Spotify Recommender Sys

DS300.M11

Team 5
Dương Văn Bình
Hà Như Chiến

Instructors:
Huỳnh Ngọc Tín
Huỳnh Văn Tín

Content

1. Data
2. Methodology
3. Experiments and Results

Problem and methodology explanation

Recommend tracks based on users’ historical listened tracks.

1. Content-based method

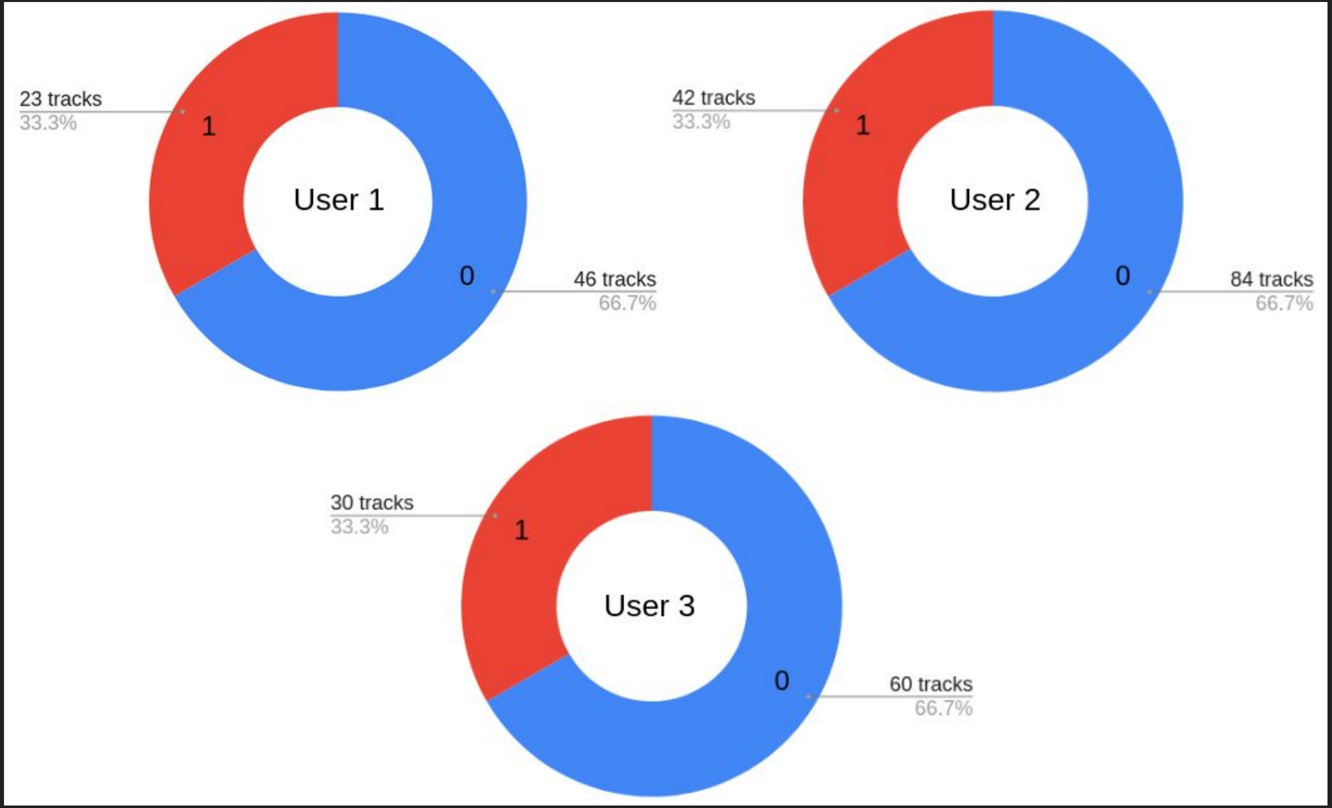
- Track names -> similarity score -> recommendation
- Acoustic features -> similarity score -> recommendation

2. Model-based method

- Track names + artist names + acoustic features -> label prediction (like or dislike a non-listened track)

1. Data

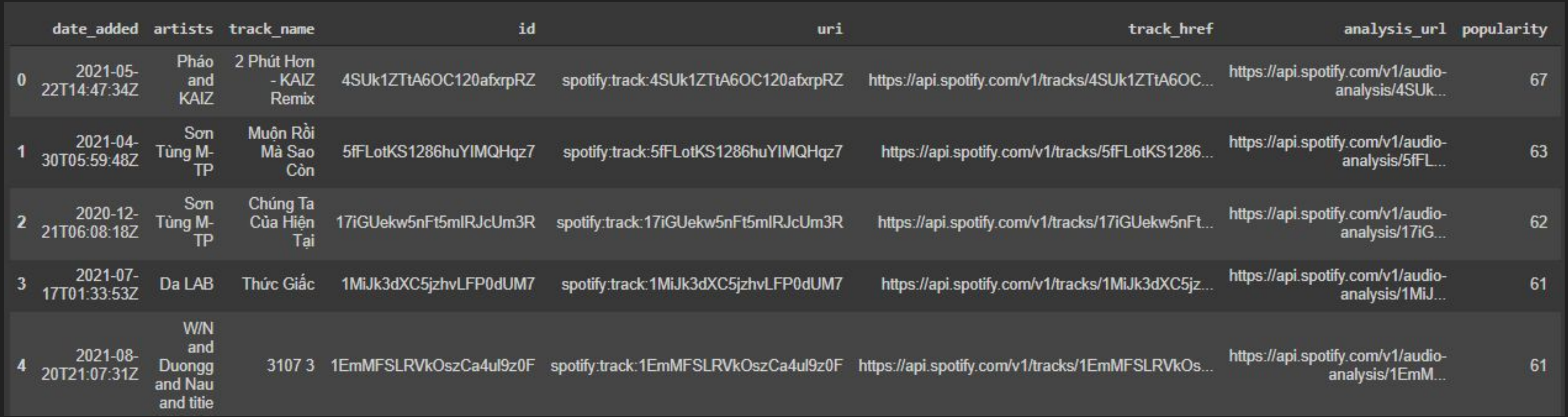
- Collected from Spotify.
- Full dataset: 1781 rows (tracks)
- Users' dataset (3 users) (real users’ listening tracks):
    - Content-based method -> **recommend songs by similarity score**:
        - user 1: 13 tracks
        - user 2: 32 tracks
        - user 3: 20 tracks
    - Model-based method -> **predict songs that user will whether like or not**:
        - user 1: 69 tracks - 40 artists
        - user 2: 126 tracks - 91 artists
        - user 3: 90 tracks - 73 artists


