
       0.12170813, 0.27491679, 0.12569209, 0.11532573, 0.11143835,
       0.53373972, 0.26304633, 0.25257709, 0.        , 0.19048182,
       0.09507484, 0.246201  , 0.06478974, 0.15039005, 0.32019634]))
('acid-drinkers', array([0.19044942, 0.28481261, 0.20048115, 0.39701152,
0.02693861,
       0.24406401, 0.14085329, 0.17261371, 0.19850576, 0.18124439,
       0.08545653, 0.31070467, 0.2789303 , 0.        , 0.18987508,
       0.30926125, 0.3538581 , 0.08630685, 0.03097973, 0.23591579]))
('bjrthrk', array([0.0638124 , 0.13765046, 0.18808624, 0.26315529, 0.11372897,
       0.10630961, 0.05368387, 0.19433006, 0.12550483, 0.04858251,
       0.82578302, 0.21052423, 0.08586546, 0.        , 0.08097086,
       0.        , 0.19028152, 0.0850194 , 0.00968813, 0.07525223]))
('anthrax', array([0.03615637, 0.23027442, 0.10692688, 0.15086945, 0.        ,
       0.1897952 , 0.10529137, 0.08734547, 0.44202101, 0.13763529,
       0.19917738, 0.19851243, 0.38760992, 0.        , 0.06881764,
       0.08129426, 0.50819183, 0.25409591, 0.14567849, 0.2402018 ]))
('bathory', array([0.09241414, 0.0830615 , 0.35035606, 0.17100897, 0.04066768,
       0.1282994 , 0.03986966, 0.17100897, 0.39576361, 0.22475465,
       0.05321616, 0.0830615 , 0.40957117, 0.        , 0.39576361,
       0.08337038, 0.42019347, 0.23941256, 0.01169209, 0.03739561]))
('aquabats', array([0.13558545, 0.30721946, 0.27924902, 0.12903217, 0.12146214,
       0.07446641, 0.14414789, 0.01228878, 0.14132095, 0.08602145,
       0.21293485, 0.18433167, 0.53367591, 0.        , 0.15360973,
       0.        , 0.2826419 , 0.16589851, 0.03675868, 0.47698946]))
('bee-gees', array([0.19302695, 0.08869597, 0.11165486, 0.08281093, 0.08790753,
       0.11292849, 0.06645916, 0.17024582, 0.20891894, 0.06431509,
       0.83243706, 0.1652015 , 0.12991021, 0.        , 0.18033446,
       0.        , 0.16057754, 0.16225898, 0.00804734, 0.10984772]))
```

```
('dmx', array([0.06921774, 0.13120004, 0.04206092, 0.0362025 , 0.0189745 ,
       0.12205915, 0.06080409, 0.02014275, 0.09227555, 0.02177594,
       0.08193331, 0.13582743, 0.11271228, 0.81881966, 0.03184731,
       0.44820344, 0.09418094, 0.01769295, 0.12441177, 0.1080352 ]))
('gordon-lightfoot', array([0.11266207, 0.29278421, 0.37066344, 0.09484559,
0.06006562,
       0.1082837 , 0.22292917, 0.11134047, 0.19793862, 0.26804188,
       0.48997118, 0.16082513, 0.29986256, 0.        , 0.14845397,
       0.        , 0.40824841, 0.13195908, 0.06907621, 0.        ]))
('chemical-brothers', array([0.04969247, 0.41210448, 0.04093824, 0.44588353,
0.00702888,
       0.01364608, 0.0137819 , 0.10133717, 0.05404649, 0.02702324,
       0.38797655, 0.32427893, 0.07505344, 0.05467958, 0.35130218,
       0.01152759, 0.39183704, 0.2432092 , 0.0404165 , 0.1034135 ]))
('billie-holiday', array([0.2079342 , 0.07420653, 0.15925808, 0.06183877,
0.01930149,
       0.09836529, 0.13876671, 0.31537775, 0.08966622, 0.0757525 ,
       0.82812925, 0.11903964, 0.18736245, 0.        , 0.12522352,
       0.        , 0.15768887, 0.0602928 , 0.00924873, 0.01352267]))
('david-guetta', array([0.32082732, 0.10065518, 0.05929982, 0.14259484,
0.43111233,
       0.08301975, 0.1625587 , 0.06878104, 0.15433794, 0.02180862,
       0.56309875, 0.2952552 , 0.03219133, 0.        , 0.26338106,
       0.02003751, 0.11239829, 0.14930519, 0.02609392, 0.32282657]))
('bob-dylan', array([0.32972174, 0.27717696, 0.24208235, 0.13771685, 0.09884737,
       0.12236163, 0.1840357 , 0.19611577, 0.12377084, 0.20657528,
       0.41771046, 0.19175764, 0.36532428, 0.        , 0.26148769,
       0.00297455, 0.34603538, 0.15602099, 0.13557653, 0.07910061]))
('2pac', array([0.10341341, 0.09881436, 0.06774146, 0.06228299, 0.03053098,
       0.07580592, 0.04113088, 0.03593249, 0.17008047, 0.03832799,
       0.12412935, 0.10240761, 0.09032195, 0.86117175, 0.03852762,
       0.23332773, 0.16169623, 0.08384249, 0.24123893, 0.08970743]))
('cannibal-corpse', array([0.00550678, 0.24106868, 0.10585549, 0.16245933,
0.00545245,
       0.03175665, 0.01603637, 0.11005309, 0.82277789, 0.04716561,
       0.06745689, 0.12053434, 0.14290492, 0.        , 0.34064053,
       0.03576876, 0.15197808, 0.1781812 , 0.        , 0.        ]))
('edith-piaf', array([0.02333948, 0.06663422, 0.17945952, 0.17769126, 0.
