

# COMP9727: Content-Based Music Recommendation

Written by: Sam Harrison (z5217483)

## Part 1: Topic Classification

### Question 1

#### Regex

The regex in the tutorial notebook removes all non-word and non-whitespace characters from the data. Whilst many of these special characters can be safely removed without affecting the meaning of the text, some characters such as the apostrophe may carry significant meaning. Additionally, this dataset contains many instances of the "&" and ":" characters, which are likely to be meaningful in the context of highly curated text like a song's metadata. As such, the following changes to the text preprocessing should be evaluated.

- Update the regex to keep the "&", "" and ":" characters
- Update the regex to keep the above 3 characters as well as the "-" character

### Question 2

#### Step 1: Load the dataset and remove any duplicate or missing values

In [260...

```
import pandas as pd

# Load the dataset
df = pd.read_csv("dataset.tsv", sep="\t")

# Concatenate the first 5 fields into a single document
df['Content'] = df[['artist_name', 'track_name', 'release_date', 'genre', 'lyric']]

# Remove any duplicate or missing values
df = df.drop_duplicates()
df = df.dropna()
```

#### Step 2: Define preprocessing function with options

In [261...

```
import nltk
import re
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import PorterStemmer, WordNetLemmatizer
from sklearn.feature_extraction.text import CountVectorizer, ENGLISH_STOP_WORDS

nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')

ps = PorterStemmer()
```

```

wnl = WordNetLemmatizer()

# Define the text preprocessing function with option type parameters
def preprocess_text(text, stopwords_type, stem_or_lemma, regex_type, use_lowercase):

    if use_lowercase:
        text = text.lower()

    if regex_type == 'strict':
        text = re.sub(r"^\w\s", '', text)
    elif regex_type == 'relaxed':
        text = re.sub(r"^\w\s'&:]", "", text)
    elif regex_type == 'relaxed_with_hyphen':
        text = re.sub(r"^\w\s'&:-]", "", text)

    tokens = word_tokenize(text)

    if stopwords_type == 'nltk':
        stop_words_set = set(stopwords.words('english'))
    else:
        stop_words_set = ENGLISH_STOP_WORDS

    tokens = [word for word in tokens if word not in stop_words_set]

    if stem_or_lemma == 'stem':
        tokens = [ps.stem(word) for word in tokens]
    else:
        tokens = [wnl.lemmatize(word) for word in tokens]

    return ' '.join(tokens)

```

```

[nltk_data] Downloading package punkt to
[nltk_data]      /Users/samharrison/nltk_data...
[nltk_data] Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]      /Users/samharrison/nltk_data...
[nltk_data] Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to
[nltk_data]      /Users/samharrison/nltk_data...
[nltk_data] Package wordnet is already up-to-date!

```

### Step 3: Evaluate combinations of preprocessing options

In [262...

```

from sklearn.naive_bayes import BernoulliNB, MultinomialNB
from sklearn.model_selection import cross_val_score
from sklearn.preprocessing import LabelEncoder
from itertools import product

label_encoder = LabelEncoder()
y = label_encoder.fit_transform(df['topic'])

# Define possible options for each preprocessing step
stopwords_options = ['nltk', 'sklearn']
shortening_options = ['stem', 'lemma']
regex_options = ['strict', 'relaxed', 'relaxed_with_hyphen']
lowercase_options = [True, False]
model_options = [('MNB', MultinomialNB(), False), ('BNB', BernoulliNB(), True)]

results = []

```

```

# Evaluate all combinations of preprocessing options
for stopwords_type, shortening_type, regex_type, use_lowercase in product(stopwords_type,
                                   shortening_type, regex_type, use_lowercase):
    processed = df['Content'].apply(lambda x: preprocess_text(x, stopwords_type,
                                                              shortening_type,
                                                              regex_type, use_lowercase))

    for model_name, model, binary in model_options:
        vectorizer = CountVectorizer(binary=binary)
        X = vectorizer.fit_transform(processed)

        acc = cross_val_score(model, X, y, cv=5, scoring='accuracy').mean()

        results.append({
            'model': model_name,
            'stopwords': stopwords_type,
            'shortening': shortening_type,
            'regex': regex_type,
            'lowercase': use_lowercase,
            'accuracy': acc
        })

results_df = pd.DataFrame(results)

top_mnb = results_df[results_df['model'] == 'MNB'].nlargest(10, 'accuracy')
top_bnb = results_df[results_df['model'] == 'BNB'].nlargest(10, 'accuracy')

print('Top 10 MNB configurations')
display(results_df[results_df['model'] == 'MNB'].nlargest(10, 'accuracy'))

print('Top 10 BNB configurations')
display(results_df[results_df['model'] == 'BNB'].nlargest(10, 'accuracy'))

```

Top 10 MNB configurations

	model	stopwords	shortening	regex	lowercase	accuracy
12	MNB	nlTK	lemma	strict	True	0.790541
14	MNB	nlTK	lemma	strict	False	0.790541
16	MNB	nlTK	lemma	relaxed	True	0.790541
18	MNB	nlTK	lemma	relaxed	False	0.790541
20	MNB	nlTK	lemma	relaxed_with_hyphen	True	0.789865
22	MNB	nlTK	lemma	relaxed_with_hyphen	False	0.789865
0	MNB	nlTK	stem	strict	True	0.785811
2	MNB	nlTK	stem	strict	False	0.785811
4	MNB	nlTK	stem	relaxed	True	0.785811
6	MNB	nlTK	stem	relaxed	False	0.785811

Top 10 BNB configurations

	model	stopwords	shortening	regex	lowercase	accuracy
25	BNB	sklearn	stem	strict	True	0.527703
27	BNB	sklearn	stem	strict	False	0.527703
29	BNB	sklearn	stem	relaxed	True	0.527027
31	BNB	sklearn	stem	relaxed	False	0.527027
1	BNB	nltk	stem	strict	True	0.526351
3	BNB	nltk	stem	strict	False	0.526351
9	BNB	nltk	stem	relaxed_with_hyphen	True	0.525000
11	BNB	nltk	stem	relaxed_with_hyphen	False	0.525000
33	BNB	sklearn	stem	relaxed_with_hyphen	True	0.525000
35	BNB	sklearn	stem	relaxed_with_hyphen	False	0.525000

### Selected Preprocessing Configuration

The results displayed above indicate that MNB achieves significantly greater accuracy than BNB. As such, it is best to optimise the preprocessing configuration for the MNB classifier. Based on the accuracy scores of the different preprocessing configurations for the MNB classifier, the NLTK stopwords and Lemmatization will be used. Whilst the strict and relaxed regex (without hyphens) result in the same accuracy, given the dataset consists of curated song metadata with minimal redundant information, the relaxed regex should be used to preserve meaningful symbols.