23 tracks
33.3%
1
User 1
0
46 tracks
66.7%
42 tracks
33.3%
1
User 2
0
84 tracks
66.7%
30 tracks
33.3%
1
User 3
0
60 tracks
66.7%

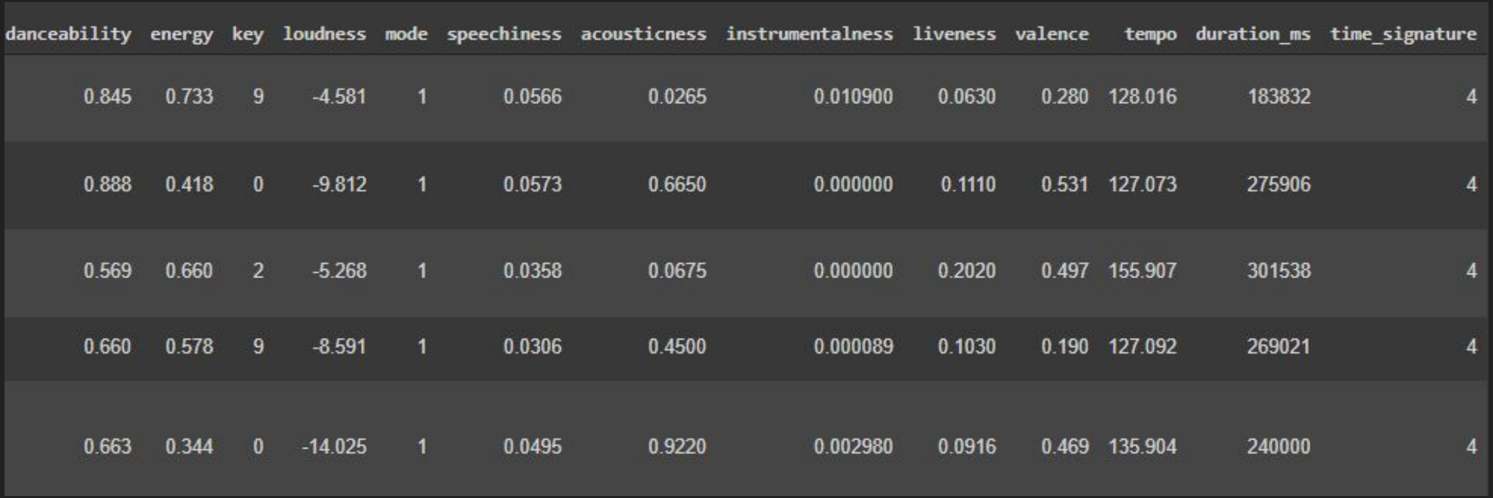
Features

21 features

- Editorial features: date_added, artists, track_name, id, uri, track_href, analysis_url.
- Acoustic features: popularity, danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, duration_ms, time_signature.



	date_added	artists	track_name	id	uri	track_href	analysis_url	popularity
0	2021-05-22T14:47:34Z	Pháo and KAIZ	2 Phút Hơn - KAIZ Remix	4SUk1ZTtA6OC120afxrpRZ	spotify:track:4SUk1ZTtA6OC120afxrpRZ	https://api.spotify.com/v1/tracks/4SUk1ZTtA6OC...	https://api.spotify.com/v1/audio-analysis/4SUk...	67
1	2021-04-30T05:59:48Z	Sơn Tùng M-TP	Muộn Rồi Mà Sao Còn	5fFLotKS1286huYIMQHqz7	spotify:track:5fFLotKS1286huYIMQHqz7	https://api.spotify.com/v1/tracks/5fFLotKS1286...	https://api.spotify.com/v1/audio-analysis/5fFL...	63
2	2020-12-21T06:08:18Z	Sơn Tùng M-TP	Chúng Ta Của Hiện Tại	17iGUekw5nFt5mIRJcUm3R	spotify:track:17iGUekw5nFt5mIRJcUm3R	https://api.spotify.com/v1/tracks/17iGUekw5nFt...	https://api.spotify.com/v1/audio-analysis/17iG...	62
3	2021-07-17T01:33:53Z	Da LAB	Thức Giấc	1MiJk3dXC5jzhvLFP0dUM7	spotify:track:1MiJk3dXC5jzhvLFP0dUM7	https://api.spotify.com/v1/tracks/1MiJk3dXC5jz...	https://api.spotify.com/v1/audio-analysis/1MiJ...	61
4	2021-08-20T21:07:31Z	W/N and Duongg and Nau and titie	3107 3	1EmMFSLRVkOszCa4ul9z0F	spotify:track:1EmMFSLRVkOszCa4ul9z0F	https://api.spotify.com/v1/tracks/1EmMFSLRVkOs...	https://api.spotify.com/v1/audio-analysis/1EmM...	61



danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration_ms	time_signature
0.845	0.733	9	-4.581	1	0.0566	0.0265	0.010900	0.0630	0.280	128.016	183832	4
0.888	0.418	0	-9.812	1	0.0573	0.6650	0.000000	0.1110	0.531	127.073	275906	4
0.569	0.660	2	-5.268	1	0.0358	0.0675	0.000000	0.2020	0.497	155.907	301538	4
0.660	0.578	9	-8.591	1	0.0306	0.4500	0.000089	0.1030	0.190	127.092	269021	4
0.663	0.344	0	-14.025	1	0.0495	0.9220	0.002980	0.0916	0.469	135.904	240000	4

2. Methodology

- Content-based method
- Model-based method

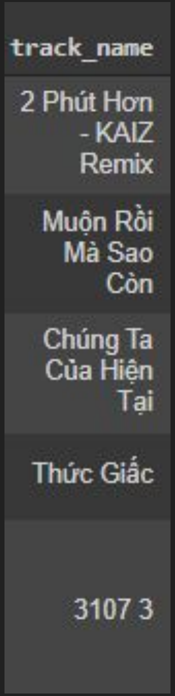
Content-based method

Preprocessing -> track_name

1. Lowercase
2. Remove punctuation
3. Remove duplicate white space
4. Number to words
5. Normalize diacritics

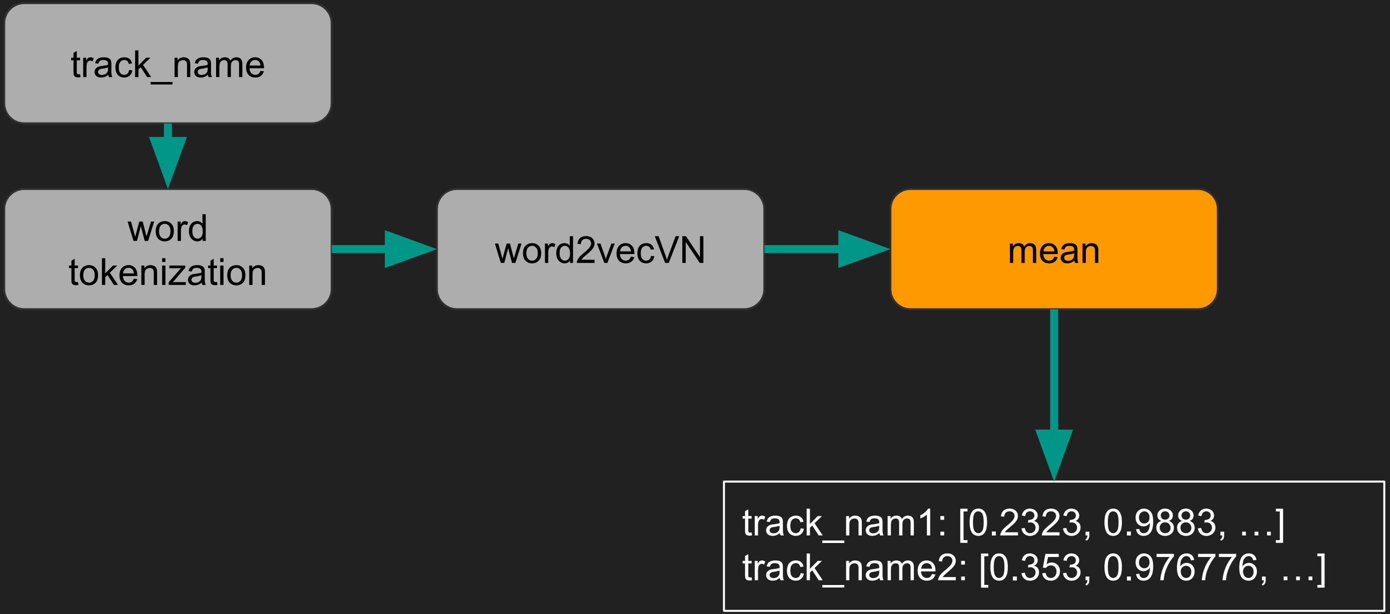
Feature extraction

1. Morpheme-based vs Phrase-base tokenization
2. Padding (max length = 17)
3. Embedding vector: word2vecVN - 400 dims



track_name
2 Phút Hơn - KAIZ Remix
Muộn Rồi Mà Sao Còn
Chúng Ta Của Hiện Tại
Thức Giấc
3107 3

Content-based method



track_nam1: [0.2323, 0.9883, …]
track_name2: [0.353, 0.976776, …]

Content-based method

Acoustic features: Popularity, Danceability, Energy, Key, Loudness, Mode, Speechiness, Acousticness, Instrumentalness, Liveness, Valence, Tempo, Duration_ms, Time_signature

[47, 0.531, 0.65, 4, -7.764, 0, 0.0809, 0.923, 0.782, 0.151, 0.386, 137.853, 262164, 3]