,
       0.        , 0.09062287, 0.37759392, 0.11105704, 0.68855362,
       0.48383662, 0.02221141, 0.2243244 , 0.        , 0.06663422,
       0.        , 0.04442281, 0.04442281, 0.02657585, 0.        ]))
('buck-owens', array([0.09929869, 0.21104836, 0.20519526, 0.0771744 ,
0.09995755,
       0.08430503, 0.17832075, 0.20317342, 0.11339912, 0.18427357,
       0.78285419, 0.14489887, 0.13679684, 0.        , 0.09607425,
       0.        , 0.28349779, 0.11182413, 0.0113068 , 0.07577075]))
('enigma', array([0.02099931, 0.17985912, 0.08073287, 0.46963215, 0.        ,
```

```
       0.232107  , 0.08153643, 0.21982781, 0.2997652 , 0.00999217,
       0.63319856, 0.01998435, 0.16146574, 0.        , 0.03996869,
       0.        , 0.32974172, 0.05995304, 0.        , 0.        ]))
('diana-ross', array([0.26103721, 0.077096  , 0.14761567, 0.12206867,
0.02339523,
       0.14923782, 0.08355309, 0.14616117, 0.13813034, 0.09637   ,
       0.85242345, 0.154192  , 0.05190881, 0.        , 0.07763139,
       0.        , 0.16918289, 0.12688717, 0.01793654, 0.07434381]))
('bruce-springsteen', array([0.47655623, 0.21602711, 0.14703223, 0.10801355,
0.26594141,
       0.10163518, 0.13617807, 0.20529284, 0.1281403 , 0.11002623,
       0.42049039, 0.10532999, 0.28796636, 0.        , 0.35892703,
       0.00457903, 0.13149476, 0.12411495, 0.0168571 , 0.30001807]))
('go-betweens', array([0.06211796, 0.22660979, 0.03980258, 0.13793639,
0.12301007,
       0.02985194, 0.14069563, 0.16749419, 0.23646239, 0.17734679,
       0.57557619, 0.31528318, 0.12935839, 0.        , 0.14778899,
       0.        , 0.30543058, 0.20690459, 0.        , 0.4093607 ]))
('al-green', array([0.55699167, 0.06129295, 0.06075654, 0.10896524, 0.05786573,
       0.04470764, 0.10767166, 0.06469811, 0.06810327, 0.06696822,
       0.64735009, 0.14528699, 0.04470764, 0.        , 0.05334756,
       0.        , 0.25198211, 0.03859186, 0.03938455, 0.35866278]))
('brian-mcknight', array([0.38970951, 0.17681184, 0.13392851, 0.14554633,
0.38362085,
       0.11432922, 0.08027717, 0.15309317, 0.21562419, 0.10889022,
       0.51346701, 0.17357747, 0.10452957, 0.02036067, 0.22856164,
       0.01471697, 0.33206125, 0.07331222, 0.05675855, 0.22986551]))
('dean-martin', array([0.19360647, 0.12949217, 0.16216817, 0.05105691,
0.07621654,
       0.11508709, 0.10566599, 0.27452339, 0.09027454, 0.12357252,
       0.82937676, 0.1339319 , 0.06501673, 0.        , 0.13763167,
       0.        , 0.20866738, 0.11395311, 0.03275807, 0.03559835]))
('chris-brown', array([0.34412352, 0.14389471, 0.05195616, 0.07716633, 0.5682033
,
       0.10297108, 0.04448823, 0.08350366, 0.06300053, 0.02777242,
       0.37944656, 0.15806051, 0.07454579, 0.38720833, 0.09375523,
       0.16124877, 0.12003652, 0.04995308, 0.16726302, 0.31772059]))
('dolly-parton', array([0.16087068, 0.18101833, 0.14100456, 0.109285  ,
0.06060782,
       0.08849251, 0.1374974 , 0.1805369 , 0.1502067 , 0.12757941,
       0.82943802, 0.16031677, 0.11669343, 0.        , 0.10061923,
       0.        , 0.24986307, 0.07654764, 0.01958505, 0.04000553]))
('elton-john', array([0.15715708, 0.28362378, 0.16200646, 0.13483753,
0.10965011,
       0.12209183, 0.13358263, 0.19915659, 0.22627908, 0.13251275,
       0.63224987, 0.16350988, 0.25670589, 0.        , 0.17125917,
       0.02247871, 0.32082034, 0.17823352, 0.0027816 , 0.20080828]))
('arrogant-worms', array([0.0697352 , 0.17697245, 0.15639187, 0.0398188 ,
```

```
0.10126911,
       0.06255675, 0.17148692, 0.07078898, 0.09733485, 0.05309174,
       0.35483846, 0.13272934, 0.45577059, 0.        , 0.11945641,
       0.        , 0.2743073 , 0.09291054, 0.03176202, 0.65305713]))
('celine-dion', array([0.19749565, 0.10048815, 0.12028224, 0.15197282,
0.03033237,
       0.15912338, 0.06706678, 0.27169018, 0.1997357 , 0.07071388,