- **Regex:** Relaxed (apostrophes, ampersands and colons) [`^\w\s'&:`]
- **Lowercasing:** True
- **Stopword removal:** NLTK stopwords
- **Word processing:** Lemmatization

### Step 4: Apply the selected preprocessing steps to the dataset

In [263...

```
df['Processed'] = df['Content'].apply(
    lambda x: preprocess_text(x, 'nltk', 'lemma', 'relaxed', True)
)

print(df['Processed'].head())
```

```
0    loving real lake 2016 rock awake know go see t...
1    incubus summer 2019 rock summer pretty build s...
2    reignwolf hardcore 2016 blue lose deep catch b...
3    tedeschi truck band anyhow 2016 blue run bitte...
4    lukas nelson promise real started 2017 blue th...
Name: Processed, dtype: object
```

### Question 3

#### Step 1: Assess the balance of the dataset

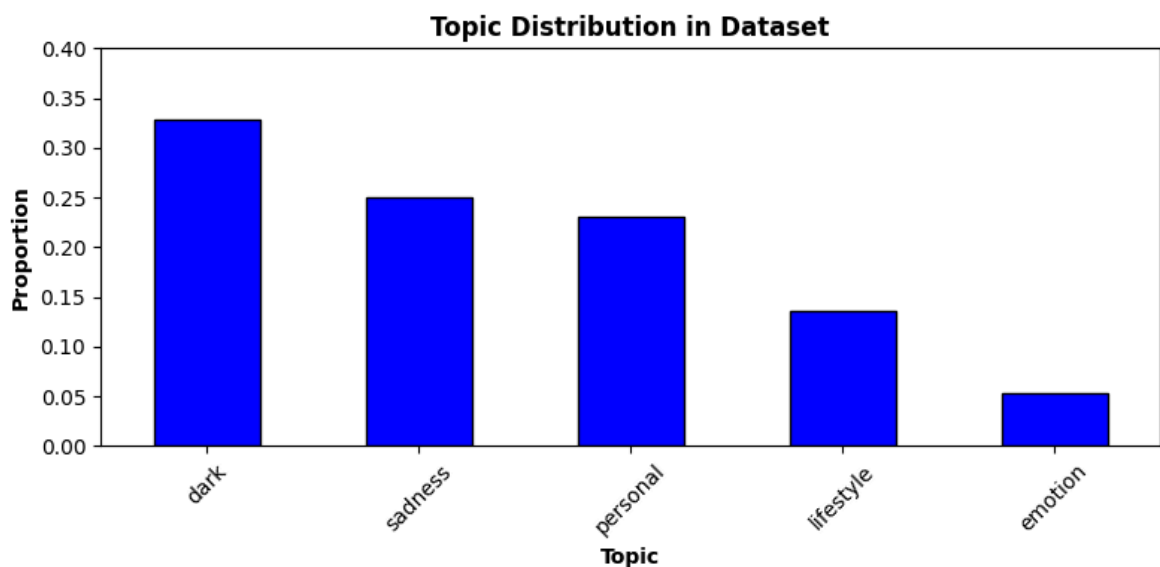
In [264...

```
import matplotlib.pyplot as plt

topic_distribution = df['topic'].value_counts(normalize=True)

plt.figure(figsize=(8, 4))
topic_distribution.plot(kind='bar', color='b', edgecolor='black')
plt.title('Topic Distribution in Dataset', fontweight='bold')
plt.ylabel('Proportion', fontweight='bold')
plt.xlabel('Topic', fontweight='bold')
plt.ylim(0, 0.4)
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()

topic_imbalance_ratio = topic_distribution.max() / topic_distribution.min()
print(f'Imbalance ratio: {topic_imbalance_ratio}')
```



Imbalance ratio: 6.1645569620253164

### Select metrics to evaluate overall accuracy of BNB and MNB

In [265...

```
from sklearn.model_selection import cross_val_predict
from sklearn.metrics import classification_report

vectorizer = CountVectorizer(binary=False)
X = vectorizer.fit_transform(df['Processed'])
y = label_encoder.fit_transform(df['topic'])

mnb = MultinomialNB()
mnb_preds = cross_val_predict(mnb, X, y, cv=5)

print('MultinomialNB Classification Report')
print(classification_report(y, mnb_preds, target_names=label_encoder.classes_))

bnb = BernoulliNB()
X_bnb = CountVectorizer(binary=True).fit_transform(df['Processed'])
bnb_preds = cross_val_predict(bnb, X_bnb, y, cv=5)

print('BernoulliNB Classification Report')
print(classification_report(y, bnb_preds, target_names=label_encoder.classes_))
```

MultinomialNB Classification Report				
	precision	recall	f1-score	support
dark	0.81	0.83	0.82	487
emotion	0.50	0.28	0.36	79
lifestyle	0.85	0.69	0.76	202
personal	0.84	0.81	0.83	341
sadness	0.74	0.89	0.81	371
accuracy			0.79	1480
macro avg	0.75	0.70	0.71	1480
weighted avg	0.79	0.79	0.78	1480

BernoulliNB Classification Report				
	precision	recall	f1-score	support
dark	0.60	0.75	0.67	487
emotion	0.00	0.00	0.00	79
lifestyle	0.08	0.00	0.01	202
personal	0.67	0.29	0.41	341
sadness	0.44	0.84	0.58	371
accuracy			0.52	1480
macro avg	0.36	0.38	0.33	1480
weighted avg	0.47	0.52	0.46	1480

Based on these results, MNB is clearly superior to BNB on precision, recall and F1 score and should therefore be used.

#### Question 4

##### Total number of features in the dataset

```
In [266... vectorizer = CountVectorizer()
X = vectorizer.fit_transform(df['Processed'])

total_features = len(vectorizer.vocabulary_)

print(f'Total number of features: {total_features}')
```

