AI-CA5-Clustering

Name: Ali Sadeghi Maharluee

Student Num: 810102471

Song Lyrics Clustering Analysis

In [102]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans, DBSCAN, AgglomerativeClustering
from sklearn.decomposition import PCA
from sklearn.metrics import silhouette_score, homogeneity_score, silhouette_samples
from sklearn.preprocessing import StandardScaler
from sentence_transformers import SentenceTransformer
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer, PorterStemmer
import re
import string
import warnings
warnings.filterwarnings('ignore')
```

install requirements:

In [83]:

```
try:
    nltk.data.find('tokenizers/punkt')
except LookupError:
    nltk.download('punkt')

try:
    nltk.data.find('tokenizers/punkt_tab')
except LookupError:
    nltk.download('punkt_tab')

try:
    nltk.data.find('corpora/stopwords')
except LookupError:
    nltk.download('stopwords')

try:
    nltk.data.find('corpora/wordnet')
except LookupError:
    nltk.download('wordnet')
```

```
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

In [84]:

```
from lyricFuncs import load_and_explore_data
from lyricFuncs import preprocess_dataset
```

First we load our data:

In [85]:

```
df = load_and_explore_data('musicLyrics.csv')
```

```
Dataset shape: (2999, 1)
Column names: ['Lyric']
Missing values: 0

First 3 lyrics samples:

Sample 1:
Cryptic psalms Amidst the howling winds A scorching source of agonizing bliss Beneath its
veil Mysteries of a life beyond Can you hear it? Sons and daughters with hearts ablaze Fo
rsaken ones in deaths...

Sample 2:
Im sleeping tonight with all the wolves Were dreaming of life thats better planned As lon
g as the wind that falls isnt longing for revenge I should be safe We should be safe Shes
two bitter ends So wa...

Sample 3:
Wings of the darkest descent Fall from the realm of dark From the blackest fall of creati
on Doomed by its end Winds of chaos blow through my soul Wings of the darkest descent sha
ll fall Lurking evil s...
```

Then we preproccess the text in our data

In [86]:

```
df = preprocess_dataset(df)
```

```
============================================================
TEXT PREPROCESSING
============================================================

Applying cleaned_only preprocessing...

Applying cleaned_no_stopwords preprocessing...

Applying cleaned_lemmatized preprocessing...

Applying cleaned_stemmed preprocessing...

Preprocessing Examples:
Original: Cryptic psalms Amidst the howling winds A scorching source of agonizing bliss B
eneath its veil Mysteries of a life beyond Can you hear it? Sons and daughters with heart
s ablaze Forsaken ones in deaths...
cleaned_only: cryptic psalms amidst the howling winds a scorching source of agonizing bli
ss beneath its veil mysteries of a life beyond can you hear it sons and daughters with he
arts ablaze forsaken ones in deaths ...
cleaned_no_stopwords: cryptic psalms amidst howling winds scorching source agonizing blis
s beneath veil mysteries life beyond hear sons daughters hearts ablaze forsaken ones deat
hs embrace chant hymn behold awe blessed cur...
cleaned_lemmatized: cryptic psalm amidst howling wind scorching source agonizing bliss be
neath veil mystery life beyond hear son daughter heart ablaze forsaken one death embrace
chant hymn behold awe blessed curse abort ...
cleaned_stemmed: cryptic psalm amidst howl wind scorch sourc agon bliss beneath veil myst
eri life beyond hear son daughter heart ablaz forsaken one death embrac chant hymn behold
awe bless curs abort law come reign us...
```

Choose best preprocessing method

In [87]:

```
best_preprocessing = 'Lyric_cleaned_lemmatized'
```

Extract features

In [90]:

```
from lyricFuncs import extract_features
features = extract_features(df[best_preprocessing].tolist())
```

```
============================================================
FEATURE EXTRACTION
============================================================
Loading SentenceTransformer model: all-MiniLM-L6-v2
Extracting features from text data...


Feature extraction complete!
Feature vector shape: (2999, 384)
Each text is represented by 384 features
```