       0.80274092, 0.17554411, 0.06452641, 0.        , 0.14142777,
       0.        , 0.177405  , 0.13708568, 0.01261711, 0.11868876]))
('ferlin-husky', array([0.16469775, 0.20324187, 0.20178534, 0.08267466,
0.0483842 ,
       0.11654843, 0.14933164, 0.26869265, 0.1842956 , 0.16018215,
       0.72139235, 0.14295827, 0.10611126, 0.        , 0.19635232,
       0.        , 0.30830759, 0.16190454, 0.00206083, 0.04519747]))
('fall', array([0.16553599, 0.33476209, 0.1829682 , 0.07285999, 0.10448793,
       0.13722615, 0.17474644, 0.06892161, 0.27962481, 0.16738105,
       0.12283663, 0.14965835, 0.65232142, 0.        , 0.11618214,
       0.05040105, 0.32294696, 0.10239782, 0.10602567, 0.21315439]))
('angus-julia-stone', array([0.        , 0.08402787, 0.14426891, 0.10923624,
0.18359095,
       0.09335047, 0.0342835 , 0.41173658, 0.05041672, 0.14284739,
       0.75712006, 0.14284739, 0.04243203, 0.        , 0.01680557,
       0.        , 0.21847247, 0.05881951, 0.24129358, 0.13781178]))
('emmylou-harris', array([0.19961884, 0.22416534, 0.17523312, 0.11018296,
0.07774202,
       0.12662831, 0.10851161, 0.27482417, 0.16970709, 0.22289887,
       0.70098219, 0.1215812 , 0.10744221, 0.        , 0.23556358,
       0.        , 0.26595888, 0.18363827, 0.00606131, 0.0387726 ]))
('aaron-neville', array([0.34513721, 0.04349412, 0.18631085, 0.10348602,
0.10610848,
       0.07119195, 0.13921169, 0.23996758, 0.10648561, 0.12148359,
       0.77963598, 0.10798541, 0.08179501, 0.        , 0.19347386,
       0.        , 0.16047832, 0.14248075, 0.01435601, 0.06559403]))
('aretha-franklin', array([0.56604155, 0.09390468, 0.07881903, 0.13451212,
0.03432713,
       0.0627989 , 0.08089791, 0.09453918, 0.11674636, 0.0716975 ,
       0.61881632, 0.12689822, 0.10765526, 0.00342359, 0.07613893,
       0.00108265, 0.13133966, 0.06344911, 0.05465995, 0.40791962]))
('barbra-streisand', array([0.0896103 , 0.08921511, 0.19544383, 0.12398276,
0.05187065,
       0.10335335, 0.13583049, 0.16859032, 0.17515025, 0.10823892,
       0.79372616, 0.18958211, 0.17888079, 0.        , 0.13447866,
       0.        , 0.26698933, 0.19286207, 0.02511659, 0.05164208]))
('devendra-banhart', array([0.25903851, 0.29787634, 0.14523339, 0.14380237,
0.02137355,
       0.11411195, 0.25668816, 0.13353077, 0.07703698, 0.16948137,
       0.61530915, 0.19002456, 0.21266318, 0.        , 0.05135799,
       0.01752665, 0.35437013, 0.06162959, 0.14133412, 0.24146167]))
```

```
('diana-krall', array([0.18972985, 0.12778064, 0.21321672, 0.07777952,
0.08092335,
       0.07013708, 0.05950155, 0.24167208, 0.1666704 , 0.14166984,
       0.83064252, 0.10833576, 0.11221933, 0.        , 0.16389256,
       0.        , 0.15555904, 0.09444656, 0.02326572, 0.00911172]))
('bobby-darin', array([0.38329859, 0.0723755 , 0.23683019, 0.12448587,
0.14156571,
       0.12280084, 0.1417407 , 0.10132571, 0.11580081, 0.23739166,
       0.69942541, 0.15343607, 0.09063871, 0.        , 0.11869583,
       0.        , 0.21133647, 0.09843069, 0.07274147, 0.21524458]))
('fear-factory', array([0.03699095, 0.20681795, 0.1866552 , 0.15841375,
0.00457824,
       0.07110674, 0.00448841, 0.1452126 , 0.74806491, 0.18481604,
       0.19606626, 0.05720496, 0.22665274, 0.        , 0.01760153,
       0.0225254 , 0.2904252 , 0.31242711, 0.00526504, 0.00962259]))
('the-blood-brothers', array([0.26204631, 0.25630802, 0.22387773, 0.05541795,
0.27387518,
       0.08395415, 0.02826326, 0.12469039, 0.13854487, 0.07619968,
       0.69275844, 0.13854487, 0.23787008, 0.        , 0.11776314,
       0.04728043, 0.13854487, 0.10390866, 0.        , 0.29539069]))
('anti-flag', array([0.02199238, 0.413356  , 0.2510101 , 0.04970737, 0.01633154,
       0.10040404, 0.12541999, 0.12296033, 0.28516332, 0.07325296,
       0.07512165, 0.3217898 , 0.10304625, 0.        , 0.1726677 ,