Total number of features: 9437

##### Test different values for N

```
In [267... n_values = [100, 250, 350, 500, 1000, 2000, 4000, total_features]

results = []

for n in n_values:
    vectorizer = CountVectorizer(max_features=n, binary=True)
    X_bnb = vectorizer.fit_transform(df['Processed'])
    acc_bnb = cross_val_score(BernoulliNB(), X_bnb, y, cv=5, scoring='accuracy')

    vectorizer = CountVectorizer(max_features=n, binary=False)
    X_mnb = vectorizer.fit_transform(df['Processed'])
    acc_mnb = cross_val_score(MultinomialNB(), X_mnb, y, cv=5, scoring='accuracy')

    results.append({'model': 'BNB', 'num_features': n, 'accuracy': acc_bnb})
```

```

    results.append({'model': 'MNB', 'num_features': n, 'accuracy': acc_mnb})

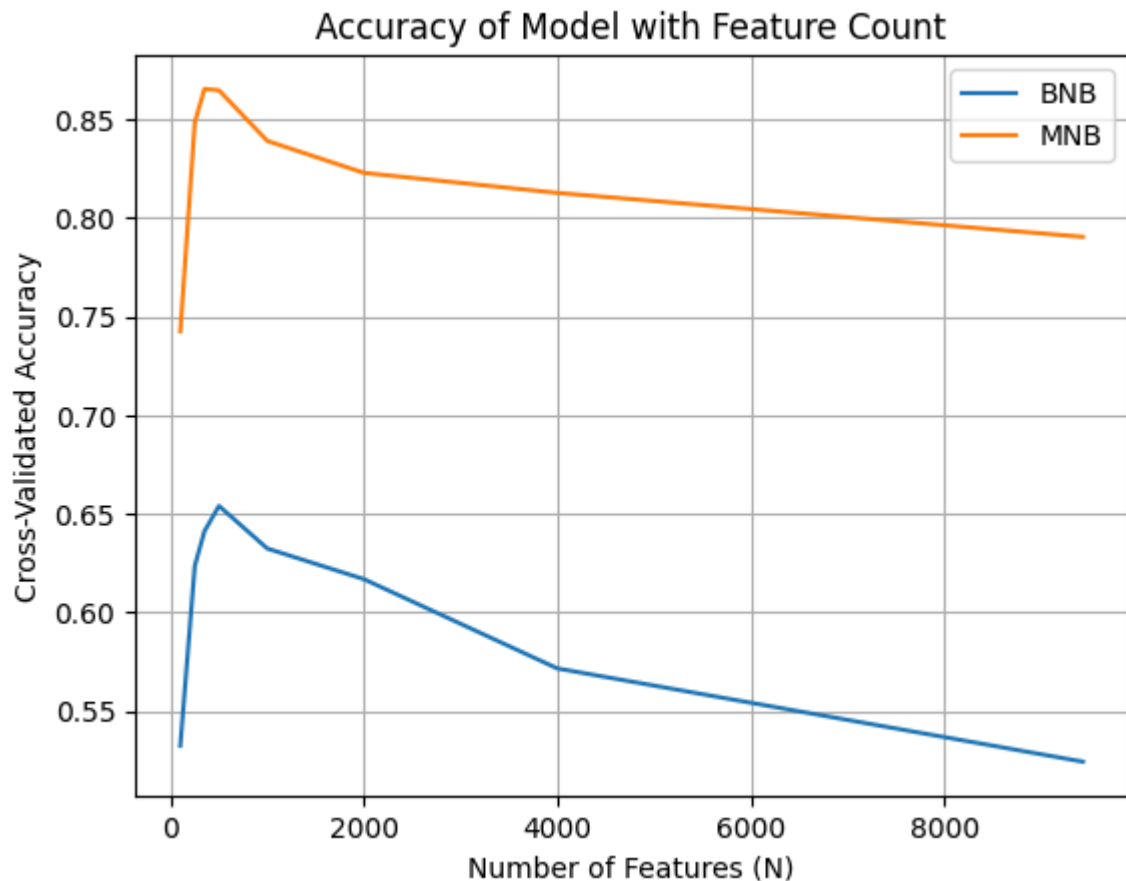
results_df = pd.DataFrame(results)
display(results_df)

for model in ['BNB', 'MNB']:
    subset = results_df[results_df['model'] == model]
    plt.plot(subset['num_features'], subset['accuracy'], label=model)

plt.xlabel('Number of Features (N)')
plt.ylabel('Cross-Validated Accuracy')
plt.title('Accuracy of Model with Feature Count')
plt.legend()
plt.grid(True)
plt.show()

```

	model	num_features	accuracy
0	BNB	100	0.532432
1	MNB	100	0.742568
2	BNB	250	0.623649
3	MNB	250	0.848649
4	BNB	350	0.641216
5	MNB	350	0.865541
6	BNB	500	0.654054
7	MNB	500	0.864865
8	BNB	1000	0.632432
9	MNB	1000	0.839189
10	BNB	2000	0.616892
11	MNB	2000	0.822973
12	BNB	4000	0.571622
13	MNB	4000	0.812838
14	BNB	9437	0.524324
15	MNB	9437	0.790541



The results of the analysis above indicates that the cross-validated accuracy is maximised with a value for N of approximately 350.

### Question 5

Support Vector Machines are supervised learning algorithms that are often used for classification tasks, given they are highly effective even when using vectors with sparse features such as those generated by TF-IDF. SVM's work by identifying the hyperplane that is most effective at separating the data into different classes, with the distance between the closest points of each class being maximised. My hypothesis is that SVM will outperform both BNB and MNB due to it's superior ability to distinguish between classes in sparse datasets like a collection of songs.

In [268...

```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.svm import LinearSVC

vectorizer = CountVectorizer(max_features=350)
X = vectorizer.fit_transform(df['Processed'])
y = label_encoder.fit_transform(df['topic'])

svc = LinearSVC(max_iter=2500)
svc_preds = cross_val_predict(svc, X, y, cv=5)

print('LinearSVC Classification Report')
print(classification_report(y, svc_preds, target_names=label_encoder.classes_))
```



LinearSVC Classification Report				
	precision	recall	f1-score	support
dark	0.85	0.88	0.87	487
emotion	0.72	0.68	0.70	79
lifestyle	0.80	0.77	0.78	202
personal	0.86	0.84	0.85	341
sadness	0.85	0.86	0.85	371
accuracy			0.84	1480
macro avg	0.82	0.81	0.81	1480
weighted avg	0.84	0.84	0.84	1480

The results of the cross-validation confirm my hypothesis that SVM is superior to both BNB and MNB on measures of precision, recall and F1 Score.

## Part 2: Recommendation Methods

### Question 1

```
In [269... model = MultinomialNB()

model.fit(X[:750], y[:750])

predicted_topics = model.predict(X)

df['PredictedTopic'] = label_encoder.inverse_transform(predicted_topics)
```

```
In [270... from sklearn.feature_extraction.text import TfidfVectorizer

user_profiles = {}

def create_user_profile(user_keywords_dict):

    liked_songs_in_topic = {topic: [] for topic in user_keywords_dict.keys()}

    for i in range(750):
        row = df.iloc[i]
        topic = row['PredictedTopic']

        if topic in user_keywords_dict:
            for keyword in user_keywords_dict[topic]:
                if keyword in row['lyrics']:
                    liked_songs_in_topic[topic].append(row['Processed'])
                    break

    user_likes_for_topic = {
        topic: ' '.join(docs) for topic, docs in liked_songs_in_topic.items()
    }

    profile_vectorizer = TfidfVectorizer()

    user_profile_matrix = profile_vectorizer.fit_transform(user_likes_for_topic)

    user_profile_df = pd.DataFrame(user_profile_matrix.toarray(), index=list(user_profiles.keys()),
                                   columns=profile_vectorizer.get_feature_names_out())
```

```

    return user_profile_df

for username in ['user1', 'user2']:

    user_keywords = pd.read_csv(f'{username}.tsv', sep='\t')

    user_keywords_dict = {
        row['topic']: [keyword.strip().lower() for keyword in row['keywords'].split()
                       for _, row in user_keywords.iterrows()]
    }

    user_profile_df = create_user_profile(user_keywords_dict)

    user_profiles[username] = user_profile_df

    print(f'Top 20 words for topic in {username} profile', end='\n\n')

    for topic in user_profile_df.index:
        top_words = user_profile_df.loc[topic].sort_values(ascending=False).head(20)
        print(f'{topic}')
        print(', '.join(top_words.index), end='\n\n')