Initialize clustering analyzer

In [91]:

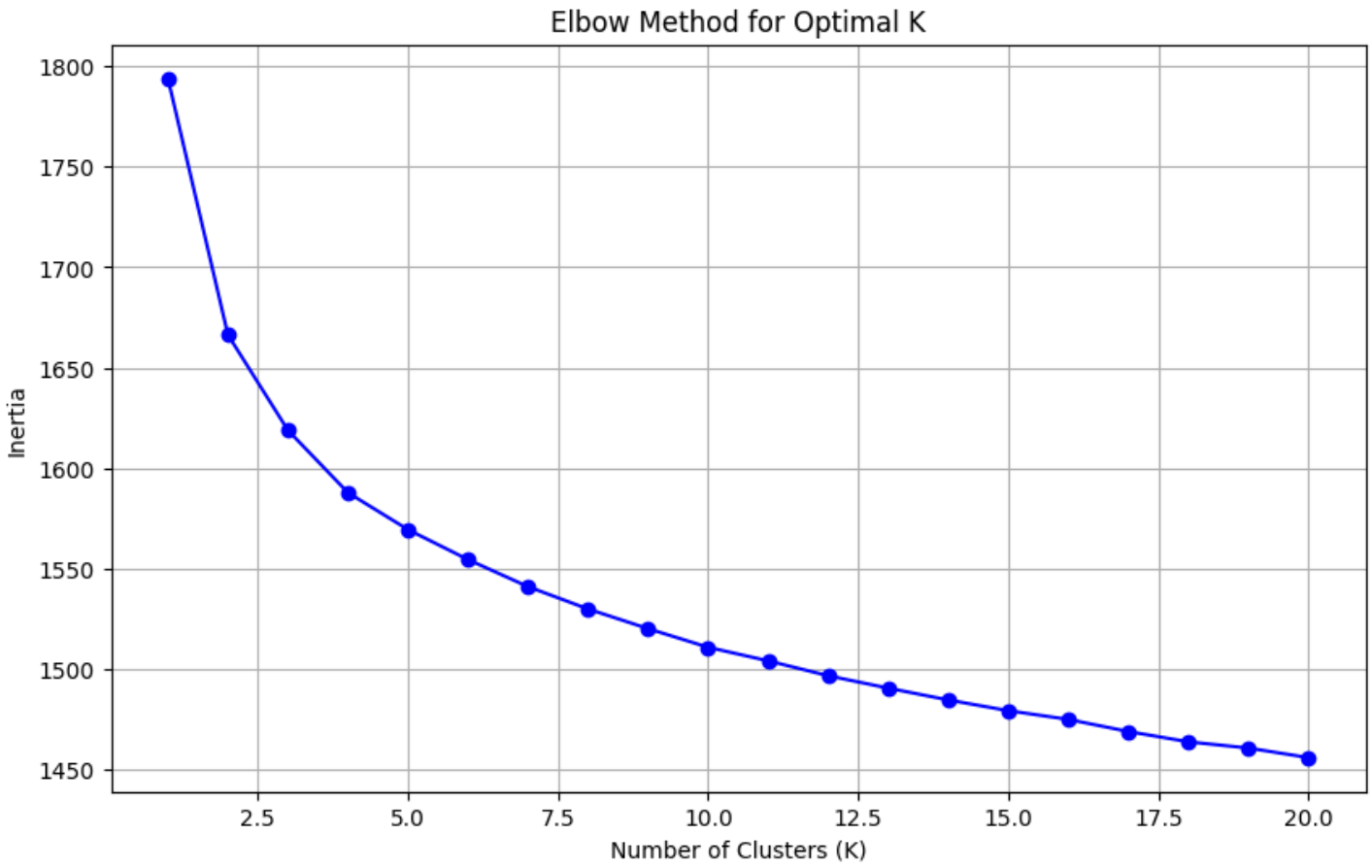
```
from lyricFuncs import ClusteringAnalyzer

analyzer = ClusteringAnalyzer(features)
```