       0.27230644, 0.34010304, 0.45259866, 0.0281722 , 0.30035038]))
('clancy-brothers', array([0.04295322, 0.10900573, 0.49540718, 0.01362572,
0.02835293,
       0.12385179, 0.22237249, 0.27251433, 0.0817543 , 0.13625717,
       0.62060771, 0.06812858, 0.33027145, 0.        , 0.2452629 ,
       0.        , 0.10900573, 0.06812858, 0.        , 0.        ]))
('dusty-springfield', array([0.44309575, 0.11484767, 0.12118417, 0.10027745,
0.05706975,
       0.13070578, 0.12763565, 0.18684174, 0.13798862, 0.08999258,
       0.70860442, 0.14227398, 0.05020487, 0.        , 0.08656429,
       0.        , 0.18855588, 0.08913551, 0.01128032, 0.30174841]))
('gary-numan', array([0.02633045, 0.15661137, 0.1075555 , 0.40405733, 0.0619178
,
       0.14551627, 0.10543115, 0.24744596, 0.14095023, 0.09083459,
       0.60476671, 0.18480142, 0.24358158, 0.        , 0.11902464,
       0.00534459, 0.43537961, 0.09709905, 0.02998157, 0.03424719]))
('b-b-king', array([0.85890965, 0.07911509, 0.07298776, 0.05591307, 0.0360119 ,
       0.06568898, 0.12531419, 0.06085776, 0.08215798, 0.08444014,
       0.37284744, 0.09128664, 0.118317  , 0.00513089, 0.10269747,
       0.00324509, 0.15252476, 0.05857559, 0.01547341, 0.09897937]))
('drake', array([0.12076425, 0.17866575, 0.04502754, 0.08652574, 0.23330413,
       0.07404529, 0.10408886, 0.02344782, 0.09709377, 0.05746366,
       0.20633523, 0.15488769, 0.1714382 , 0.57468547, 0.08553499,
       0.51223454, 0.18295902, 0.03203434, 0.11656743, 0.34050812]))
('celtic-woman', array([0.02550599, 0.0825289 , 0.37752725, 0.04854641,
```

```
0.02525432,
       0.10296198, 0.19311832, 0.42235379, 0.14563924, 0.15049388,
       0.47587438, 0.06796498, 0.10786493, 0.        , 0.48546413,
       0.        , 0.25729599, 0.17962173, 0.        , 0.00530797]))
('bing-crosby', array([0.11007067, 0.13093826, 0.4135545 , 0.05713669,
0.12879999,
       0.10579301, 0.11170264, 0.28568347, 0.13093826, 0.18807495,
       0.68124244, 0.12855756, 0.0721316 , 0.        , 0.26425721,
       0.        , 0.20235913, 0.15712591, 0.02563643, 0.01041203]))
('bill-miller', array([0.04363811, 0.33223156, 0.12582664, 0.05537193,
0.12962261,
       0.1118459 , 0.11295914, 0.51219032, 0.09690087, 0.16611578,
       0.3700781 , 0.12458683, 0.25165327, 0.        , 0.31838858,
       0.        , 0.41528945, 0.06921491, 0.        , 0.18162792]))
('bill-anderson', array([0.2033449 , 0.21616215, 0.20999656, 0.09469961,
0.05354747,
       0.08212737, 0.13859132, 0.20278069, 0.18528185, 0.1842525 ,
       0.70732632, 0.16572432, 0.18088812, 0.        , 0.18939922,
       0.        , 0.30571504, 0.16984169, 0.00615803, 0.04389314]))
('babyface', array([0.45309098, 0.07839848, 0.10777094, 0.14024617, 0.32037835,
       0.12932513, 0.0968486 , 0.23171107, 0.09843365, 0.04355471,
       0.69000956, 0.16463681, 0.06466256, 0.01175064, 0.08928716,
       0.01486367, 0.18162315, 0.07273637, 0.04064816, 0.08714811]))
('elvis-costello', array([0.16602063, 0.25026618, 0.19148232, 0.18580368,
0.25117647,
       0.16084514, 0.13279323, 0.19970735, 0.23130662, 0.16937206,
       0.52063131, 0.19717941, 0.28211728, 0.        , 0.14662059,
       0.        , 0.38551103, 0.20476324, 0.00756168, 0.09121214]))
('damien-rice', array([0.09362875, 0.05265199, 0.06544732, 0.10125383,
0.05478017,
       0.12680418, 0.23547675, 0.00405015, 0.05265199, 0.05265199,
       0.46919923, 0.11745444, 0.76900601, 0.        , 0.0688526 ,
       0.01382173, 0.17415658, 0.02430092, 0.0533059 , 0.17270614]))
('akon', array([0.28219877, 0.2159136 , 0.20931281, 0.09996   , 0.41392145,
       0.09624351, 0.0938028 , 0.0519792 , 0.179928  , 0.0726376 ,
       0.35631059, 0.2972144 , 0.1406636 , 0.42250251, 0.0839664 ,
       0.09665289, 0.1979208 , 0.0806344 , 0.22166184, 0.25574893]))
('craig-cardiff', array([0.06031561, 0.47355302, 0.08695736, 0.10045064,