```

Top 20 words for topic in user1 profile

dark

fight, blood, know, like, come, grind, na, stand, yeah, black, tell, kill, head, gon, hand, drown, time, build, dilly, lanky

sadness

cry, club, tear, know, steal, mean, think, baby, say, true, leave, eye, music, feel, write, want, heart, smile, face, time

personal

life, live, know, world, na, yeah, change, like, time, dream, wan, come, thank, teach, need, thing, think, lord, learn, feel

lifestyle

night, song, come, time, tonight, long, home, na, right, sing, yeah, wan, spoil, wait, like, closer, stranger, know, tire, mind

emotion

good, feel, touch, know, hold, loove, kiss, video, want, go, gim, morning, lovin, miss, vision, lip, like, baby, time, yeah

Top 20 words for topic in user2 profile

sadness

heart, inside, break, away, like, step, know, fall, open, tear, blame, come, leave, go, na, fade, think, feel, scar, violence

emotion

touch, good, kiss, hold, vision, video, loove, morning, luck, lovin, lip, sunrise, feel, know, gim, time, knock, foot, addiction, go

## Top 20 words evaluation

The top 20 words per topic for both User 1 and User 2 generally appear to be reasonable based on the keywords in the each user's dataset. The first few words in each user's top

20 are particularly relevant, with the later words appearing less so. This is to be expected given the dataset contains far less than 20 keywords per topic. Additionally, song lyrics often contain juxtapositions and other literary techniques that can make it difficult to assess the true meaning of a song purely based on each word viewed in isolation.

### Create User 3 profile

In [271...

```
user3_keywords_dict = {}

user3_keywords_dict['lifestyle'] = ['dance', 'weekend', 'party']
user3_keywords_dict['emotion'] = ['smile', 'heartbeat', 'joy', 'comfort']
user3_keywords_dict['personal'] = ['journey', 'change', 'soul']
```

In [272...

```
user3_profile_df = create_user_profile(user3_keywords_dict)

user_profiles['user3'] = user_profile_df

print(f'Top 20 words for topic in User 3 profile', end='\n\n')

for topic in user3_profile_df.index:
    top_words = user3_profile_df.loc[topic].sort_values(ascending=False).head(20)
    print(f'{topic}')
    print(', '.join(top_words.index), end='\n\n')
```

Top 20 words for topic in User 3 profile

lifestyle

song, night, na, spoil, time, country, tire, wan, tonight, ready, come, home, play, yeah, ring, like, bring, dance, need, charmer

emotion

hold, love, oohoh, dance, one, heart, faceless, low, speak, high, crowd, fluently, light, matter, lead, yeah, mean, baby, smile, sure

personal

change, life, world, know, live, na, day, yeah, like, time, way, come, mind, feel, dream, gon, away, go, grow, think

## Question 2

### Choice of N (Number of songs shown in total)

I have chosen to use an N value of 15, therefore showing 15 songs to the user in total across all topics. This number is based on the length of time a user is likely to want to spend listening to or reading new songs that they have been recommended before they are likely to prefer to listen to songs they already like. Additionally, within the 15 recommended songs it is likely that the user will "like" at least a couple, presuming the recommender performs at a reasonable level.

### Choice of recommendation algorithm

I have chosen to use cosine similarity to compute the top N song recommendations for each user. Cosine similarity is particularly useful for this task as it measures the angle between two vectors rather than their magnitude, which results in recommendations

based on the relative importance of the words in the user profile and song, rather than simply the count of each word.

## Metrics for evaluating performance of recommendations

The primary metric for evaluating the performance of my recommender is precision, which provides an indication of the quality of the recommendations by measures the proportion of the songs which the user liked out of all the songs that were recommended to them. Recall is another important metric which measures how many of the songs the user would indeed like were recommended to them. This helps to assess the coverage of the recommendations. F1 score is a harmonic mean of both precision and recall, providing a balance between the quality and coverage of the recommender.

```
In [273... from sklearn.metrics.pairwise import cosine_similarity
from sklearn.feature_extraction.text import TfidfVectorizer
import numpy as np
import pandas as pd

N = 15
M_values = [10, 20, 50]

def evaluate_user(user_keywords_dict, user_profile_df, M_values):
    results = []
    test_df = df.iloc[750:1000]

    for M in M_values:
        total_true_positives = 0
        total_false_positives = 0
        total_false_negatives = 0

        for topic in user_profile_df.index:
            topic_songs = test_df[test_df['PredictedTopic'] == topic]
            if topic_songs.empty:
                continue

            top_m_words = user_profile_df.loc[topic].sort_values(ascending=False)
            keywords = user_keywords_dict.get(topic, [])
            if len(top_m_words) == 0:
                continue

            vectorizer = TfidfVectorizer(vocabulary=top_m_words.index)
            song_vectors = vectorizer.fit_transform(topic_songs['Processed'])

            profile_vector = top_m_words.values.reshape(1, -1)
            profile_vector /= np.linalg.norm(profile_vector)
            similarities = cosine_similarity(song_vectors, profile_vector).flatten()

            top_indices = np.argsort(similarities)[-N:][::-1]
            recommended = topic_songs.iloc[top_indices]

            true_positives = 0

            for _, row in recommended.iterrows():
                if any(kw in row['lyrics'].lower() for kw in keywords):
                    true_positives += 1

            false_positives = N - true_positives
```

```

        matching_songs = topic_songs[
            topic_songs['lyrics'].apply(lambda text: any(kw in text.lower()
        ])
        false_negatives = len(matching_songs) - true_positives

        total_true_positives += true_positives
        total_false_positives += false_positives
        total_false_negatives += false_negatives

    precision = total_true_positives / (total_true_positives + total_false_positives)
    recall = total_true_positives / (total_true_positives + total_false_negatives)
    f1 = 2 * precision * recall / (precision + recall) if (precision + recall) > 0 else 0

    results.append({
        'M': M,
        'precision': precision,
        'recall': recall,
        'f1': f1
    })

return pd.DataFrame(results)

for username in ['user1', 'user2']:
    user_keywords = pd.read_csv(f'{username}.tsv', sep='\t')
    user_keywords_dict = {
        row['topic']: [kw.strip().lower() for kw in row['keywords'].split(',')]
        for _, row in user_keywords.iterrows()
    }

    profile_df = user_profiles[username]
    result_df = evaluate_user(user_keywords_dict, profile_df, M_values)

    print(f'{username} results')
    display(result_df)