Find optimal K using elbow method

In [92]:

```
k_range, inertias = analyzer.elbow_method(max_k=20)
```

```
============================================================
ELBOW METHOD FOR K-MEANS
============================================================
```



seems that the best k (the elbow) is 4 or 5. so we test our KMeans algorithm for k=4 and k=5

Perform clustering with different algorithms

In [93]:

```
kmeans_labels = analyzer.kmeans_clustering(n_clusters=4)
```

```
Performing K-Means clustering with 4 clusters...
```

In [94]:

```
dbscan_labels = analyzer.dbscan_clustering(eps=0.6, min_samples=10)
```

```
Performing DBSCAN clustering with eps=0.6, min_samples=10...
DBSCAN found 1 clusters
```

In [95]:

```
hierarchical_labels = analyzer.hierarchical_clustering(n_clusters=4)
```

```
Performing Hierarchical clustering with 4 clusters...
```

Dimensionality reduction for visualization

In [96]:

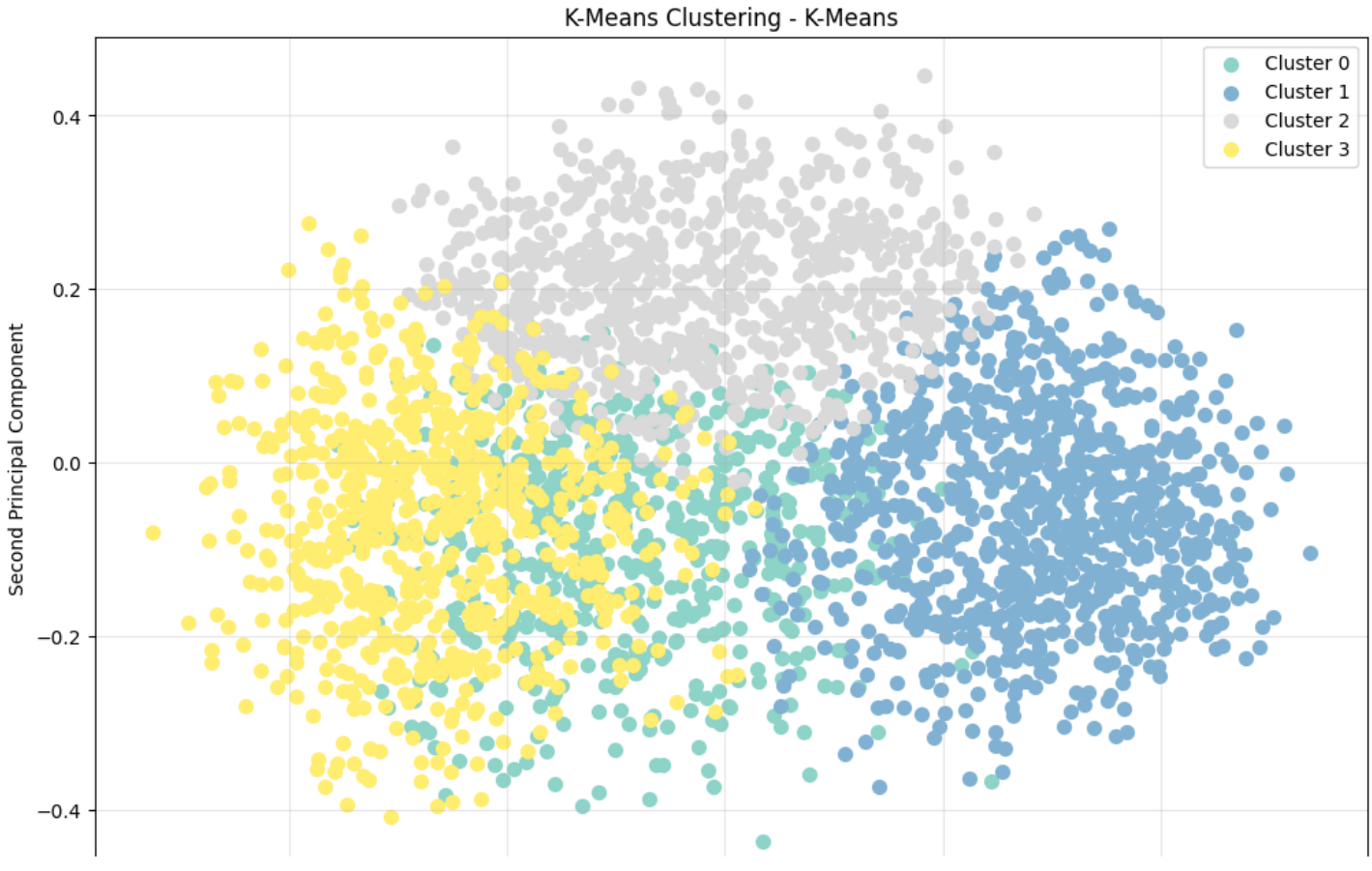
```
from lyricFuncs import perform_pca

features_2d, pca = perform_pca(features, n_components=2)
```