0.0746506 ,
       0.01449289, 0.1610086 , 0.27265174, 0.08610055, 0.25830165,
       0.46888743, 0.22960146, 0.04347868, 0.        , 0.28700183,
       0.02448587, 0.44485283, 0.10045064, 0.        , 0.09414075]))
('eminem', array([0.12146791, 0.36446485, 0.08490714, 0.12109022, 0.12002089,
       0.13893896, 0.0696737 , 0.03391482, 0.11941836, 0.05684314,
       0.12746468, 0.19369659, 0.22288125, 0.26289827, 0.04752851,
       0.69891791, 0.19369659, 0.09768422, 0.15288541, 0.19898842]))
('blind-guardian', array([0.        , 0.34294655, 0.17146501, 0.37690165,
0.01413103,
```

```
       0.10630831, 0.0207806 , 0.0984698 , 0.3599241 , 0.13582042,
       0.15129256, 0.11205184, 0.13031341, 0.        , 0.23429022,
       0.        , 0.57384126, 0.32257349, 0.01218814, 0.02227553]))
('frank-sinatra', array([0.12637004, 0.10767659, 0.26410257, 0.09089582,
0.04655745,
       0.09462498, 0.18685452, 0.29646023, 0.09788781, 0.14543332,
       0.75600245, 0.17899485, 0.05649253, 0.        , 0.21395479,
       0.        , 0.24192273, 0.12445736, 0.05354165, 0.01834778]))
('agoraphobic-nosebleed', array([0.07326409, 0.116205  , 0.04694456, 0.185928  ,
0.07254119,
       0.07041683, 0.11852952, 0.069723  , 0.069723  , 0.023241  ,
       0.04602404, 0.185928  , 0.11736139, 0.        , 0.023241  ,
       0.83278909, 0.34861499, 0.209169  , 0.        , 0.        ]))
('eddy-arnold', array([0.14212624, 0.10905082, 0.20831929, 0.0735459 ,
0.06068717,
       0.06915517, 0.12416621, 0.26713224, 0.1470918 , 0.17836994,
       0.78429382, 0.13272076, 0.1289189 , 0.        , 0.14032896,
       0.        , 0.26459617, 0.17667923, 0.0111261 , 0.00369718]))
('frank-turner', array([0.07766103, 0.34266171, 0.33250021, 0.08062629,
0.05592346,
       0.08142862, 0.07538585, 0.188128  , 0.38969371, 0.23516   ,
       0.40581232, 0.20156571, 0.11535722, 0.        , 0.20828457,
       0.01146453, 0.40985029, 0.16125257, 0.        , 0.21304195]))
('all-american-rejects', array([0.06267745, 0.29824026, 0.23236057, 0.07242978,
0.11525246,
       0.22805759, 0.06953284, 0.20450761, 0.14059898, 0.22154991,
       0.24046012, 0.23007106, 0.07745352, 0.        , 0.170423  ,
       0.02907965, 0.48570556, 0.28119796, 0.03568432, 0.4518681 ]))
('beach-boys', array([0.40802737, 0.24744266, 0.17162153, 0.08720118,
0.24271099,
       0.05118537, 0.28432158, 0.11850417, 0.11254169, 0.10657922,
       0.58520721, 0.13937282, 0.11742526, 0.        , 0.16322272,
       0.        , 0.22210215, 0.09018242, 0.02496926, 0.30722006]))
('daft-punk', array([0.00257204, 0.08077508, 0.02224879, 0.16644562, 0.        ,
       0.10629977, 0.09237765, 0.01713411, 0.16155016, 0.13952059,
       0.06543757, 0.08811827, 0.00988835, 0.        , 0.50912777,
       0.        , 0.26925026, 0.73431889, 0.        , 0.12846251]))
('alan-jackson', array([0.12964111, 0.24536405, 0.12180282, 0.07901554,
0.0793248 ,
       0.0560013 , 0.20926717, 0.17050722, 0.15248613, 0.16218979,
       0.76040928, 0.13169257, 0.20580477, 0.        , 0.15525861,
       0.        , 0.23566038, 0.0998091 , 0.00663451, 0.20613336]))
('connie-smith', array([0.12303392, 0.17335003, 0.18428957, 0.08667501,
0.03322362,
       0.07985881, 0.09926606, 0.26154565, 0.13989651, 0.13077283,
       0.81304167, 0.10188116, 0.07525157, 0.        , 0.10036054,
       0.        , 0.22657153, 0.22809214, 0.00363882, 0.01995134]))
('faithless', array([0.06515942, 0.29351426, 0.14612998, 0.24804022, 0.06881759,
```

```
       0.1377797 , 0.06746718, 0.14055612, 0.14055612, 0.21910219,
       0.63855077, 0.26871023, 0.18788141, 0.        , 0.10335009,
       0.01410788, 0.21496819, 0.17776215, 0.31161802, 0.09944095]))
('cseh-tamxi-xi', array([0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 1., 0., 0., 0.,
0., 0., 0.,
       0., 0., 0.]))