```

user1 results

	M	precision	recall	f1
0	10	0.706667	0.609195	0.654321
1	20	0.706667	0.609195	0.654321
2	50	0.666667	0.574713	0.617284

user2 results

	M	precision	recall	f1
0	10	0.200000	0.545455	0.292683
1	20	0.166667	0.454545	0.243902
2	50	0.200000	0.545455	0.292683

The results of the recommendations clearly show that the recommendation algorithm performs far better with User 1 than User 2 on precision, recall and F1 score. This is likely due to User 1 having a much larger dataset of keywords from which their "liked" songs

were initially determined during training, resulting in a more accurate profile that was used to generate the recommendations.

The appropriate M value based on these results is 10, as this results in the highest precision, recall and F1 scores for both User 1 and User 2.

## Part 3: User Evaluation

### Week 1

```
In [ ]: # Choose 15 random songs from 1-250
week1_sample = sampled_df = df.iloc[:250].sample(n=15, random_state=1)

# Print the lyrics to each song for user feedback
for i in range(15):
    print(f'Song {i + 1}: {week1_sample.iloc[i]['lyrics']}')
```

SONG 1: stay inside livin life instead watchin window playin safe gonna hesitate know mental let givin control yeah ready wind blow wanna road gonna lead time tickin world waitin dodgin raindrops long cloud head dodgin raindrops tryin dodgin raindrops long dodgin raindrops head head head head sooner later come quick come blockin moment run away past pushin zone pullin unknown let givin control yeah ready wind blow wanna road gonna lead time tickin world waitin dodgin raindrops long cloud head dodgin raindrops tryin dodgin raindrops long dodgin raindrops head head head dodgin raindrops head head dodgin raindrops head head ready outside livin life instead watch window dodgin raindrops long cloud head dodgin raindrops tryin dodgin raindrops long dodgin raindrops head head head head dodgin raindrops head head dodgin raindrops head head

SONG 2: cities gold hold promise ruin fountain youth illusion glass half till remember forget life lesson action pain weapon sorrow attraction heartbeat know house build home feel come force spring porch mean thing sheet thread lonely stop remember forget life lesson action pain weapon sorrow attraction heartbeat know house build home feel whoooah stay calm light realise life lesson action pain weapon attraction heartbeat know house build home feel house build home feel

SONG 3: mind know want plan offer try help load little lighter share wealth good luck good luck know problems eye want live morning brother surprise good luck good luck yeah yeah yeah yeah yeah yeah yeah yeah yeah time go receive time want believe breathe good luck good luck good luck good luck

SONG 4: stay late night nothin drink help drink mind look oooooo face wonder wonder little space heart cause cold miss babe come rescue shake pain soul fool heart baby crazy cause mess pill pain come rescue baby right away fool leave leave mean night mmmmmm heart pay strand lonely road lose lead home need baby miss babe come rescue shake pain soul fool heart baby crazy cause mess pill pain come rescue baby right away drown baby need save need save yeah leave like throw line baby time come rescue shake pain soul fool heart baby crazy cause mess pill pain come rescue baby right away right away right away right away

SONG 5: excuse death dishonor flame reach stand alter preach choir lord streets slaughter excuse death dishonor flame reach stand alter preach choir lord streets slaughter excuse death dishonor drug dealers pill poppers real partner mobbin partner poppin feel come killin year feel like hell leave like believe lean size clip know feds come pick negotiatin goin months partner blow trial game right tell judge right give life flame reach stand alter preach choir lord streets slaughter excuse death dishonor flame reach stand alter preach choir lord streets slaughter excuse death dishonor bear gangster father heart daughter start hustle smart plug ones trust ones stab ones close toast time pledge family loyalty strong suit curse want prosper want spot playin like actin costner imposter jealous hearts conquer appear strong shots revolver dress like robber half million dollars reign karma click death dishonor flame reach stand alter preach choir lord streets slaughter excuse death dishonor flame reach stand alter preach choir lord streets slaughter excuse death dishonor

SONG 6: devil think excuse slip fell outta hand devil think wasn hold head catch believe lie say moment weakness stumble devil think excuse slip fell outta hand devil think reason stay blame want time devil think excuse slip fell outta hand devil think spend forever prove devil think devil

SONG 7: corner store smell like pall malls cashier wear overalls cop blood hand moonshine take moonlight dirty south yeayeayeayah dirty south dirty south dirty south cars burn dark lead know dead body tree soul dead roads dead roads lead small trail headstones ghost lurk feel chill spine pin dirty south yeayeayeayah dirty south dirty south dirty south dirty south raise dirty south dirty south hate time hate dirty south yeayeayeayah dirty south dirty south dirty south woahoh woahoh woahohoh dirty south dirty south dirty south dirty south dirty south

SONG 8: cross bridge like leviathan travel coast tie whip post get high cross bridge like leviathan travel coast tie whip post get high yeah yeah river sink boat soft sound violin follow ghost deliver shadow hide river sink boat soft sound violin follow ghost deliver shadow hide cross bridge like leviathan travel coast tie whip post get high cross bridge like leviathan travel coast tie whip post get high yeah trust yeah thief night yeah lose touch yeah thief night beast rain river s

ink boat lose follow ghost deliver shadow hide river sink boat lose hard road felt invisible break silence trust power abstain thief night beast rain lose touch regain search truth excuse shame cross bridge like leviathan travel coast tie whip post get high cross bridge like leviathan travel coast tie whip post get high

SONG 10: think think different set apart sober mind sympathetic hearts swear oath  
swear oath know life play play distance great height sure kill innocents build dr  
op death longer human be longer people target screen real drop death tell send se  
nd chaplain say say say shut mouth sure kill innocents build drop death longer hu  
man be longer people target screen real drop death tell tell safe shoot dream jus  
tify resolve rise avenge innocent defend ones survive ask

SONG 12: fight thoughts suicide alive comprise give realize life homie time doubt  
cause dream mind work time get closer happiness inside aren movement surely leave  
back falcon punch game leave ash shed tear cause legacy mixtapes couple single co  
uple fan cool life get pretty hard feel like give start relight passion heart kno  
w life thankful oldest dream come conquer supreme wait money purchase things real  
ity dangerous wish fall asleep life quick women aren speed fall dream mean bless  
life teach lessons pretty thankful live future hand drop guess learn spill solely  
focus make kill track multisyllable verse provide music crack sixteens pointless  
rhyme favorite album okay like music ahead dance prowl freakin problem nerdy love  
cartoon ahead doubt like flame instrumentals word rhythms slowly captivate mental  
southern rapper different kill beat guess childish different believe stop chase d  
ream