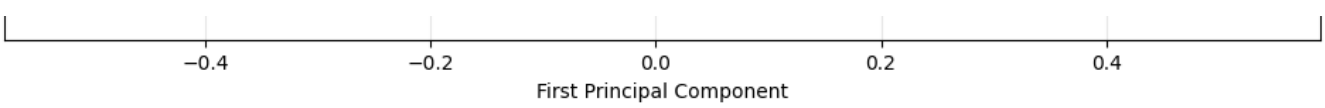
```
============================================================
DIMENSIONALITY REDUCTION (PCA)
============================================================
Original features shape: (2999, 384)
Reduced features shape: (2999, 2)
Explained variance ratio: [0.09424974 0.04659624]
Total explained variance: 0.141
```

Visualize clusters

In [97]:

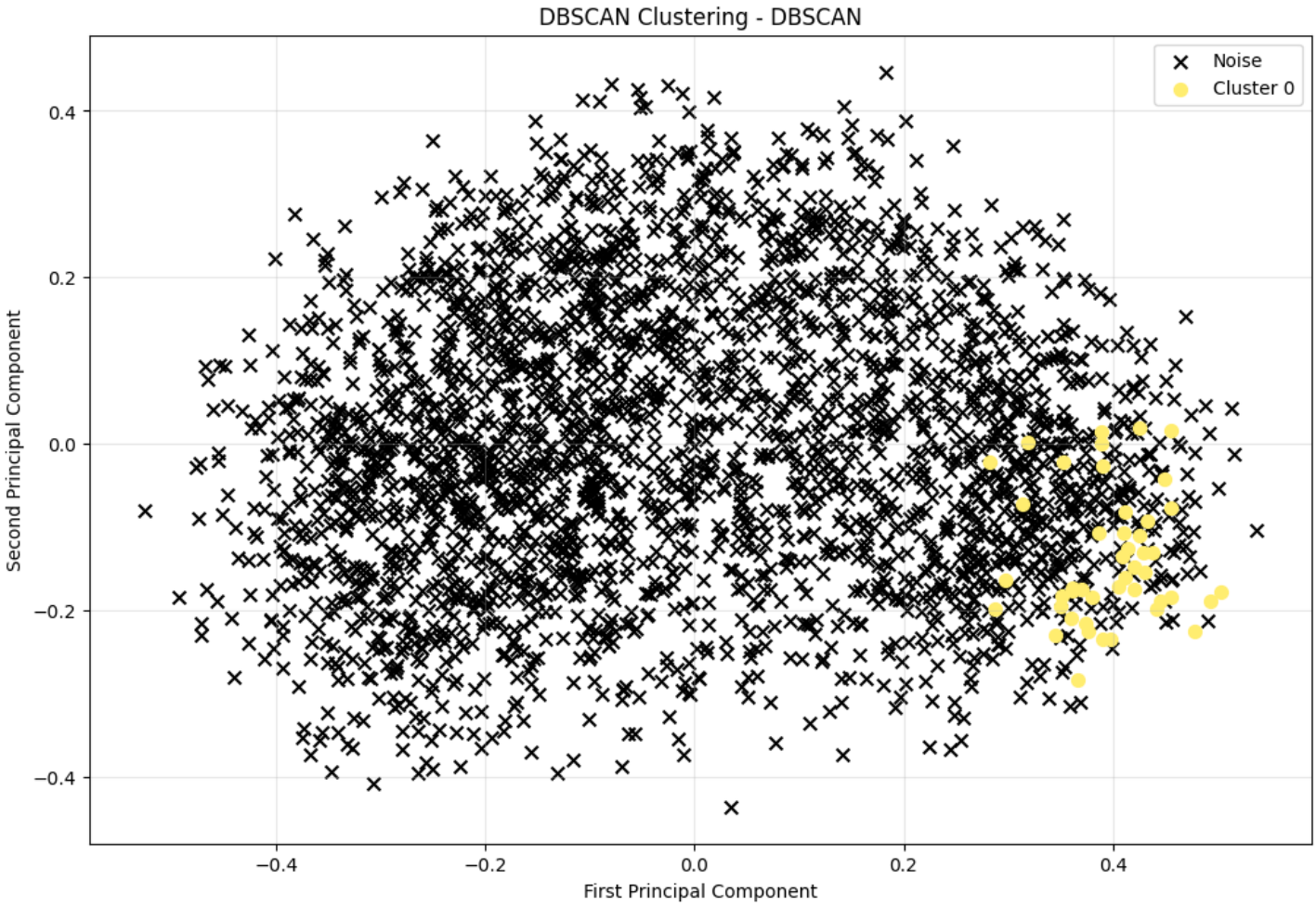
```
from lyricFuncs import visualize_clusters

visualize_clusters(features_2d, kmeans_labels, "K-Means Clustering", "K-Means")
```