('christina-aguilera', array([0.3609974 , 0.14546021, 0.11737854, 0.13449588,
0.14905818,
       0.15207534, 0.05890063, 0.09356235, 0.08186705, 0.03508588,
       0.37418095, 0.21709389, 0.09744634, 0.        , 0.09356235,
       0.03367564, 0.18785565, 0.10745051, 0.07346519, 0.70170954]))
```

[46]:
```
genre_features_collection =␣
 ↪extract_features_by_genre(artist_lyrics,artist_genre,load_stopwords_file("stopwords.
 ↪txt"),20)
for genre_features in genre_features_collection:
  print(genre_features)
```

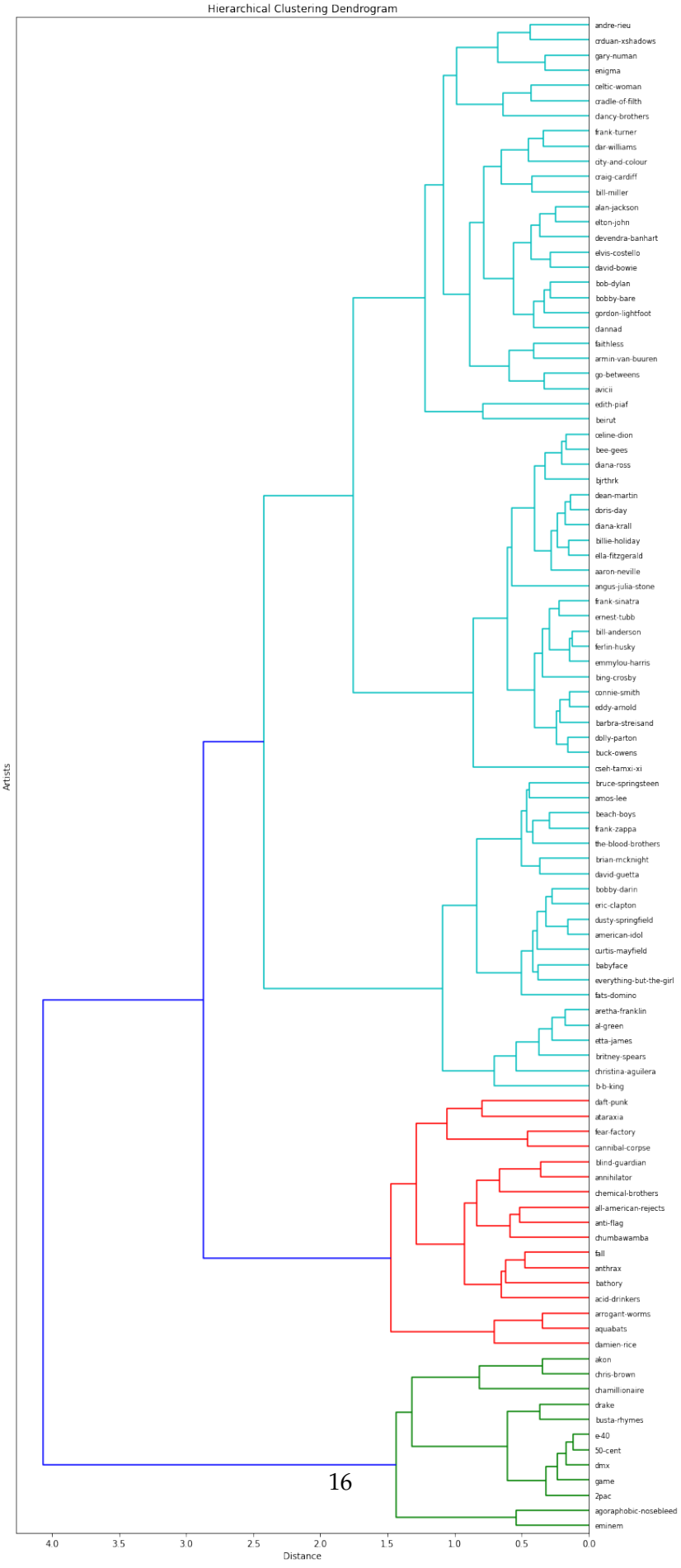
```
/usr/local/lib/python3.7/dist-packages/sklearn/feature_extraction/text.py:385:
UserWarning: Your stop_words may be inconsistent with your preprocessing.
Tokenizing the stop words generated tokens ['aren', 'couldn', 'didn', 'doesn',
'hadn', 'hasn', 'haven', 'isn', 'quelqu', 'shouldn', 'wasn', 'weren', 'won',
'wouldn'] not in stop_words.