SONG 14: bull fight lead gonna fight gonna fight brother feelin like superman col  
d beer hand waitin week long jukebox shadowbox nothin lipstick shoot little fireb  
all shoot little ball wall crowd need smoke toke cause tonight burnin step right  
pay little cover drink bunch redneck mother light lead gonna fight gonna fight br  
other gonna fight gonna fight brother yonder come plow trash talkin cowboy actin  
like feet tall kissin ontrea hell date bathroom stall roof higher proof howlin li  
ke bluetick hound steady beer flow go circus come yeah step right pay little cove  
r outdrink bunch redneck mother light lead gonna fight gonna fight brother gonna  
fight gonna fight brother goin fine till moonshine spill drink know better hell k  
now better rared take swing say step right pay little cover outdrink bunch rednec  
k mother light lead gonna fight gonna fight brother gonna fight gonna fight broth  
er bull fight lead gonna fight gonna fight brother



SONG 15: crabb tell loneliness need escape reality free crabb thompson away ocean shore away ocean shore feel dream crabb look place life live abundantly perfect crabb thompson away spirit soar lonely nevermore look away tide roll soon begin laughter solo thompson loneliness want free crabb world beckon despite maybe go temporarily crabb thompson away ocean shore life mean away tide roll laughter

The user liked songs 1, 3 and 13.

## Week 2

```
In [ ]: # Choose 15 random songs from 251-500
week2_sample = sampled_df = df.iloc[250:500].sample(n=15, random_state=1)

# Print the Lyrics to each song for user feedback
for i in range(15):
    print(f'Song {i + 1}: {week2_sample.iloc[i]['lyrics']}')
```

SONG 1: heaven heaven mater nights raise sword holy blood burn pyre defend command holy mater stand fight deus regnum heaven heaven brother believe sword nights blood thunder strike reward come kill plunder fate crusade holy mater stand fight deus regnum heaven follow follow mater mater deus regnum sanctus christus deus vult sanctus iesus deus vult sanctus christus deus vult cantus lupus agnus christus sanctus lupus

SONG 2: people scheme dream people live live look world understand tenement flat second land rain today care heart daydream share people scheme dream people live live believe roll take laugh hesitation know dreamer sort follow road course people scheme dream people live live good people scheme dream people live live people people

SONG 3: million piece breathe hard speak cause like bitter pill blow mind make heart beat harder harder harder harder hard clown laugh drill look outside world end faster faster faster faster fast echo news ring loud echo echo sound drown echo break heart break heart million piece break heart million piece gonna break leave till morning wanna know break heart break heart million piece million piece go mean drink dance right disaster want talk echo news ring loud echo echo music drown echo break heart break heart million piece break heart million piece gonna break leave till morning wanna know break heart break heart million piece million piece fear loathe mind pull pull ceiling break heart break heart million piece break heart million piece gonna break leave till morning wanna know break heart break heart million piece million piece million piece million piece

SONG 4: think weight chest night say couldn't think break wouldn't better start think easy hat cause afraid hurt think good goodbye heart have hard time hard time make sense right inside yeah know mind heart heart have hard time shouldn't tough change lock screen sink kiss smile million reason leave think feel feel hat cause afraid hurt think good goodbye heart have hard time make sense right inside yeah know mind heart heart have hard time miss wish work wrong love hat cause afraid hurt think good goodbye heart have hard time make sense right inside yeah know mind heart heart have hard time heart have hard time

SONG 5: pick sewer train future come inside inside deep inside body behavior turn past blue sky past dark sky hand move count rat automatic spell pressure cold care stand line know centric fashion mall body behavior turn past blue sky past dark sky hand move hand move hand move

SONG 6: nothin' world somethin' nothin' thank beautiful girl bring world yeah remember leave home cadillac search kinda proud money understand sacrifice late nights fuss fight home sorry things wrong remember leave home cadillac search kinda proud wanna wanna proud nothin' world somethin' nothin' thank beautiful girl world yeah

SONG 7: stop baby tear asunder shove go live stuff plunder come strong roll thunder easy rifle easy time world thunder move world asunder glue dude gun look wonder greasy moment crack wall wonder goodbye know blunder goodbye holler awful daddy easy run world go goodbye daddy young ones leave ratty ratty world stop baby tear asunder shove go live stuff plunder come strong roll thunder

SONG 8: weight shoulder think easier heart grow colder warmth kiss dismiss past leave bruise hide truth truth right hold hold strong try play fake keep hold hold strong try break heaven know shake hold hold hold hold see time harder remember taste bitterness help father help fall miss past leave bruise hide truth truth right hold hold strong try play fake keep hold hold strong try break heaven know shake whoa whoa hold hold hold hold strong try play fake keep hold hold strong try break heaven know shake hold hold keep hold keep hold keep hold hold hold hold hold hold hold hold

SONG 9: buju banton yout brownin song hear tell cause cyan shoe tongue grun hear stop black woman girls dark complexion cause stop black woman nuff ting gwan complexion black beauty colla million birth natural smooth lika grape true lotion teki teki care complexion whola plan wrong bcaw black woman stop black woman girls dark complexion cause stop black woman nuff ting gwan complexion true lite skin want woman baddah worry intension wedah black buju right spread nation backitive buju banton stop black woman girls dark complexion cause stop black woman nuff ting gwan complexion black proud folla buju banton shout loud black stand crowd like line behind dark cloud stop black woman girls dark complexion cause stop black woman

SONG 10: come eye smile shine days heart strong kindness face patiently take hand  
gently hold space like blossom silently radiate high butterfly gently high butter  
fly gently come hand tie know lick wound rebuild pride leave seed germinate need  
need rain locate trust time trust learn trust fate high butterfly gently high but  
terfly gently high butterfly gently high butterfly gently beneath canopy jungle f  
ace belly animals centipedes snake shipibo song fade road candle burn reach desti  
nation strength inside days answer lie deep inside natural mystic come

SONG 12: wait long days january sit say email today kind think forget wanna like  
lonely baby know time tell anybody lonely place face face rule like lonely like l  
onely like lonely ready steady like long friend wanna come strong cause people wa  
nna know gotta like lonely baby know time tell anybody lonely place face face rul  
e like lonely like lonely like lonely lonely baby know time tell anybody lonely p  
lace face face rule like lonely like lonely like lonely lonely lonely like

SONG 14: stubborn come settle stick rule ways gotta freedom gotta space meet girl swear start feel things felt spend time try figure hold start fall hold pride start build wall inside push away cause know leave know shake head tell wrong fool's song know cause know meet girl swear start feel things felt spend time try figure hold start fall hold pride start build wall inside push away cause know leave know best song repeat maybe like meet girl swear start feel things felt spend time try figure hold start fall hold pride start build wall inside push away cause know leave know