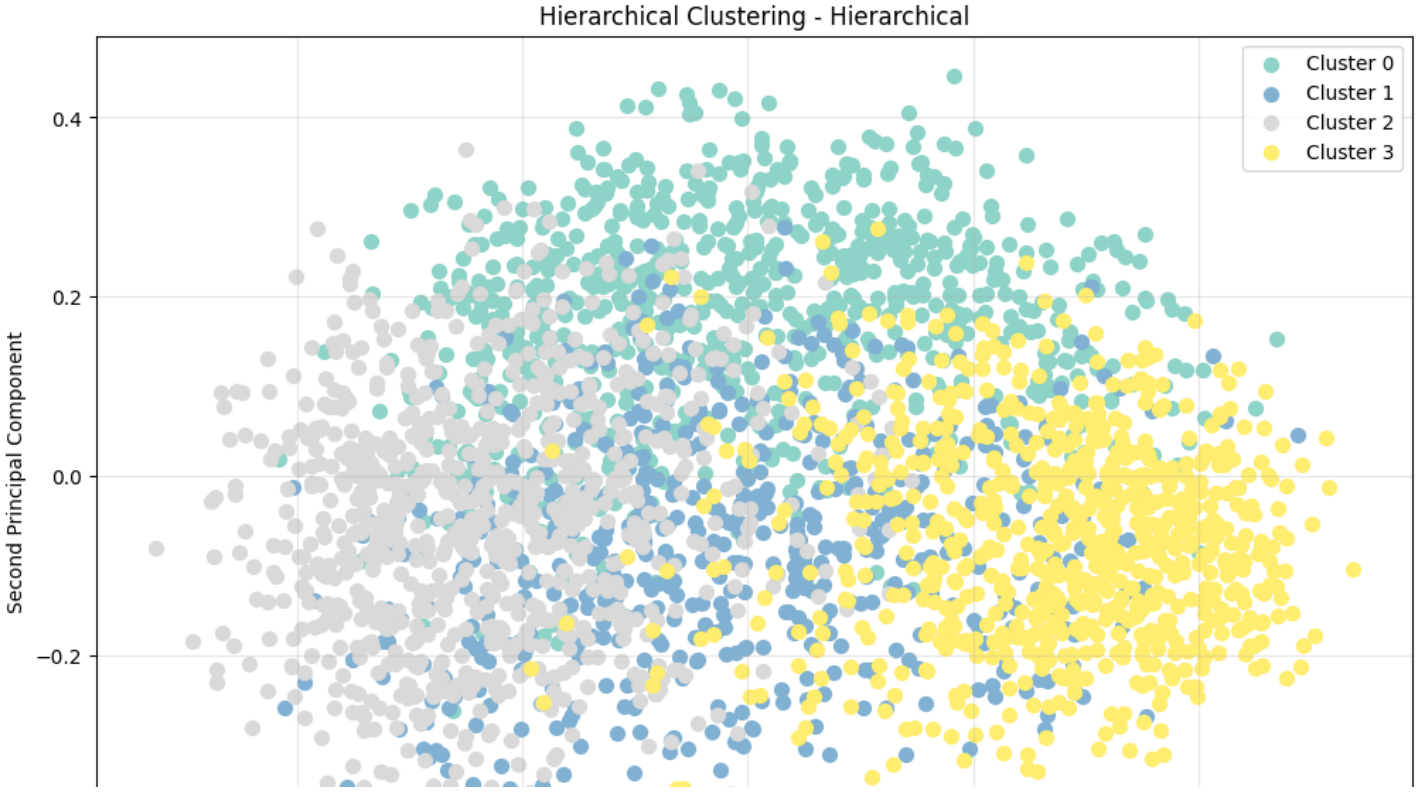
In [98]:

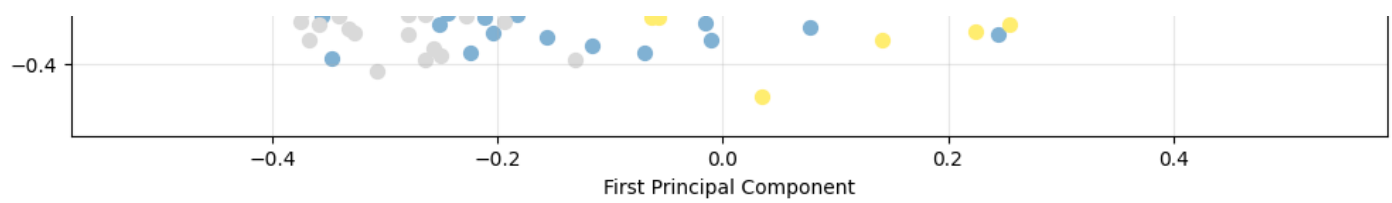
```
visualize_clusters(features_2d, dbscan_labels, "DBSCAN Clustering", "DBSCAN")
```



In [99]:

```
visualize_clusters(features_2d, hierarchical_labels, "Hierarchical Clustering", "Hierarch
ical")
```