Content-based method

Experiment on 3 types of vector

- Track name -> Morpheme-based tokenization -> ... -> similarity score -> recommendations

- Track name -> Phrase-based tokenization -> ... -> similarity score -> recommendations

- Acoustic features -> similarity score -> recommendations

Content-based method

Similarity metrics:

- Pearson:

$$sim(A,B) = \cos\varphi = \frac{A \cdot B}{|A| \cdot |B|},$$

- Cosine:

$$sim(A,B) = \frac{\sum_{i=1}^{N}(A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum_{i=1}^{N}(A_i - \bar{A})^2}\sqrt{\sum_{i=1}^{N}(B_i - \bar{B})^2}},$$

Model-based method

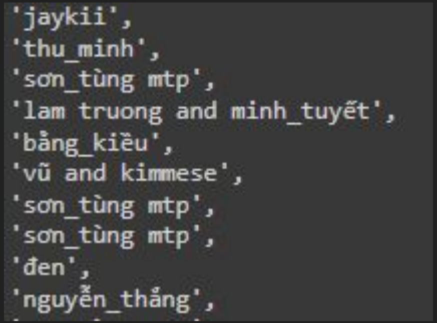
Preprocessing

1. Text Preprocessing for artists and track_name features:
2. Vectorize artists and track_name using word2vec:
3. Concatenate artists and track_name vectors and other features:

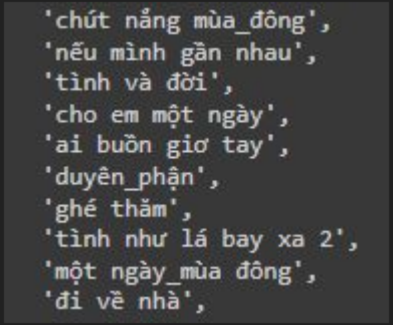
Model-based method

Preprocessing

1. Text Preprocessing for artists and track_name features:
   a. Remove punctuations
   b. Tokenize word level using pyvi
   c. Lowercase



```
'jaykii',
'thu_minh',
'sơn_tùng mtp',
'lam truong and minh_tuyết',
'bằng_kiều',
'vũ and kimmese',
'sơn_tùng mtp',
'sơn_tùng mtp',
'đen',
'nguyễn_thắng',
```

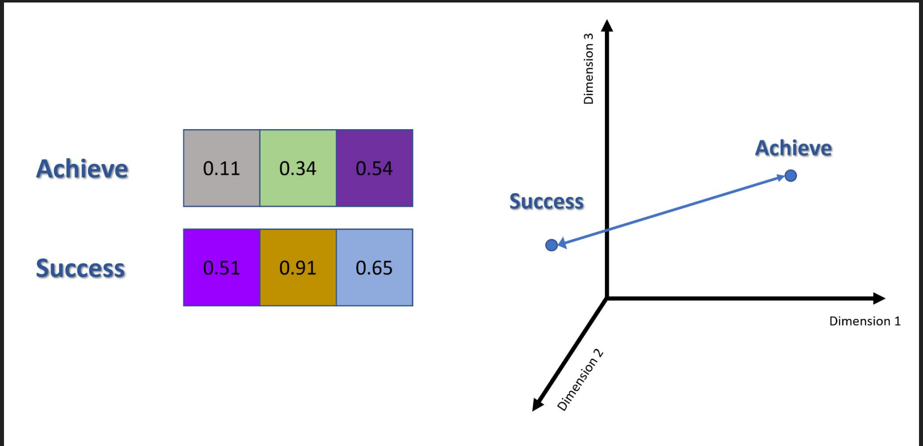


```
'chút nắng mùa_đông',
'nếu mình gần nhau',
'tình và đời',
'cho em một ngày',
'ai buồn giơ tay',
'duyên_phận',
'ghé thăm',
'tình như lá bay xa 2',
'một ngày_mùa đông',
'đi về nhà',
```

Model-based method

Preprocessing

2. Vectorize artists and track_name using Word2Vec:



Achieve	0.11	0.34	0.54
Success	0.51	0.91	0.65

Model-based method

Preprocessing

2. Vectorize artists and track_name using word2vec:

Training word2vec model using gensim library:

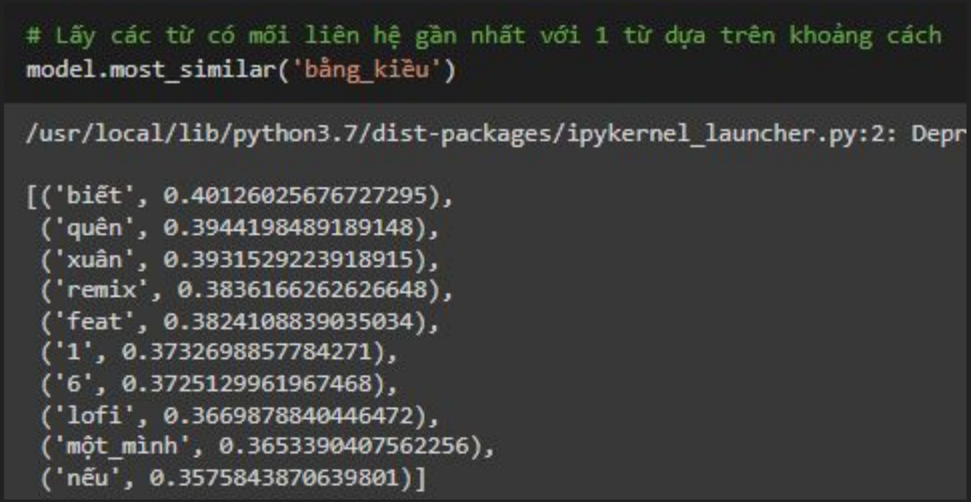
- Input: artists or track_name
- Output: 1D vector (150,)