The user liked songs 2, 6, 10 and 14.

```
In [ ]: # Choose 15 random songs from 501-750
week3_sample = sampled_df = df.iloc[500:750].sample(n=15, random_state=1)

# Print the Lyrics to each song for user feedback
for i in range(15):
    print(f'Song {i + 1}: {week3_sample.iloc[i]['lyrics']}')
```

SONG 1: drifters high rift plain ash acid rain turn dust eye choke death smog lie  
look sky darkness realize kill fear power lie hypnotize turn clock glass sand tim  
e blacken land silent child climb mound char plant seed grow star look sky darkne  
ss realize kill fear power lie hypnotize look sky darkness realize kill fear powe  
r lie hypnotize look sky darkness realize kill fear power lie hypnotize

SONG 2: golden magic yeah know name middle madness neon grey crowd yeah write sto  
ry blood sweat heartbeats fame go history yeah legends love baby heaven wonder qu  
estion close eye take world remember crazy tragic epic amaze wear crown give stay  
lose forever remember legends like write permanent marker brightest fade come whi  
chever hell high water write story blood sweat heartbeats fame legends love baby  
heaven wonder question close eye take world remember crazy tragic epic amaze wear  
crown give stay lose forever remember legends woah legends love baby heaven wonde  
r question close eye take world remember baby crazy tragic epic amaze wear crown  
give stay lose forever remember legends yeah legends yeah write story

SONG 3: hearts valve pumpin blood flood body live life stickin lettin makin right  
season change life grow dream temporary slide know bird different directions suns  
ets sunrises livin dream watchin leave lyric commercial

SONG 4: wrong right come away say perfect place like paradise ask speak like game  
hide seek take control paradise give little leave middle reminisce think leave de  
ceive betray slay think leave deceive betray slay tell enemy string strangle drea  
m leave scar paradise crack mirror wall take call suddenly go fault paradise impo  
ssible unreasonable think leave deceive betray slay think leave deceive betray sl  
ay give little leave middle reminisce think leave deceive betray slay think leave  
deceive betray slay

SONG 5: daddy pontiac beigeer yellow young little fella play bench seat listen so  
ngs sing couldn tune bucket sing lungs wayward hippie radio songs baby birth roll  
band stand bounce road pontiac hippie radio remember seventeen maybe eighteen rig  
ht forget baby blue glow dashboard light heart week right start bust brother mout  
h makin white wed rebel yell hippie radio werewolf london lady marmalade soul cra  
nk band hand pull road girl pontiac hippie radio years seven days tie can bumper  
pace maternity floor baby mother hand shake leave take home heart head hear long  
long song cradle silver spoon hippie radio blink go take hand smile look say know  
girl pontiac hippie radio

SONG 6: tell lonely cause selfish hold cause mend helpless thing song sing mind c  
hange space stay thing fall fall float ocean drain feet know tear fall vain thing  
song sing mind change space stay thing fall fall wonder change heart wander quest  
ion cause help sleep longer fall

SONG 7: take lifetime lose lifetime yesterdays waste time hand turn sand fade win  
d cross line small crimes take fine fee friction right fine fee friction right lo  
se long fail follow place cross line small crimes take fine fee friction right fi  
ne fee friction right fine fee friction right fine fee friction right fine fee fr  
iction right fine fee friction right fine fee friction right

SONG 8: mornin walk rise drink like whiskey tree see thousand time branch watch a  
way wind beautiful world clear days breath break heart poundin chest beatin days  
forget give days days alive days livin days livin go fire coffee walk kitchen lik  
e tshirt kill kiss like time laugh look like lose mind say baby live know days br  
eath break heart poundin chest beatin days forget give days days alive days livin  
days livin like blue little bluer high little high feel miss days live beautiful  
world clear days start singin need reason world right clear eye blinkin heart gra  
teful give days yeah days alive days livin days livin livin like blue little blue  
r high little high feel missin days live days live

SONG 9: havana half heart havana take east nanana heart havana somethin bout mann  
ers havana walk doin come room say girls know forever minute summer night papa sa  
y malo feelin like know love leave feelin like tell nanananana havana half heart  
havana take east nanana heart havana heart havana havana graduate fresh campus fr  
esh east manners fresh east bump bumper like traffic quick girl like uncle shawt  
y cravin eatin wait shawty cakin bacon history makin homie homie point blank clos  
e range cost million gettin mula baby havana half heart havana take east nanana h  
eart havana heart havana havana nanana like nanana yeah babe like nanana yeah yea  
h like nanana yeah babe havana havana half heart havana yeah take east nanana hea



```

row = df.iloc[song_idx]
topic = row['PredictedTopic']
liked_songs_in_topic[topic].append(row['Processed'])

for song_idx in week3_liked:
    row = df.iloc[song_idx]
    topic = row['PredictedTopic']
    liked_songs_in_topic[topic].append(row['Processed'])

user_likes_for_topic = {
    topic: ' '.join(docs) for topic, docs in liked_songs_in_topic.items()
}

profile_vectorizer = TfidfVectorizer()

user_profile_matrix = profile_vectorizer.fit_transform(user_likes_for_topic.values())

user_profile_df = pd.DataFrame(user_profile_matrix.toarray(), index=list(user_likes_for_topic.keys()),
                                columns=profile_vectorizer.get_feature_names_out())

```

### Present 15 song recommendations to the user

In [278...

```

N = 15

test_df = df.iloc[750:1000].copy()

vectorizer = TfidfVectorizer(vocabulary=user_profile_df.columns)
X_test = vectorizer.fit_transform(test_df['Processed'])

combined_user_vector = user_profile_df.sum(axis=0).values.reshape(1, -1)
combined_user_vector = combined_user_vector / np.linalg.norm(combined_user_vector)

similarities = cosine_similarity(X_test, combined_user_vector).flatten()

top_indices = np.argsort(similarities)[-N:][::-1]
recommended_songs = test_df.iloc[top_indices].copy()

for i in range(15):
    print(f'Recommended Song {i + 1}: {recommended_songs.iloc[i]['lyrics']}')

```

Recommended Song 1: say reason fell love walk float free fall hold free fall hold hold hold hold cold cold summer know hold hold tell ways want believe want believe hold know fall hold barely hold hold say reason fell love walk float free fall hold free fall hold

Recommended Song 2: tremble like earth fall feet like felt grind indication head know settle score lie tell know stack problems like pile bricks gotta knock wall like break put grind heavens spare crop winter fall hide reaper call go lead loam return hesitate grimace hold chin like think sentiment keep let know irresistible hear voice hush cold time peace mind buy sell cause like break grind heavens spare crop winter fall hide reaper call go lead loam return want fall apart pick piece finally build perfect pawn lose dark lose dark