  'stop_words.' % sorted(inconsistent))

('Metal', array([0.07090992, 0.2916229 , 0.19065845, 0.19065845, 0.04273379,
       0.13806302, 0.09485962, 0.15309028, 0.4696021 , 0.15402949,
       0.2009897 , 0.16436073, 0.36300242, 0.        , 0.2225914 ,
       0.06501061, 0.44847001, 0.26485558, 0.04789941, 0.11458291]))
('Hip-Hop', array([0.17999033, 0.23455642, 0.07350872, 0.08941204, 0.19260697,
       0.12570697, 0.0773962 , 0.04353978, 0.11849746, 0.0455542 ,
       0.23162314, 0.20960586, 0.17447718, 0.60462893, 0.0611748 ,
       0.4187969 , 0.17150856, 0.0651683 , 0.232542  , 0.24862203]))
('Electronic', array([0.15500641, 0.22797486, 0.1129568 , 0.26013045,
0.15583092,
       0.12449983, 0.11172005, 0.16654945, 0.20076629, 0.09646676,
       0.62579717, 0.25683244, 0.06967044, 0.00198639, 0.24652616,
       0.00760948, 0.26878772, 0.29269828, 0.04699663, 0.17479447]))
('R&B', array([0.46915234, 0.1442577 , 0.12508864, 0.13138504, 0.18259583,
       0.1123559 , 0.11067686, 0.15685052, 0.13054552, 0.08577107,
       0.64853001, 0.18595391, 0.11137646, 0.00471935, 0.12229023,
       0.01752544, 0.22932917, 0.10466029, 0.04071677, 0.26626809]))
('Pop', array([0.33953673, 0.12848781, 0.13330109, 0.15435919, 0.11979717,
       0.13430386, 0.0976026 , 0.16806366, 0.16712775, 0.08236055,
       0.7212565 , 0.21479258, 0.10388661, 0.02523241, 0.12949058,
       0.01163449, 0.21633015, 0.13249888, 0.05354773, 0.24460817]))
('Indie', array([0.10743405, 0.25643018, 0.20467363, 0.13723328, 0.09175025,
       0.09959215, 0.14507518, 0.1913424 , 0.19761592, 0.16781669,
```

```
       0.62892051, 0.18820564, 0.28152427, 0.        , 0.16389574,
       0.0113731 , 0.350533  , 0.13252814, 0.04548303, 0.19761592]))
('Folk', array([0.10323377, 0.23348199, 0.28751088, 0.09937457, 0.08007853,
       0.12445941, 0.18427711, 0.27110925, 0.14857945, 0.18331231,
       0.54414811, 0.12349461, 0.29137009, 0.        , 0.26628525,
       0.        , 0.34732858, 0.16015707, 0.02797925, 0.20743235]))
('Rock', array([0.46626424, 0.22118472, 0.16897683, 0.11848215, 0.16455855,
       0.11244083, 0.17880526, 0.1539186 , 0.15319724, 0.14445084,
       0.54543268, 0.16600125, 0.23732498, 0.00072412, 0.19837195,
       0.01497935, 0.27267126, 0.13624546, 0.03606763, 0.19025673]))
('Country', array([0.15266148, 0.194138  , 0.17810506, 0.09062097, 0.06436412,
       0.09143423, 0.14162431, 0.23003784, 0.15335856, 0.17322547,
       0.77446074, 0.14336702, 0.1434832 , 0.        , 0.15301002,
       0.        , 0.2718629 , 0.13918451, 0.01034008, 0.04833118]))
('Jazz', array([0.19617224, 0.09725629, 0.21426257, 0.07866806, 0.06522478,
       0.09792015, 0.12148738, 0.27832559, 0.10638444, 0.13111343,
       0.80809022, 0.14107141, 0.09841805, 0.        , 0.17409871,
       0.        , 0.19650417, 0.10771217, 0.02721848, 0.0390021 ]))