Evaluate clustering results

In [106]:

```
from lyricFuncs import calculate_metrics


methods = {
    'K-Means': kmeans_labels,
    'DBSCAN': dbscan_labels,
    'Hierarchical': hierarchical_labels
}

results_summary = {}

for method_name, labels in methods.items():
    metrics = calculate_metrics(features, labels)
    results_summary[method_name] = metrics

    print(f"\n{method_name} Results:")
    print(f"  Silhouette Score: {metrics['silhouette']:.3f}")
    print(f"  Number of Clusters: {metrics['n_clusters']}")
    print(f"  Noise Points: {metrics['noise_points']}")
```

```
K-Means Results:
  Silhouette Score: 0.049
  Number of Clusters: 4
  Noise Points: 0

DBSCAN Results:
  Silhouette Score: -1.000
  Number of Clusters: 1
  Noise Points: 2953

Hierarchical Results:
  Silhouette Score: 0.032
  Number of Clusters: 4
  Noise Points: 0
```

Analyze samples from each cluster

In [107]:

```
from lyricFuncs import analyze_cluster_samples

analyze_cluster_samples(df, kmeans_labels, best_preprocessing, n_samples=4)
```

```
============================================================
CLUSTER SAMPLE ANALYSIS
============================================================

--- Cluster 0 (Size: 584) ---
Sample 1: im phone hello hello arrivederci sample include excerpt casino dir martin scors
ese close encounter third kind dir steven spielberg know sometimes prepared adversity hap
pens sometimes caught short dont...
Sample 2: alright party night friend came party bang bang gang blow whistle party night p
arty night come people gon na show rock come people gon na show got come people gon na sh
ow rock blow whistle come people...
Sample 3: eight mile high touch youll find stranger known sign street say youre going som
ewhere nowhere warmth found among afraid losing ground raingrey town known sound embrace
small face abound round square s...
Sample 4: ask question there multiple choice youve one shot get right say distance voice
dont get head spin could see mess youre enemy closing see white eye he taking prey sight
```

matter time may game begin best ...

--- Cluster 1 (Size: 934) ---
Sample 1: blame try shame still ill care run around even put still ill saturday id rather chillin wannatake karen millen buy new dress co goin dinner none fit co used slimmer fatter somethin dont know anymore s...
Sample 2: produced zck intro one time dollar two time cent middle finger landlord collect in rent one time scuffle two time knuckle middle finger police cause theyre government muscle verse one time dollar two t...
Sample 3: verse yoh whats happening im anakin light saber panicking mace windu im hit ninjitsu wait thats wrong nation patient vacant relation hating im waiting edward kenway freeing plantation go away cause im...
Sample 4: st st st st never said dont like religion dont like tv say got bad attitude around come naturally say need compassion forgive cant forget say control temper feel like shit feel like shit cause born be...

--- Cluster 2 (Size: 757) ---
Sample 1: bridge zara larsson used happy without feel low watched left never seem let go cause upon time everything clear see time hasnt changed thing buried deep inside feel there something know hook zara lars...
Sample 2: verse light around christmas tree dont burn bright around world isnt silent night outside hear voice sing sweetest sound caroling somehow there sadness song heart know somethings wrong chorus differen...
Sample 3: tonight im writing letter remembering every sky picture shot heart drove home said id never forget like time clock read said make wish wished wed never go separate way say thousand bitter thing see wi...
Sample 4: got secret hidden corner mind want free found cant decide crumpled paper holding piece word play youre unaware cant hear everything try say chorus late late late go x broken wish falling fold hand loo...

--- Cluster 3 (Size: 724) ---
Sample 1: life nothing reflection solitude end giving purpose hope isolated light abyss ignorance brief moment must fly dream fall deep subconscious glimpsing yearning soul alike heart united medium sorrow sepa...
Sample 2: strapped chopping block victim gain consciousness scream terror pendulum swing five foot across fifteen foot high contraption forged meticulous perfection swaying back forth metronomic momentum taunti...
Sample 3: came vision life using spirit love might victim still burn look damage done real time might walking forever dont one faking animal charge real waiting someone steer ship together everlasting soul fort...
Sample 4: yearsof war confusion never find way home lost without trace go rest soul battlefield go voyage far home world war neverending fight survive destiny take u sanctuary place well free evil darkness salv...