Recommended Song 3: hear howl afar rush moment go scream wild darkness kill light remember run burn house tree remember run blind fear river flow beneath skin like savage horse keep waste sand like break diamonds hand remember run remember fall knees remember glide shore touch ocean floor river flow beneath skin like savage horse keep waste sand like break diamonds hand river grow inside river grow inside river grow inside river grow inside

Recommended Song 4: wish atlas path feel like go astray know long hard journey go wouldn't heart guide focus mission step time analyze vision fall line know gonna okay word ring true want feel spirit darkest days follow heart days hard remember purpose long learn journey season reason follow heart follow heart feel follow heart cause everybody look future want cause crumble tomorrow feel sorrow wonder cause know change near afraid move know grow gonna stop know feel like like remember stay strong weather follow heart days hard remember purpose long learn journey

Recommended Song 5: wanna smile want morning perform tell know lead place go skin demons drive heavenly limbs know beautiful half gold treasure soul cause want finger mouth fail fault fall fall think survive wild plant root dream electric fan baby kill look eye feel drain hand baby think past wonder stupid little thing know beautiful half gold treasure soul cause want finger mouth fail fault fall fall

Recommended Song 6: dream wish fantasy need breath truly madly deeply strong faithful cause count begin reason live deeper mean wanna stand mountain wanna bathe wanna like forever fall fall fall star shine brightly wish send heaven want tear pleasure certainty surround comfort protection highest power lonely hours tear devour baby close eye cause stand right need surely come wanna stand mountain wanna bathe wanna like forever fall wanna stand mountain wanna bathe wanna like forever fall fall fall fall fall

Recommended Song 7: crossfire fall fall fall fall defend good define give time crossroad better crossfire better crossfire crossroad better fall fall fall fall begin shootin want wanna crossroad better crossfire turn bitter serve crossroad better crossfire better crossfire crossroad better crossfire better crossfire better crossfire fall think truly special dream stay like baby baby like explain like explain

Recommended Song 8: heart like fortress feel lock away easier feel safe wander darkness wilderness eye know afraid gotta fortress fall fall rise say come fortress cave safe believe know truly know pray feel felt want fail tell awful truth face darkness fortress fall fall rise say fortress cave safe earth spin round behold happen night fall prey sunrise fortress siege fortress siege fortress fall fall rise say fortress cave safe fortress cave safe fortress cave safe come come

Recommended Song 9: hop pray feel pain save fear linger fall quicker look tell fine cause right feel like stick fade slowly fall faster yeah yeah fall faster yeah yeah days week sleep thoughts runnin deep control crashin know breakin crumb tryin drug takin control energy pullin holdin alternate reality reality hop pray feel pain save fear linger fall quicker look tell fine cause right feel like stick fade slowly fall faster fall faster need tree insecurity hold yeah feelin guess smoke blunt look pray want know gotta recenter actually yeah catch midair yeah start floatin hate choices hold grudge yeah hop pray feel pain save fear linger fall quicker look tell fine cause right feel like stick fade slowly fall faster fall faster

Recommended Song 10: calendar days gettin longer outta shade pretty girls water spin slow songs cause everybody want dance summer bring memory july magic bare feet radio crank doors jeep bull farm hood star livin large small teens wish freeze

time recover night summer summer night summer summer sunset smile midnight kiss y  
eah remember burnin backroads head september yeah better july magic bare feet rad  
io crank doors jeep bull farm hood star livin large small teens wish freeze time  
recover night summer summer night summer summer bridge night summer long july mag  
ic bare feet radio crank doors jeep bull farm hood star livin large small teens w  
ish freeze time recover night summer summer night summer summer outtro night summ  
er summer night summer summer source lyric power record info provide itunesapple  
music hambrick official youtube channel hambrick official soundcloud

Recommended Song 11: cry zero near somersault backflip pool summer summer pool po  
ol summer kiss hair want whiskey soda dip inflatables sink pool summer summer poo  
l pool summer kiss dive dive dive pool belong sink like beat stone crowdin round  
loud lyric commercial

Recommended Song 12: summer take neverland husband want voice little voice buzz p  
oison backward noise swango say psycho say neverland voice dirty say play ball di  
zzy dizzy speak whisper murder heart string break pull stretch infinitely summer  
silence get violent summer silence summer silence get violent summer silence play  
summer wait favourite cheshire grin dear even clear dream paperpale skin summer s  
ilence get violent summer silence summer silence get violent summer silence summe  
r silence get violent summer silence summer silence get violent summer silence su  
mmer silence summer silence summer silence

Recommended Song 13: bloodshot eye watch sleep warmth feel slowly fade hear call  
hold know shame different go wrong path walk wrong direction hang anybody help t  
hings better tear fall crash conscience call guilty come home tear fall crash con  
science call guilty come home moments hear scream visions leave inside slowly fad  
e hear call hold know shame different go wrong path walk wrong direction hang an  
ybody help things better tear fall crash conscience call guilty come home tear fa  
ll crash conscience call guilty come home yeah batter room see break bone heal br  
eath choke hop world time hear call hold know shame different go wrong path walk  
wrong direction hang anybody help things better tear fall crash conscience call  
guilty come home tear fall tear fall crash conscience call conscience call guilty  
come better tear fall crash conscience call conscience call guilty come home

Recommended Song 14: shame poison shame shame know unring try fall away hear exac  
tly make wonder break lord wonder go someday know long someday train come go reme  
mber pick right try fall away pain make wonder break lord wonder shame hurt sham  
e shame shame know murder truth try fall away leave wonder break wonder hell go w  
onder wonder care shame shame shame let shame shame shame shame shame shame

Recommended Song 15: walk home step gate chicken plate food cold cover face know  
want trouble bubble come trouble walk live room say gotta tell say wanna know thi  
ngs say gotta tell black blue say want trouble bubble come trouble heart pump che  
st scream mind run freeze hand hand tell ready fight punch face cause like pain t  
ime curse know want satisfaction gonna happen knock kick grind gonna come lightni  
ng thunder suffer suffer square leave chest expose throw quick leave hook break n  
ose blood run clothe sick sick look cause like say want trouble bubble come troub  
le heart pump chest scream mind run nose bleed hand hand tell ready fight punch f  
ace cause like pain time curse know want satisfaction gonna happen knock kick gri  
nd gonna come lightning thunder suffer walk home blood hand break nose like scar  
house cause pop home drown trouble whiskey bubble look trouble excuse things home  
deal drink bubble belt buckle break bubble punch face cause like pain time curse  
know want satisfaction gonna happen knock kick grind gonna come lightning thunder  
suffer suffer

The user liked songs 1, 4, 5, 6, 9 and 10.

### **Compute performance metrics**