Model-based method

Preprocessing

2. Vectorize artists and track_name using word2vec:

Training word2vec model using gensim library:



```
# Lấy các từ có mối liên hệ gần nhất với 1 từ dựa trên khoảng cách
model.most_similar('bảng_kiều')
```

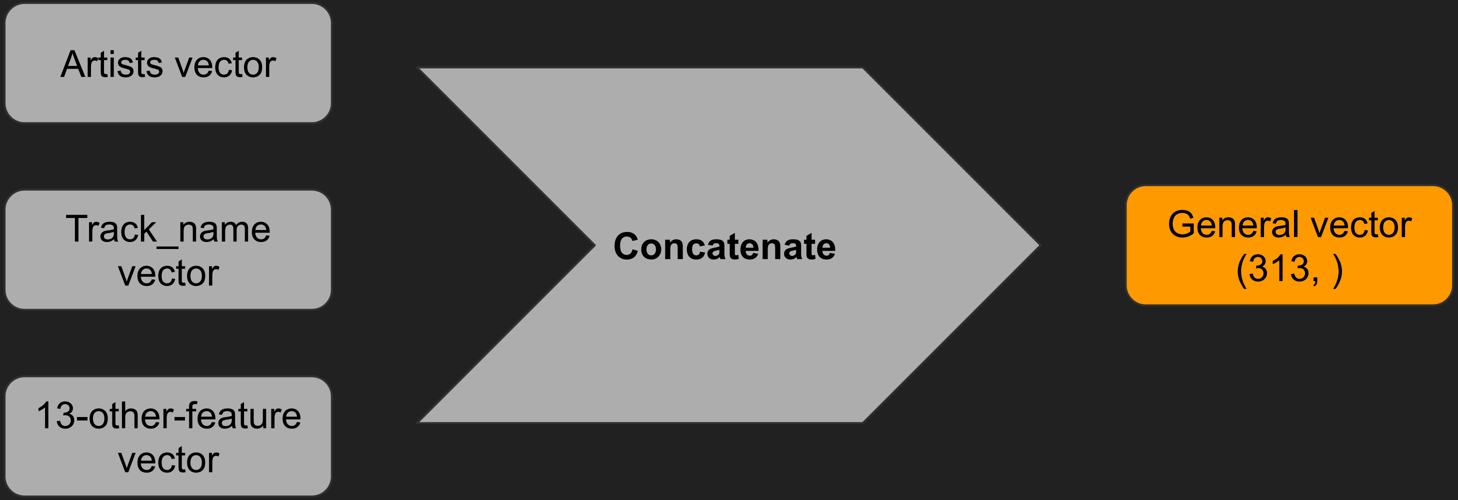
```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:2: Depr

[('biết', 0.40126025676727295),
 ('quên', 0.3944198489189148),
 ('xuân', 0.3931529223918915),
 ('remix', 0.3836166262626648),
 ('feat', 0.3824108839035034),
 ('1', 0.3732698857784271),
 ('6', 0.3725129961967468),
 ('lofi', 0.3669878840446472),
 ('một_mình', 0.3653390407562256),
 ('nếu', 0.3575843870639801)]
```

Model-based method

Preprocessing

3. Concatenate artists and track_name vectors and other features:



Artists vector

Track_name vector

13-other-feature vector

Concatenate

General vector (313,)

Model-based method

Models:

1. Logistic Regression
2. SVM
3. LightGBM
4. Boosting Decision Tree

Model-based method

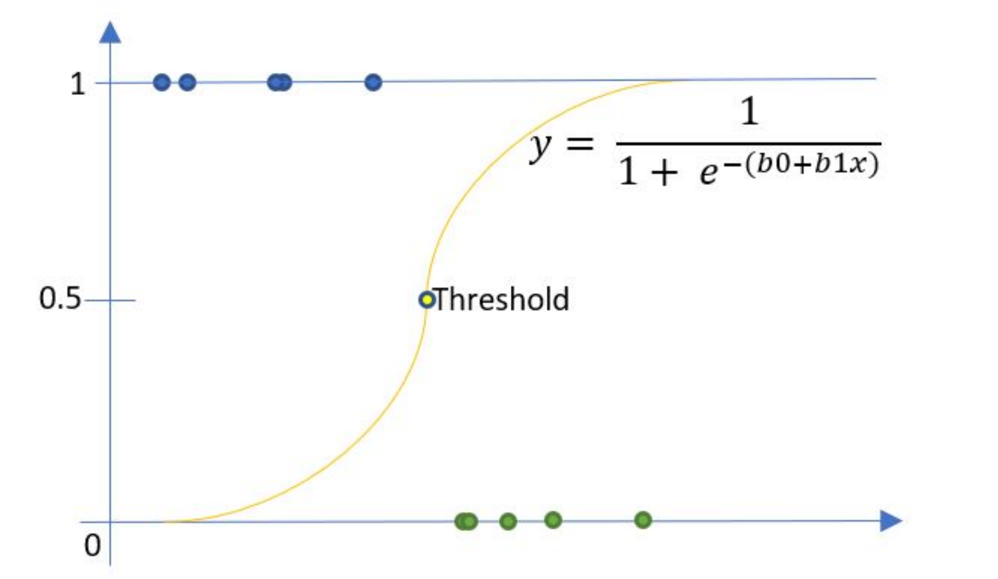
Models:

1. Logistic Regression

Sử dụng hàm Sigmoid

$$f(x) = b0 + b1x,$$

$$y(f(x)) = \frac{1}{1 + e^{-f(x)}}.$$

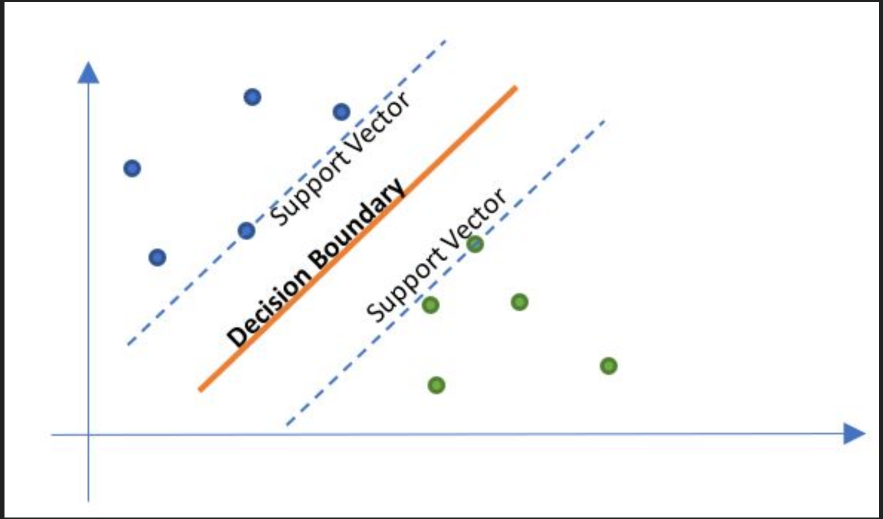


Model-based method

Models:

2. SVM

Sử dụng Decision Boundary
-> Phân tách 2 miền dữ liệu.

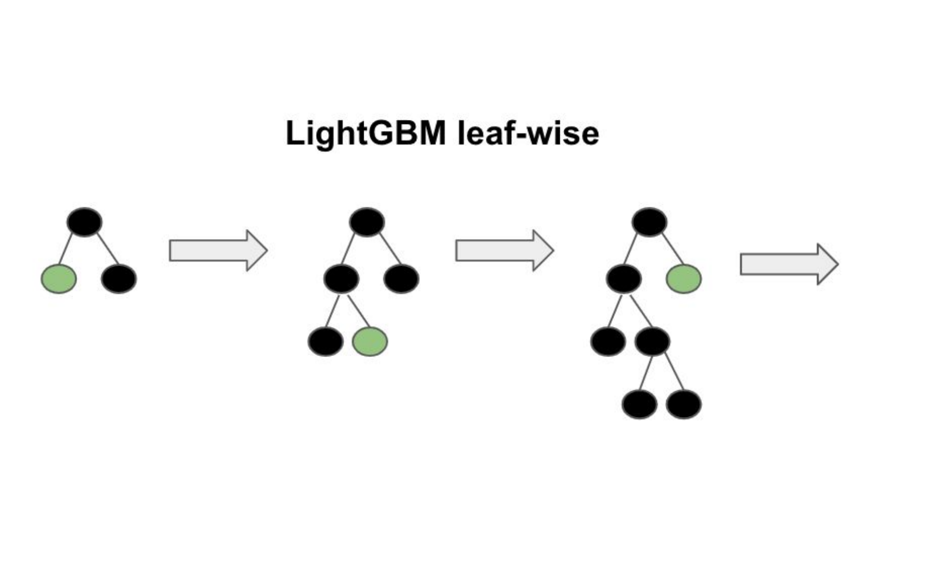


Model-based method

Models:

3. LightGBM

LightGBM



LightGBM mở rộng cây quyết định theo hướng leaf-wise.

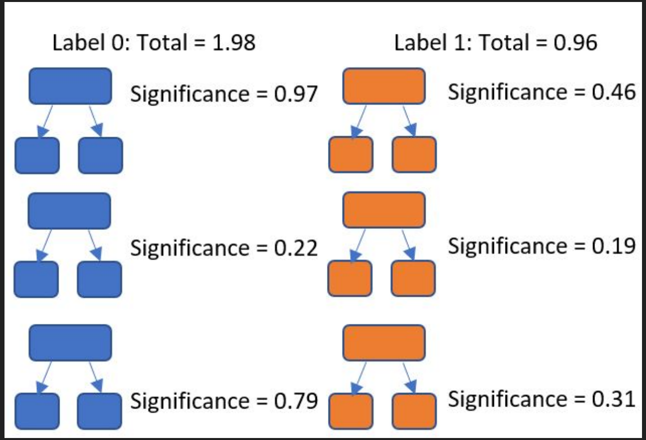
Model-based method

Models:

4. Boosting Decision Tree

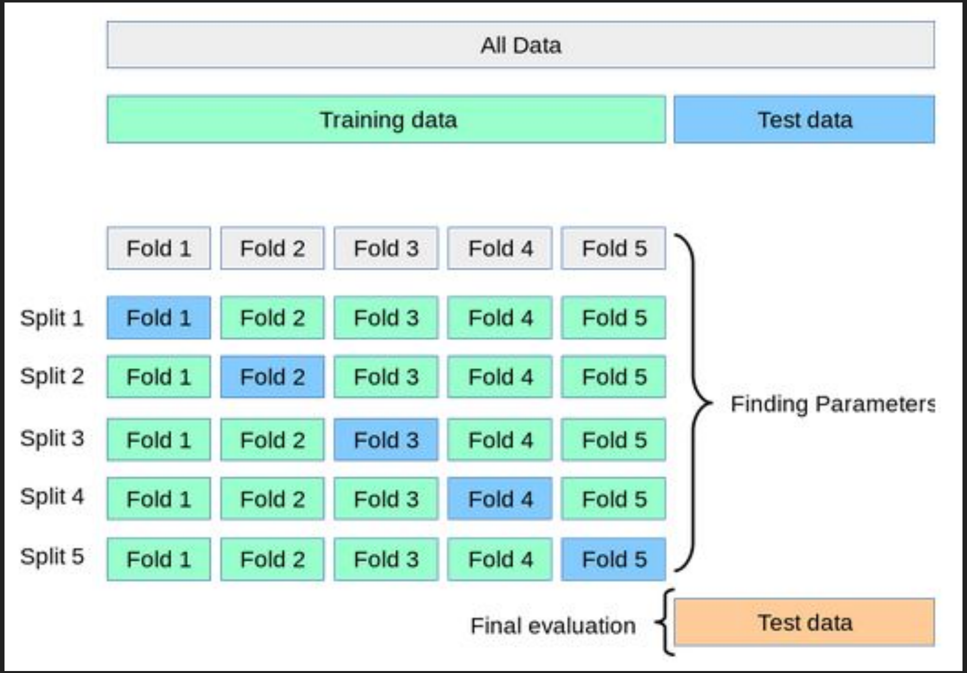
- Sử dụng AdaBoostClassifier và DecisionTreeClassifier đến từ sklearn.

1.Khởi tạo trọng số cho các điểm dữ liệu: weight = 1/(số điểm dữ liệu).
2.Xây dựng cây quyết định cho từng feature.
3.Tính mức độ quan trọng của các cây quyết định dựa trên kết quả phân loại.
4.Cập nhật trọng số ở bước 1(và chuẩn hóa trọng số).
5.Lặp lại các bước trên(số lần lặp = số estimators).
6.Sử dụng rừng cây quyết định với độ quan trọng của chúng để đưa ra dự đoán phân loại.



Model-based method

Evaluation:



All Data

Training data	Test data

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 2	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 3	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 4	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

Finding Parameters

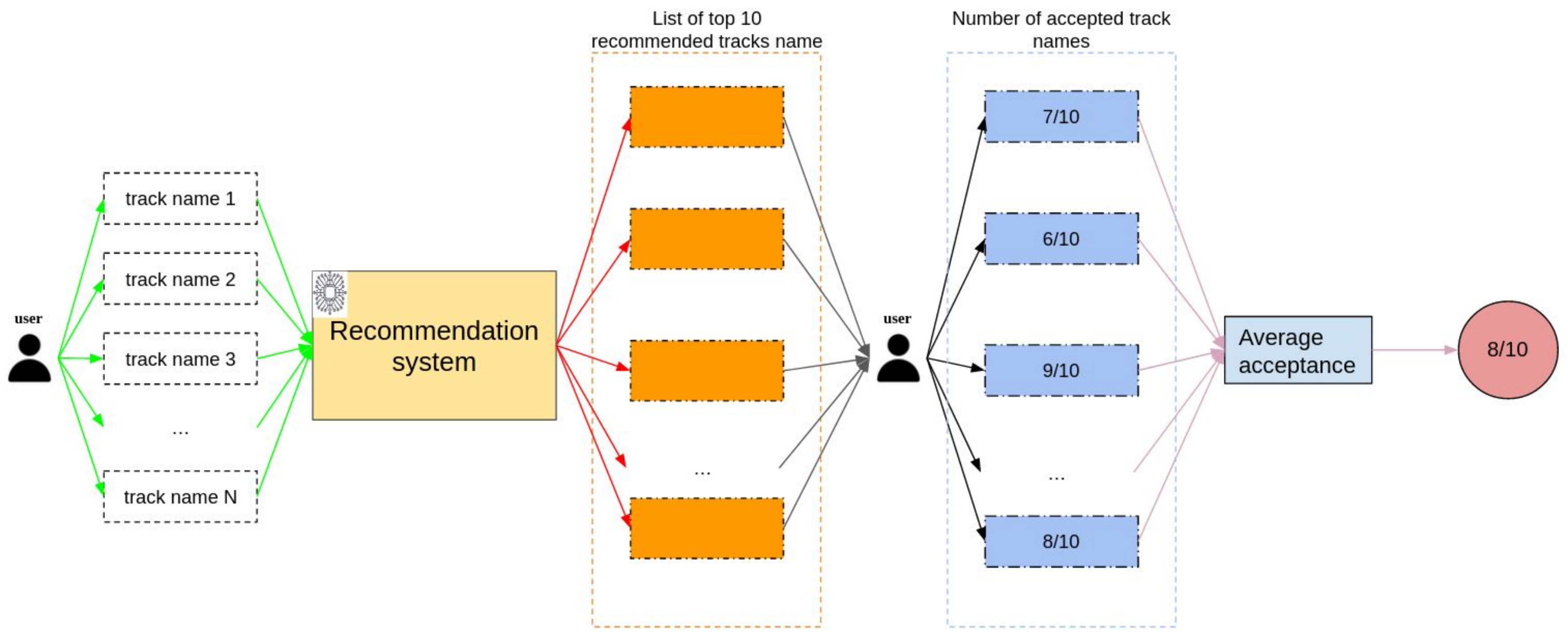
Final evaluation { Test data

Metrics

- Human evaluation
- Accuracy
- F1-score

Metrics

- Human evaluation



3. Experiments and Results

Content-based method

- Human evaluation rate:

	Morpheme-based tokenization	Phrase-based tokenization	Acoustic features
User 1	6/10	7/10	8/10
User 2	5/10	**8/10**	8/10
User 3	**8/10**	7/10	**9/10**

3. Experiments and Results

Model-based method

- Human evaluation rate:

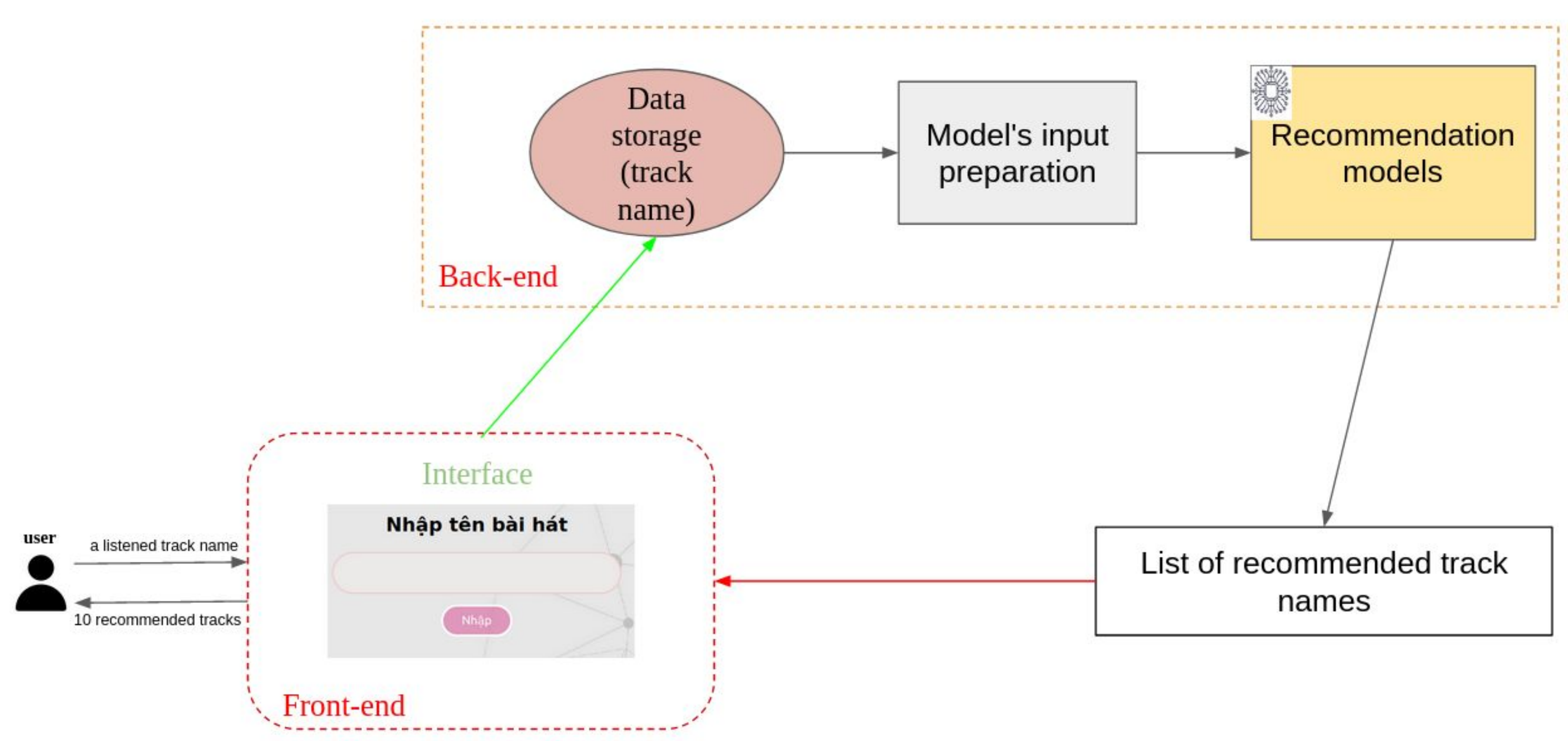
	SVM	Logistic regression	Boosting decision tree	LightGBM
User 1	**7/10**	7/10	**8/10**	**9/10**
User 2	**7/10**	**8/10**	**8/10**	8/10
User 3	5/10	6/10	7/10	7/10

Model-based method

Evaluation: 10-folds training result

	Logistic Regression		SVM		LightGBM		Boosting Decision Tree	
	f1	acc	f1	acc	f1	acc	f1	acc
User 1	0.66	0.82	0.09	0.68	1.0	1.0	0.95	0.97
User 2	0.16	0.69	0.17	0.66	0.95	0.97	0.88	0.93
User 3	0.12	0.65	0.00	0.69	0.32	0.74	0.27	0.6

3. Recommendation Operation Interface:



Thank you