```

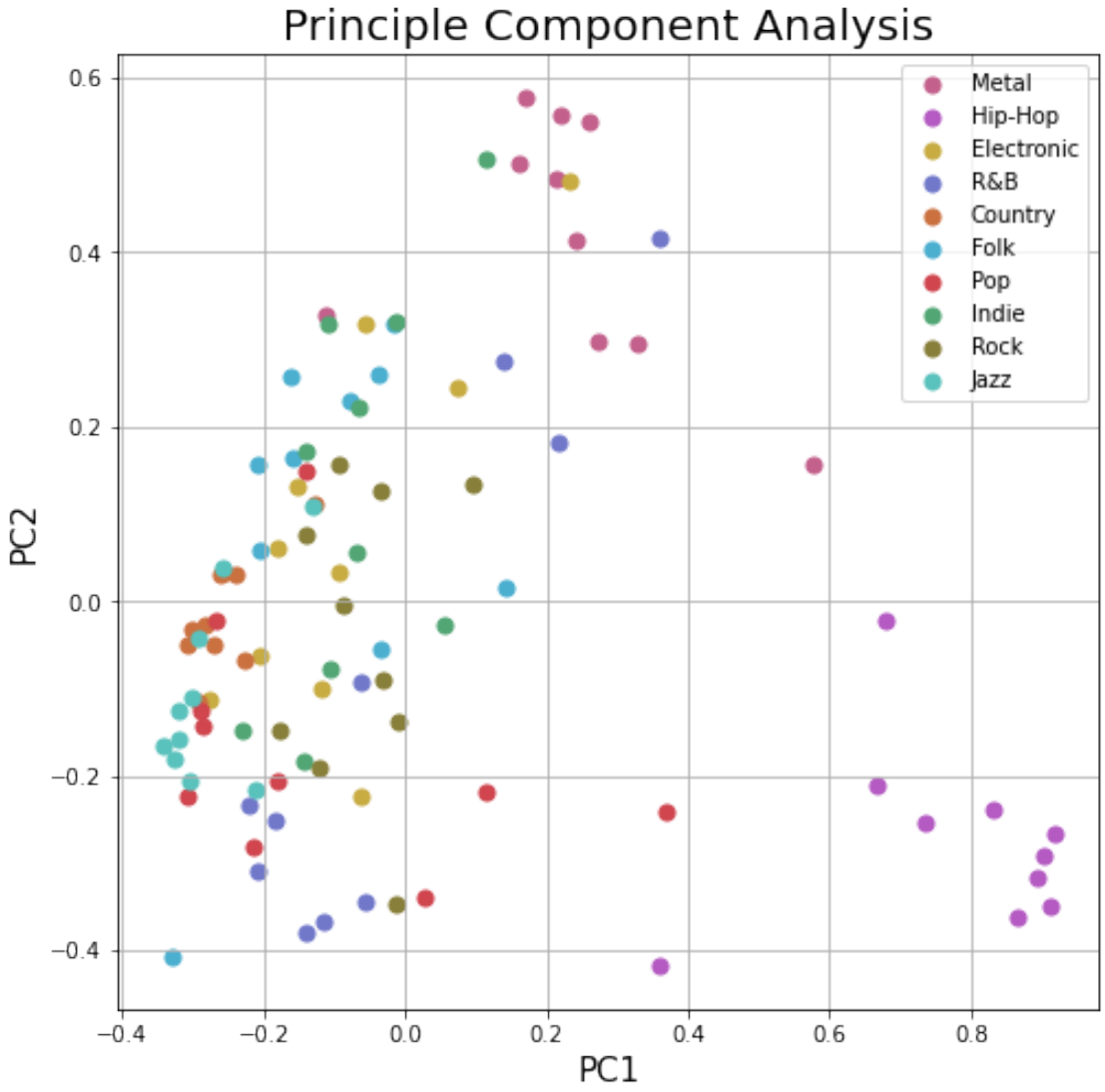
[47]:
```
apply_hierarchical_clustering(artist_lyrics,load_stopwords_file("stopwords.
↪txt"),20)
```

```
/usr/local/lib/python3.7/dist-packages/sklearn/feature_extraction/text.py:385:
UserWarning: Your stop_words may be inconsistent with your preprocessing.
Tokenizing the stop words generated tokens ['aren', 'couldn', 'didn', 'doesn',
'hadn', 'hasn', 'haven', 'isn', 'quelqu', 'shouldn', 'wasn', 'weren', 'won',
'wouldn'] not in stop_words.
  'stop_words.' % sorted(inconsistent))
```


Hierarchical Clustering Dendrogram
Artists
Distance
andre-rieu
crduan-xshadows
gary-numan
enigma
celtic-woman
cradle-of-filth
clancy-brothers
frank-turner
dar-williams
city-and-colour
craig-cardiff
bill-miller
alan-jackson
elton-john
devendra-banhart
elvis-costello
david-bowie
bob-dylan
bobby-bare
gordon-lightfoot
clannad
faithless
armin-van-buuren
go-betweens
avicii
edith-piaf
beirut
celine-dion
bee-gees
diana-ross
bjrthrk
dean-martin
doris-day
diana-krall
billie-holiday
ella-fitzgerald
aaron-neville
angus-julia-stone
frank-sinatra
ernest-tubb
bill-anderson
ferlin-husky
emmylou-harris
bing-crosby
connie-smith
eddy-arnold
barbra-streisand
dolly-parton
buck-owens
cseh-tamxi-xi
bruce-springsteen
amos-lee
beach-boys
frank-zappa
the-blood-brothers
brian-mcknight
david-guetta
bobby-darin
eric-clapton
dusty-springfield
american-idol
curtis-mayfield
babyface
everything-but-the-girl
fats-domino
aretha-franklin
al-green
etta-james
britney-spears
christina-aguilera
b-b-king
daft-punk
ataraxia
fear-factory
cannibal-corpse
blind-guardian
annihilator
chemical-brothers
all-american-rejects
anti-flag
chumbawamba
fall
anthrax
bathory
acid-drinkers
arrogant-worms
aquabats
damien-rice
akon
chris-brown
chamillionaire
drake
busta-rhymes
e-40
50-cent
dmx
game
2pac
agoraphobic-nosebleed
eminem
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0

[48]:
```
apply_pca(artist_lyrics,load_stopwords_file("stopwords.txt"),20)
```

```
/usr/local/lib/python3.7/dist-packages/sklearn/feature_extraction/text.py:385:
UserWarning: Your stop_words may be inconsistent with your preprocessing.
Tokenizing the stop words generated tokens ['aren', 'couldn', 'didn', 'doesn',
'hadn', 'hasn', 'haven', 'isn', 'quelqu', 'shouldn', 'wasn', 'weren', 'won',
'wouldn'] not in stop_words.
  'stop_words.' % sorted(inconsistent))
```



[49]:
```
apply_kmeans(artist_lyrics,load_stopwords_file("stopwords.txt"),20)
```

```
/usr/local/lib/python3.7/dist-packages/sklearn/feature_extraction/text.py:385:
UserWarning: Your stop_words may be inconsistent with your preprocessing.
Tokenizing the stop words generated tokens ['aren', 'couldn', 'didn', 'doesn',
'hadn', 'hasn', 'haven', 'isn', 'quelqu', 'shouldn', 'wasn', 'weren', 'won',
'wouldn'] not in stop_words.
  'stop_words.' % sorted(inconsistent))
```