This clustring works well (why? see the Report)

In [108]:

```
analyze_cluster_samples(df, hierarchical_labels, best_preprocessing, n_samples=4)
```

```
============================================================
CLUSTER SAMPLE ANALYSIS
============================================================

--- Cluster 0 (Size: 803) ---
Sample 1: heart go bang night close friend told never let heart falll careless hand said thanks thats nice appreciate good advice thing dont always go way planned oh take heart shake take heart break get doctor...
Sample 2: verse young first saw close eye flashback start im standing balcony summer air see light see party ball gown see make way crowd say hello little know prechorus romeo throwing pebble daddy said stay aw...
Sample 3: there cold eye he waiting fall hope forum new world gone stand edge vision clear courage one last step time come one decision wrong right end tunnel show light also path far away paradise follow star ...
Sample 4: well want collapsing star ill wait ill wait good else dont want ghost dont want fading light dont want weight carry want man come home every night well want dont someone youre kid wouldnt know u dont ...

--- Cluster 1 (Size: 582) ---
Sample 1: produced j dilla chorus ayo yall dance real slow cause fantastic ayo yo yo yyo yo yo yoyyo fantastic yo yo yoyo fantstic yo yo yoyo fantastic yo yo yyo yo yo say fantastic
```

```
Sample 2: intro phoneeg action movie playing fuck popcorn g fucking soda soda opening oh
message tone who texting dawg ting kinda sweet bum nice yo phonee plug chuckle gunshot ba
ckground movie boring g dialogue...
Sample 3: youll never half man claim loud word mean nothing youre good youll never half m
an claim loud word mean nothing youre good always always youve crossed line face come tes
t wish never prosper know know k...
Sample 4: woke late today thats third time week already wednesday im letting slip away sl
ip away say say quick quick write another line fill another page fill another day least w
rite feel like im something youn...

--- Cluster 2 (Size: 837) ---
Sample 1: heard end see every death reassemble war youll lay dead truly see end life gone
cooperation lay dead xa extropian council overseer prepare walk way terror unbeknownst ma
n towering machine death visito...
Sample 2: cant get laid one want salvation lie every bottle lustful thought alone wallow
dirt freak dumpster child keep rat
Sample 3: age unholy revelation near rising apocryphal sign blazing hate raised warfare b
aalberiths black spear carried beast summoned beyond gate towards bestial armageddon brin
g forth horror death satariel bl...
Sample 4: ray rising sun veil ice reddenit strewed chained coldness weaved web dream hoar
frost blood glimmering sparkling melt scattering crimson ash bloody scab clotting blacken
ing growing dim disappearing dus...

--- Cluster 3 (Size: 777) ---
Sample 1: verse yo could love yo could got vacant spot right next nature vibe girl feel c
ould fill wan na deal adored first day saw quality material many guy yo major traffic fac
t played back planned next move ...
Sample 2: verse coming new overlordian boy within man try never needed comp never wanted
comp feel exceeded skill needed im rough stuff enough puff got shot got proof aloof type
fella helluva guy love high roll...
Sample 3: pocohantas first verse think got aero sway bought gucci bag always tryin start
fad swag something aint never second verse nickname equal dumbo get wifi next always lyin
g face promising true im allergi...
Sample 4: verse state mississippi many year ago boy fourteen year got taste southern law
saw friend ahangin color crime blood upon jacket put brand upon mind chorus many martyr m
any dead many lie many empty wor...
```

This method of clustering is also good but not as Kmeans (why? see the report.)

Find best method based on silhouette score

In [109]:

```
best_method = max(results_summary.keys(),
          key=lambda x: results_summary[x]['silhouette'])

print(f"Best performing method: {best_method}")
print(f"Best silhouette score: {results_summary[best_method]['silhouette']:.3f}")
```

```
Best performing method: K-Means
Best silhouette score: 0.049
```